applied sciences

MDPI

Article

Long-Term Categorization Capability Assessment of Semantically Reduced Concept Lattices

Milena Frtunić Gligorijević *, Miloš Bogdanović, Darko Puflović and Leonid Stoimenov

Faculty of Electronic Engineering, University of Niš, 18000 Niš, Serbia; milos.bogdanovic@elfak.ni.ac.rs (M.B.); darko.puflovic@elfak.ni.ac.rs (D.P.), leonid.stoimenov@elfak.ni.ac.rs (L.S.)

* Correspondence: milena.frtunic.gligorijevic@elfak.ni.ac.rs

Abstract

The findability and usability of open data have never been more important due to the volume of available data. Therefore, methods have been developed to improve the dataset metadata quality, enhancing its findability and usability. One effective approach for filling in missing category information is based on the Formal Concept Analysis method. Since this method relies on a knowledge base consisting of concept lattices produced for each category on an open data portal, the focus of this research is to analyze the usability of the semantically reduced concept lattice over time. In particular, we are focused on (1) the impact of the semantic reduction threshold on the long-term usability and expressiveness of the generated concept lattices and (2) the identification of metadata change indicators that would require recreation of the knowledge bases. We perform this analysis using Ireland's and Canada's open data portals—two portals with different approaches to dataset categorization. For both portals, datasets from 2020, 2021, and 2023 are used for knowledge bases creation, while semantically reduced concept lattices are evaluated using six test datasets per portal, containing new datasets from 2020 to 2025. We will show that semantically reduced concept lattices maintain good long-term usability under regular dataset growth. Among other findings, we will demonstrate that dataset categorization based on two-year-old concept lattices achieves over 83% accuracy and that at the category level, even four-year-old reduced concept lattices attain accuracy exceeding 90%.

Keywords: open data; formal concept analysis; semantic similarity; categorization; natural language processing; open data portal; concept lattice; reduction

Check for updates

Academic Editor: Stefan Fischer

Received: 23 April 2026
Revised: 25 May 2026
Accepted: 26 May 2026
Published: 1 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The release of various content on the web has become increasingly common due to the many projects undertaken in the past. The idea of open data, which allows for public access to data for various uses, is the key motivation behind these initiatives. It is believed that open data will benefit the public, businesses, and individuals, as it is seen as a driver of economic development, knowledge enhancement, and innovation [1]. In a proposal for a directive of the European Parliament and of the Council on the reuse of public sector information (recast), estimated total value of public sector information could rise to 194 billion by 2030 [2].

Government agencies are disseminating information from various domains, such as statistics, transportation, the environment, permits, licenses, budgets, geography, and the economy, to enhance transparency and engage civil societies [3]. This initiative is referred to as Open Government Data (OGD). Data is made available through Open Data Portals

(ODPs) to maximize its potential benefits. These portals are similar to digital libraries that offer catalogs and databases for organizing files. Each dataset published on the ODP is supported by metadata, which is structured descriptive information containing various details about the respective resource. The keys signify property names, while the values denote the information linked to those properties. Metadata is organized in the form of key value pairs.

Open data portals provide metadata-based search tools to facilitate data discoverability. Some of the most straightforward and user-friendly search options are based on the category and tags. Category-based search selects datasets that are linked to the same topic, whereas tag-based search provides more concentrated browsing based on the phrases used for the dataset's description. However, incomplete or erroneous information hinders the effectiveness of the search process, yields only partial results, and makes it more difficult to locate facts. This problem is becoming increasingly apparent due to the amount of publicly available datasets and the growth of the ODP.

The quality of search results on open data portals depends on how complete the metadata entries are. Due to the increasing number of datasets on open data portals, inadequate metadata has a substantial impact on the discoverability, use, and reuse of datasets, particularly in cases when no classification is provided. Numerous studies indicate the importance of metadata for proper dataset discovery [4,5] and interpretation [6]. Since then, improving metadata quality has received a lot of attention from the scientific community. However, several studies have shown that information is often missing [7], which has led to the creation of techniques for filling in the missing metadata.

One approach to addressing the gaps in category information is to utilize the existing metadata, specifically the tags that can be used to categorize datasets. The Formal Concept Analysis (FCA) technique has proven its open data categorization capabilities by constructing concept lattices for each available category on an open data portal [8]. These concept lattices act as a knowledge base for categorizing datasets based on the relevant tags that define them and outline a hierarchy of tag usage within an individual category. However, as the number of datasets on a portal increases, the variety of tags and their combinations used to describe the datasets also expands. Consequently, it is essential to update the knowledge base or parts of it, specifically the concept lattices for categories that have experienced significant changes, by incorporating new datasets with varying tags to maintain the accuracy of the categorization process over time. In such cases, the high computational cost of knowledge processing is a significant obstacle in practical FCA implementations. The task of creating concept lattices that are appropriately sized and structured to represent crucial contextual features continues to be a major objective in FCA [9,10]. Numerous methods for reducing concept lattices have been discussed in existing literature [11] and their usage has facilitated FCA application in various fields of data science. Nevertheless, the question of durability, usability, and efficiency of the reduced concept lattices for the task of the categorization remains open for evolving metadata environments such as open data portals.

In this research, we analyze the usability and efficiency of the semantically reduced concept lattices for the task of open data categorization in open data portals with evolving metadata. Since, in this case, concept lattice acts as a knowledge base for categorizing datasets, our analysis is focused on measuring the usability and efficiency of knowledge bases generated from data originating from different periods of time. The measurements are performed by evaluating the precision of the FCA-based categorization algorithm [12], resulting in a twofold contribution:

- Determining the influence of a reduction threshold on the knowledge base durability and usability.
- Discovering a set of metadata volume and structure change indicators that impose knowledge base recreation.

In order to perform an in-depth analysis of the long-term usability of the reduced concept lattices, it is essential to evaluate their behavior across portals with different dataset categorization approaches, particularly regarding the number of categories that may be assigned to a single dataset. Accordingly, this study utilizes two national open data portals with contrasting categorization approaches: Ireland's open data portal that assigns only one category per dataset, and Canada's open data portal that assigns multiple categories per dataset. Datasets from Ireland and Canada ODPs across 2020, 2021, and 2023 are used to construct knowledge bases with different reduction thresholds, while six derived test datasets per portal, capturing newly available data across different year intervals (from 2020 to 2025), are used for categorization purposes.

2. Related Work

2.1. Open Data—Metadata Usage

The goal of the open data movement is to encourage the reuse of public sector data by making it publicly accessible online [13]. The ultimate benefit might be enormous, despite the initial objective being rather modest. The quantity of publicly available data has grown dramatically in the past 20 years. The market potential for strategic planning, cost reduction, and productivity would reach billions of dollars if this data could be utilized properly [14].

Open data is typically made accessible for consumption through so-called "one-stop-shop" ODPs, which centralize data from heterogeneous sources. Users may visually identify and use data through the ODP's user interface, or developers can utilize it to create specialized applications for open data analysis and reuse. In most cases, ODPs are based on open-source platforms like Drupal Knowledge Archive Network (DKAN) or Comprehensive Knowledge Archive Network (CKAN), while some utilize proprietary software like Opendatasoft or Socrata [15]. When it comes to developers' use of open data, the dataset format that is accessible through APIs is more important than the platform itself. The structure of the dataset is not intended for human reading but rather for machine processing. It depends on metadata, which is similar to DNA for data—basic, useful components that define the purpose and content. Open data portals mostly rely on the Data Catalog vocabulary (DCAT) [16,17]. Consequently, datasets accessible through APIs share a consistent structure across several metadata keys, such as title, description, groups, publisher, tags, resources, and so on. These keys may be used for several purposes, including programmatically exploiting datasets, analyzing and comprehending datasets to uncover hidden information.

Maintaining high-quality data and related metadata is crucial as the quantity and diversity of data sources grow [5]. End consumers are not interested in poorly explained data, as it is unable to provide new value for them. Poorly described data impacts both the integration with other datasets across portals as well as the discovery and consumption of a dataset inside a single portal [18]. Thus, we may designate data harmonization—which entails conforming data to open data publishing standards—as the most crucial stage prior to releasing open data. The eight open government data principles outline the fundamental characteristics of data that must exist for it to be classified as "open" [19]. The openness qualities of datasets cannot be automatically determined without values for metadata keys like organization, group, or tags.

For data segmentation and quality assessment, a number of meta-keys are employed, but groups and tags are among the most crucial. Groups combine data under a certain subject or category, whereas tags are mostly used to quickly search and retrieve related datasets. Thus, both are meant for dataset consumers. The process of tagging datasets is fraught with issues including synonyms, incoherence, misspellings, and ambiguity [20], which impede the datasets' findability.

The manual tag creation process is laborious and prone to human mistakes, as tags may not always be correct or pertinent to the dataset in question [21,22]. Furthermore, because tagging is time-consuming and one of the obstacles to data openness, publishers may completely ignore it [19]. Although tags can appear to be a trivial aspect, the present practice demonstrates that OGD portals, especially "trend-setters" like the Estonian open data portal, sometimes struggle with both their presence and relevance to the actual dataset [23]. Search engines are unable to use semantic relations across datasets because values for metadata keys frequently indicate provenance information rather than concepts from the reported dataset's domain [24].

Since metadata keys and values impact how datasets are understood and used, it is crucial to maintain them continuously. Cleaning and reconciling metadata, then using common metadata values across ODPs, would be the solution to this issue. Formal Concept Analysis (FCA) has proven its value for determining the similarity of metadata usage and open data portal metadata alignment [25]. In the context of open data portals, FCA generates knowledge structures that have been successfully used for metadata alignment by taking advantage of semantic similarity among existing metadata values. Still, FCA results, e.g., knowledge structures, depend on the actual metadata values (groups, categories, tags), which raises the question regarding their durability. Open data portals are evolving metadata ecosystems and the usability of knowledge structures generated at a particular point of time in such a system should be investigated.

2.2. Formal Concept Analysis in a Nutshell

Rudolf Wille formalized Formal Concept Analysis (FCA) in 1982 [26], presenting it as a lattice theoretic method (sometimes nicknamed "applied lattice theory") for building concept hierarchies and extracting key relationships. Its information-preserving property led to rapid popularity in data analysis, albeit at the expense of potentially significant computational resources. A core step in the Formal Concept Analysis (FCA) process is examining the relationships between objects and their attributes within a specific domain. These relationships are captured in a binary table known as a formal context—essentially an object-by-attribute matrix where rows represent objects, columns represent attributes, and a cross (or 1) in a cell indicates that the object possesses that attribute. When FCA is applied to this formal context, it generates sets of objects and attributes that embody the domain's knowledge. These sets are then organized into a concept lattice, which visually reveals various rules, implications, and connections.

The following definitions constitute the foundation of FCA formal description:

Definition 1 ([26]). *A formal context is a triple* $K \coloneqq (G, M, I)$ *which consists of a set G of objects, a set M of attributes, and a binary relation* $I \subseteq G \times M$. $(g, m) \in I$ *is read as "object g has attribute m".*

Definition 2 ([26]). *For* $A \subseteq G$, *let* $A^I \coloneqq \{m \in M | \forall g \in A : (g, m) \in I\}$, *and dually, for* $B \subseteq M$, *let* $B^I \coloneqq \{g \in G | \forall m \in B : (g, m) \in I\}$.

If the following conditions are met: $A \subseteq G$, $B \subseteq M$, $A^I = B$, $B^I = A$, *then a pair (A, B) is a formal concept. Set A is named concept extent while set B is named concept intent.*

Definition 3 ([26]). *The set $S(C)$ of all concepts of a formal context C together with a partial order $(A_1, B_1) \leq (A_2, B_2) :\leftrightarrow A_1 \subseteq A_2$ (which is equivalent to $B_1 \supseteq B_2$) is a complete lattice of C.*

FCA is considered as a data analysis method based on the usage of the relationship between a particular set of objects and a particular set of attributes. In most cases, an incidence matrix is used to visualize the relationship between these two distinct sets. Thus, an incidence matrix visualizes a formal context. An example of formal context visualization is given in Table 1 where airlines of the Star Alliance group represent the set of objects, while the attribute set contains their destinations [27]. Each row within the incidence matrix represents an object, each column represents an attribute while matrix field values (usually bool or yes/no values) denote the relationship between object and attribute. In the case of formal context presented in Table 1 incidence matrix describes which destinations are served by which Star Alliance member.

Table 1. Incidence matrix—a formal context about the destinations of the Star Alliance members [27].

	Latin America	**Europe**	**Canada**	**Asia Pacific**	**Middle East**	**Africa**	**Mexico**	**Caribbean**	**The United States**
Air Canada	X	X	X	X	X		X	X	X
Air New Zealand		X		X					X
All Nippon Airways		X		X					X
Ansett Australia				X					
The Austrian Airlines Group		X	X	X	X	X			X
British Midland		X							
Lufthansa	X	X	X	X	X	X	X		X
Mexicana	X		X				X	X	X
Scandinavian Airlines	X	X		X		X			X
Singapore Airlines		X	X	X	X	X			X
Thai Airways International	X	X		X				X	X
United Airlines	X	X	X	X			X	X	X
VARIG	X	X		X		X	X		X

The most significant output of FCA is the concept lattice. FCA derives concepts from the incidence matrix and generates a collection of formal concepts. This collection is logically organized into a hierarchy of concepts interconnected using subconcept–superconcept relations—a concept lattice [27]. In Figure 1 we can see a concept lattice generated for a formal context depicted in Table 1. It is easy to notice that the concept lattice reflects the generalization and specialization between formal concepts within a single formal context [27]. If formal concepts are considered the basic units of thought, then it is no surprise FCA has found its various applications.

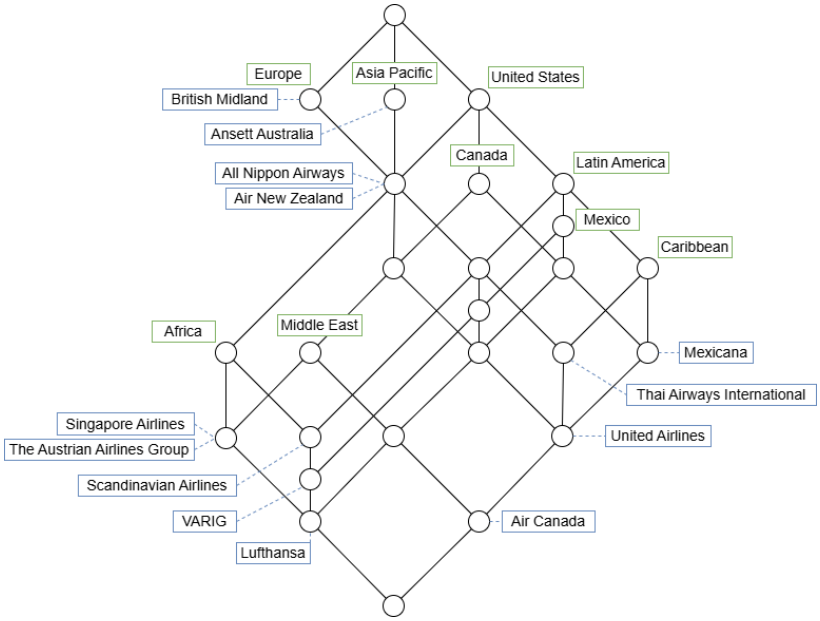


Figure 1. The concept lattice of the formal context in Table 1.

The value of FCA largely stems from the fact that experts can visually inspect the concept lattice to reveal meaningful patterns and dependencies. Therefore, generating all formal concepts and their lattice is essential for FCA applications. Over the years, FCA has been successfully applied to knowledge discovery [28–32], knowledge representation [33], pattern recognition tasks [34,35], web usage mining [36,37], and many other data science fields. To elaborate on the literature on FCA, Poelmans et al. presented a survey [28]. Their study examined 702 articles published between 2003 and 2009 that referenced formal concept analysis in their abstracts, with particular emphasis placed on a subset of 140 publications centered solely on data mining and knowledge discovery. A comprehensive 2013 survey [29] re-examined and visualized the key research areas in formal concept analysis. From 2011 to 2016, Singh et al. [38] analyzed over 350 publications and highlighted the following major directions: FCA combined with granular computing, fuzzy FCA, interval-valued fuzzy approaches, possibility theory, rough sets, triadic concepts, factor concepts, and incomplete formal contexts. The study reaffirmed FCA's prominent role in knowledge discovery, logical reasoning, decision contexts, and ontology engineering. Azibi et al. [39] reviewed various FCA-based classification approaches and grouped them into three main categories—distributed classifiers, ensemble classifiers, and mono-classifiers, thereby highlighting the effectiveness of Formal Concept Analysis in machine learning.

With the rise of big data, FCA has emerged as a valuable technique for handling large-scale datasets. This makes efficient and reliable FCA methods critical for knowledge extraction and representation. In response, researchers have proposed a variety of parallel and distributed algorithms to accelerate the computation of all relevant formal concepts. The researchers proposed parallel implementations based on the Close-by-One algorithm [40] and, in 2009, presented the earliest distributed FCA algorithm [41] using the MapReduce paradigm [42]. Separately, Chunduri et al. [43] proposed a distributed approach called UNConceptGeneration to derive novel insights from binary object-attribute relational data.

Numerous techniques have been proposed to build reasonably sized concept lattices that highlight the most important features of the data. These falls into two main groups: (a) methods that reduce complexity with a context-level representation, and (b) lattice-level pruning techniques, such as iceberg reduction. In general, the goal is often to remove as

many objects or attributes as possible without altering the core lattice structure. Commonly used approaches, as reviewed in [11], include context preprocessing [44–47], non-essential distinction elimination [48–50], and selective concept filtering [51,52]. The method applied in this research for reducing concept lattices falls within the context of preprocessing techniques. This method employs an attribute reduction process by replacing redundant attributes with its semantically similar attribute, thereby reducing the total number of used attributes.

Current state-of-the-art approaches, such as machine learning or embedding-based approaches, confirm FCA to be complementary rather than comparable to them. FCA has successfully combined decision trees to create ML models generating modern classifiers and regressors [53]. The authors of [54] induce a decision tree from the concept lattice, while [55] compares the ways the concept lattice and the decision tree can be used for supervised learning. Moreover, FCA has been successfully combined with artificial neural networks in neuro-symbolic approaches. Integration of these approaches is bidirectional. FCA has been used for modernization of the neural network architecture based on concept lattices [56], as well as improving low dimensional node embeddings without parameter tuning [57]. On the other hand, the formal lattice graph of FCA has been reproduced using generative neural models for graphs [58].

3. Research Problem Definition

Open data metadata completeness, such as metadata for category information, is one of the key prerequisites for open data discoverability and its ability to reach its full potential on open data portals. Therefore, filling out missing metadata values such as category information, in order to improve metadata quality and consequentially open data findability, has been recognized as an important process for improving the quality of open data. Methods relying on Formal Concept Analysis have proven effective for determining appropriate categories for datasets based on their tags [12]. One of the core parts of these methods is the knowledge base that consists of concept lattices generated based on the combinations of tags used for describing already categorized datasets. Therefore, in order to maintain the quality of FCA-based categorization methods, it is essential to periodically update the concept lattices. Furthermore, concept lattice generation algorithms commonly face exponential complexity as the number of attributes and/or objects increases, resulting in the generation of extremely large concept lattices with potentially a significant number of concepts with arguable usability. Therefore, the algorithm for the reduction of set of attributes, in this case set of tags, has shown great potential [25].

For that reason, within this research, we have decided to address two important aspects of FCA application for open data categorization bringing usability and durability conclusions and recommendations:

- The impact of the semantic reduction measure on the usability of created concept lattices over time—a lower reduction threshold retains more general concepts, but can also reduce the expressiveness of the used terms which can lead to losing specific terms in the particular field. Alternatively, a higher reduction threshold will keep the expressiveness of the terms but will result in a more complex concept lattices. Therefore, within this part of the analysis, we will examine how the reduction threshold influences the long-term usability of the generated concept lattices.
- Metadata change indicators, due to which it is necessary to recreate the knowledge base for the selected reduction measure.

3.1. Formal Context Reduction

The main idea behind formal context reduction is to reduce the number of attributes describing objects in the formal context. For FCA-based open data categorization, for each category on an open data portal, a formal context is defined with a set of datasets that belong to a particular category and a set of tags used for describing those datasets. Therefore, the reduction of such formal context represents a reduction in the number of distinct tags appearing in the context by comparing their meanings. For example, if in the "Arts" category there are tags "paint", "pigment", and "rouge", the formal context reduction algorithm could mark these tags as synonyms and reduce the tags space by replacing those three tags with only one.

For that reason, in this research, we employ [25] algorithm that relies on the semantic similarity for reducing the number of distinct tag values, which results in the reduction of the formal context. At the beginning, all tags are converted into 300 dimensional vectors using GloVe (Global Vectors for Word Representation) [59] model which allows similarity comparison on the word level. The GloVe model used within this algorithm is trained on Common Crawl data, with 840 billion generated tokens and vocabulary containing 2.2 million entries. This model has a broad vocabulary and heterogeneous linguistic coverage, and allows similarity comparison on the word level.

Within this algorithm, tags are processed as they appear in the formal context. For each pair of tags within one formal context a cosine similarity is calculated using the following definition:

Definition 4 ([25]). *Similarity S_T between two tags, T_X and T_Y, is defined as the largest similarity between any two words W_X and W_Y, whereas W_X and W_Y represent parts of T_X and T_Y, respectively.*

After computing tag similarity results for all tag pairs, tags whose similarity values exceed a predefined threshold are considered highly similar. Consequently, for each tag in the formal context C, a corresponding group of similar tags $L_{ST} = [s_1, s_2, \ldots, s_k]$ is defined. If the list of similar tags contains more than two tag values, the tag is identified as a reduction candidate tag.

For each formal context C, reduced formal context C_R is computed using the following algorithm [25]:

For each pair (T_I, L_{STI}) from C:

a. Find all intersection tag sets $L_{INT} = [u_1, u_2, \ldots, u_j]$ between L_{STI} and all other L_{ST} sets in category C;
b. Among all L_{INT} tag sets, find tag set L_{SUB} which has the maximal number of elements and is equal to L_{STJ} coupled with tag T_J from C;
   - if L_{SUB} cannot be found, iteration ends;
   - if more than one tag has an equal maximal number of elements, the first one is taken;
c. Since L_{SUB} is a subset of L_{STI}, within L_{STI} replace L_{SUB} with single tag—T_J;
d. Repeat steps from a to c until iteration ends.

The proposed algorithm operates through an iterative process of tag list comparison and substitution, designed to optimize the representation of data within a formal context for concept lattice construction. Specifically, the algorithm begins by generating lists of similar tags based on predefined similarity criteria. These lists are then compared against existing tag lists within the formal context to determine the largest common subset (i.e., the maximal matching subset) that fully corresponds to the tags associated with an existing tag. The identified largest subset of tags is replaced by its corresponding tag, and this

step is repeated until all subsets are found and replaced by the appropriate tags. In cases where multiple tags correspond to subsets of equal maximum size, the first identified tag is selected. As the result of the performed algorithm the size of the formal context used for generating concept lattice is reduced.

3.2. Methodology

From the formal context reduction algorithm it can be concluded that the threshold value depends on the embedding model used for generating word vectors, as the semantic similarity assigned to the same word pair may vary across different models. Furthermore, it can be concluded that the choice of threshold value directly influences the extent of the reduction. A lower threshold value increases the number of replacement candidates, thereby enhancing the reduction potential. Accordingly, it will result with a larger number of replaced tags and smaller number of attributes in the formal context, and thereby simpler concept lattice. Consequently, higher threshold results in the replacement of only highly similar tags, thereby limiting the overall reduction. Finding an appropriate reduction threshold can be challenging. The authors of [60] reported the effects of the word embedding model dimensionality on similarity threshold value by conducting the experiments in different vector dimensions. They suggested that a threshold value of 0.7 would be appropriate for the 300-dimension model. Furthermore, the authors of [61] analyzed open government metadata and manually checked tag pairs with semantic similarity between 0.7 and 0.8. They concluded that threshold value set to 0.7 may harm the expressivity of formal context. The authors gave example for tags "water transport", "pipeline transportation occurrences", where setting reduction threshold to 0.7 would result in substitution of these tags with more general one, e.g., water transport tag. Therefore, within this research among 0.7, we will analyze and more restricted threshold values 0.8 and 0.9. Our goal is to investigate the number of reduced tags depending on the reduction threshold, as well as its influence on the usability of the knowledge base over time.

For the purpose of the long-term usability analysis, we will combine the formal context reduction algorithm with a FCA-based dataset categorization algorithm presented in [12]. Since different open data portals have different dataset categorization approaches in terms of the number of categories that can be assigned to a single dataset, within this research, Ireland's and Canada's open data portals will be investigated. Ireland's open data portal was chosen as a representative of the portals where datasets are assigned only one category, and Canada's open data portal was chosen as a representative of the portals where datasets are assigned multiple categories. Since the number of assigned categories can have a direct impact on the knowledge base usability and categorization algorithm performance, this portal selection ensures that in the evolving metadata environment both categorization approaches are analyzed.

The long-term usability analysis of the reduced concept lattices will be performed on the 5-year period. To enable an over time analysis, six distinct datasets were utilized for the construction of reduced concept lattices corresponding to the open data portals (ODPs) of Ireland and Canada:

- IrelandDS 2020 and CanadaDS 2020—sets containing all datasets that were available on the Ireland's ODP and Canada's ODP in 2020 respectively;
- IrelandDS 2021 and CanadaDS 2021—sets containing all datasets that were available on the Ireland's ODP and Canada's ODP in 2021 respectively;
- IrelandDS 2023 and CanadaDS 2023—sets containing all datasets that were available on the Ireland's ODP and Canada's ODP in 2023 respectively.

For each of these six datasets, reduced concept lattices were generated using Formal Concept Analysis with reduction thresholds set to 0.7, 0.8, and 0.9. Each threshold value

yields a distinct set of reduced lattices, e.g., a corresponding knowledge base, resulting in a total of 18 knowledge bases across the six input datasets. For testing the usability of these knowledge bases, we will perform categorization using six test datasets for each open data portal, where test datasets contain all new data that became available in the following time periods: 2021–2020, 2023–2020, 2025–2020, 2023–2021, 2025–2021, and 2023–2025. To ensure that this analysis reflects the real-world growth of new datasets, no duplicate removal or category balancing was performed on the test data.

4. Dataset and Formal Context Reduction

For the purpose of this research we have chosen two open data portals with different dataset categorization approaches—Ireland ODP and Canada ODP. Ireland open data portal organizes its data into 14 categories and in the period from 2020 to 2023 the upward trend in the number of datasets and tags used for their description was noted. In 2020, the total number of datasets on Ireland ODP was 10,151; in 2021 this number increased to 12,949, in 2023 it was 17,844; and in 2025 there were 20,325 datasets on the portal. The total number of tags more than tripled in this period; in 2020, there were 5070 tags; in 2021, there were 11,778 tags; in 2023, there were 15,593 tags; and the number of tags in 2025 was 17,231. Furthermore, it was noticed that in the period from 2020 to 2021 there were 7984 datasets with new identification numbers, resulting in significant changes in the distribution of datasets by category and usage of tags. For example, the number of datasets was substantially reduced in the *Economy* and *Society* categories, and notably increased in the *Government* category. The detailed distribution of the number of available datasets by category for 2020, 2021, and 2023 is presented in Table 2 (ND column).

On the datasets from 2020, 2021, and 2023 reduction was performed using reduction thresholds 0.7, 0.8, and 0.9, resulting in three reduced formal contexts per year. The percentage of reduced tags per category for each threshold and each dataset is presented in Table 2 (RT [%] 0.7, RT [%] 0.8, and RT [%] 0.9). Based on generated reduced contexts, concept lattices were produces for each category on the portal. The structure of the generated concept lattices, e.g., the number of generated nodes per lattice for each reduction threshold is presented in Table 2 (NN 0.7, NN 0.8, and NN 0.9).

From the presented results it can be noted that the *Crime* category did not have any reduction in the number of tags for any of the years and threshold values, due to the specifics of the used tag values for describing datasets. Furthermore, the *Towns* category did not have any reductions in 2021, while in 2020 within this category there was reduction only for 0.7 threshold. Moreover, it can be noted that for the 0.9 threshold, there were multiple categories, like *Agriculture*, *Economy*, *Housing*, etc., whereas for multiple datasets the reduction was equal to 0. It can be concluded that, overall, reduction percentage was below 1% for most of the datasets in most of the categories, and only in two cases the reduction percentage was between 3% and 4%. As expected, the reduction percentages for threshold set to 0.8 were higher, and besides the *Crime* category, only the *Towns* category had reduction equal to 0 in 2020 and 2021. The highest reduction percentages were obtained for threshold 0.7, where in 16 cases reduction was higher than 10%, and in 2023 for the *Society* category it was higher than 20%.

Furthermore, from the presented number of nodes in concept lattices, it can be noticed that some of the concept lattices have had significantly different structures between the years. The differences are due to the sizable changes that occurred in this period of time, in terms of available datasets within categories and tags and combination of tags used within them.

Table 2. Ireland ODP—reduction information for formal contexts in 2020, 2021, and 2023 for reduction thresholds of 0.7, 0.8, and 0.9 and corresponding concept lattices.

Category	Year	ND	NT-BR	RT [%] 0.7	RT [%] 0.8	RT [%] 0.9	NN 0.7	NN 0.8	NN 0.9
Agriculture	2020	123	105	14.29	8.57	0.00	33	35	33
	2021	131	270	7.78	4.07	0.00	152	154	151
	2023	155	380	6.05	2.63	0.00	196	198	195
Arts	2020	135	297	7.07	2.36	1.01	250	251	260
	2021	65	215	10.70	3.72	0.47	129	133	152
	2023	97	243	14.81	6.58	0.82	144	150	170
Crime	2020	8	13	0.00	0.00	0.00	6	6	6
	2021	93	100	0.00	0.00	0.00	96	96	96
	2023	382	407	0.00	0.00	0.00	407	407	407
Economy	2020	1149	202	9.41	5.45	0.00	123	123	123
	2021	321	416	6.25	4.33	0.00	324	326	324
	2023	425	561	8.91	4.63	0.18	589	635	831
Education and Sport	2020	182	154	9.74	3.25	1.30	83	86	87
	2021	225	364	5.49	1.92	0.55	253	259	263
	2023	262	463	7.56	2.81	0.22	370	393	426
Energy	2020	123	243	12.35	7.00	0.82	240	297	313
	2021	125	243	9.88	5.35	0.41	202	224	284
	2023	189	484	18.80	8.26	1.03	395	434	512
Environment	2020	3098	1712	15.07	8.00	0.88	3678	3996	4259
	2021	3444	2144	13.34	7.98	0.93	4361	5080	5139
	2023	3502	2243	13.60	8.02	1.03	4748	5438	5471
Government	2020	495	1409	14.27	6.96	3.55	922	977	1035
	2021	5775	6772	3.03	1.48	0.72	6252	6257	6267
	2023	7486	9174	2.43	1.28	0.58	8673	8720	8752
Health	2020	1416	427	14.29	4.92	0.23	220	219	211
	2021	1817	971	5.66	2.06	0.10	780	773	764
	2023	1069	1125	9.51	3.91	0.36	884	887	946
Housing	2020	424	347	6.92	3.46	0.00	466	490	502
	2021	345	388	5.67	3.87	0.00	519	553	558
	2023	727	717	8.37	3.77	0.42	961	1030	1068
Science	2020	329	387	10.34	3.10	0.52	356	358	302
	2021	127	376	12.50	5.59	0.27	444	450	362
	2023	113	421	11.88	4.99	0.48	469	476	388
Society	2020	2203	459	4.14	3.70	0.00	373	376	373
	2021	86	195	18.97	13.33	1.54	82	90	88
	2023	137	363	20.66	11.02	1.38	205	233	272
Towns	2020	50	52	1.92	0.00	0.00	31	31	31
	2021	18	34	0.00	0.00	0.00	22	22	22
	2023	60	162	13.58	5.56	3.09	87	92	92
Transport	2020	411	330	6.36	3.33	0.61	243	256	263
	2021	370	542	7.01	4.06	0.55	458	483	495
	2023	396	733	7.91	4.37	0.82	638	743	957

ND—number of datasets, NT-BR—number of tags before reduction; RT [%] 0.7—tag reduction percentage with reduction threshold set to 0.7; RT [%] 0.8—tag reduction percentage with reduction threshold set to 0.8; RT [%] 0.9—tag reduction percentage with reduction threshold set to 0.9; NN 0.7—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.7; NN 0.8—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.8; NN 0.9—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.9.

The Canada open data portal organizes its data into 19 categories, where one dataset may belong to multiple categories. Furthermore, within this portal in the period from 2020 to 2025 there were significant changes made, which is particularly important for the analysis

in this research. In 2020, there were 85,546 available datasets on the portal. However, in 2021, due to the data consolidation, the number of available datasets was 31,270. From that point, an upward trend in the number of datasets was detected, and in 2023 there were 38,956 datasets, and in 2025 the number of dataset was equal to 42,051. The analyses of used tags showed that the number of tags increased constantly throughout the analyzed period; in 2020 the number of tags was 17,225, in 2021 it was equal to 27,551, in 2023 it was equal to 38,943, and in 2025 there were 43,053 used tags. The detailed distribution of the number of available datasets by category for 2020, 2021, and 2023 is presented in Table 3 (ND column).

Table 3. Canada ODP—reduction information for formal contexts in 2020, 2021, and 2023 for reduction threshold of 0.7, 0.8 and 0.9 and corresponding concept lattices.

Category	Year	ND	NT-BR	RT [%] 0.7	RT [%] 0.8	RT [%] 0.9	NN 0.7	NN 0.8	NN 0.9
agriculture	2020	712	714	20.17	8.26	0.14	446	528	528
	2021	858	936	19.98	10.26	0.53	578	689	675
	2023	1110	1271	17.78	8.10	0.63	867	1036	1008
arts_music_literature	2020	28	149	15.44	6.04	0.00	36	37	35
	2021	43	216	14.81	5.56	0.00	59	58	56
	2023	143	530	16.79	8.30	0.00	214	210	207
economics_and_industry	2020	66,465	3365	24.99	11.71	1.58	2406	3056	3313
	2021	7751	4277	24.06	11.43	1.75	2935	3780	4153
	2023	9978	5944	22.56	10.73	1.58	4400	5477	6003
education_and_training	2020	290	541	24.03	9.80	1.48	199	264	264
	2021	465	723	23.24	8.99	1.52	305	384	401
	2023	767	1236	22.49	10.03	2.51	652	874	943
form_descriptors	2020	69,475	2203	19.56	11.12	2.41	18,895	12,594	15,753
	2021	11,721	4763	15.89	9.66	1.93	17,697	16,574	18,219
	2023	12,005	6288	14.09	7.49	1.15	16,838	16,174	18,594
government_and_politics	2020	64,789	2904	23.73	11.74	1.72	2051	2255	2162
	2021	6149	4490	22.32	10.29	1.27	2983	3296	3310
	2023	7145	6841	21.90	10.57	1.39	4724	5335	5375
health_and_safety	2020	1669	4086	19.53	9.37	1.74	2100	2366	2508
	2021	3288	10,805	22.36	11.07	2.09	8221	9596	10,704
	2023	5150	15,299	23.31	11.56	2.37	10,847	12,400	13,671
history_and_archaeology	2020	102	191	11.52	1.05	0.00	92	87	86
	2021	13	89	12.36	2.25	0.00	17	15	14
	2023	17	114	13.16	5.26	0.00	20	20	18
information_ and_communications	2020	678	1010	19.31	9.60	0.89	579	649	663
	2021	899	1326	18.63	9.73	0.75	858	957	971
	2023	1004	1721	18.83	8.77	0.58	952	1074	1105
labour	2020	691	750	28.00	9.87	1.73	361	508	529
	2021	806	907	26.35	9.04	1.76	540	732	740
	2023	1126	1292	24.07	9.83	2.09	862	1206	1300
language_and_linguistics	2020	52	176	18.75	5.68	0.00	74	71	71
	2021	91	279	15.05	5.38	0.00	120	139	129
	2023	234	354	16.38	6.78	0.56	205	279	250
law	2020	522	488	22.75	9.84	0.82	280	342	322
	2021	735	757	19.42	8.45	0.92	501	618	604
	2023	1004	1442	18.10	7.77	0.76	1031	1189	1206
military	2020	65	219	5.94	1.37	0.00	89	94	95
	2021	103	328	10.06	1.83	0.00	133	133	135
	2023	134	398	13.07	4.02	1.01	189	196	190

Table 3. *Cont.*

Category	Year	ND	NT-BR	RT [%] 0.7	RT [%] 0.8	RT [%] 0.9	NN 0.7	NN 0.8	NN 0.9
nature_and_environment	2020	73,139	7786	18.56	10.29	2.48	34,313	37,355	38,481
	2021	14,468	10,156	17.26	9.39	1.95	28,138	22,978	25,380
	2023	15,651	13,398	15.26	8.28	1.64	24,542	24,910	28,279
persons	2020	2486	768	20.44	9.77	0.78	765	885	902
	2021	2614	1050	19.05	9.14	0.76	964	1133	1135
	2023	2970	1340	19.70	8.73	0.97	1368	1676	1831
processes	2020	121	375	19.73	8.27	0.80	154	164	152
	2021	158	526	18.63	7.98	0.57	217	229	218
	2023	200	655	20.92	7.79	0.31	286	307	296
science_and_technology	2020	7603	3333	19.77	10.59	2.49	19,264	13,114	16,371
	2021	10,110	6424	17.12	9.26	1.82	20,597	18,200	19,948
	2023	10,758	8606	14.36	7.55	1.31	18,220	18,570	21,146
society_and_culture	2020	1620	1801	21.43	10.16	1.50	1428	1567	1612
	2021	2053	2516	18.52	8.66	1.31	2013	2303	2331
	2023	2978	4349	18.85	9.77	1.75	2963	3354	3511
transport	2020	944	909	16.61	8.03	1.65	549	566	535
	2021	1128	1218	15.19	6.73	0.90	744	721	708
	2023	1467	1744	15.37	7.97	1.15	891	955	967

ND—number of datasets; NT-BR—number of tags before reduction; RT [%] 0.7—tag reduction percentage with reduction threshold set to 0.7; RT [%] 0.8—tag reduction percentage with reduction threshold set to 0.8; RT [%] 0.9—tag reduction percentage with reduction threshold set to 0.9; NN 0.7—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.7; NN 0.8—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.8; NN 0.9—number of nodes in the concept lattice generated based on the reduced formal context produced with a reduction threshold set to 0.9.

On the datasets from 2020, 2021, and 2023 reduction was performed using reduction thresholds of 0.7, 0.8, and 0.9, resulting in three reduced formal contexts per year. The percentage of reduced tags per category for each threshold and each dataset is presented in Table 3 (RT [%] 0.7, RT [%] 0.8, and RT [%] 0.9). Based on generated reduced contexts concept lattices were produced for each category on the portal. The structure of the generated concept lattices, e.g., the number of generated nodes per lattice for each reduction threshold, is presented in Table 3 (NN 0.7, NN 0.8, and NN 0.9).

From the presented reduction information it can be noted that for the threshold set to 0.9, categories *arts_music_literature* and *history_and_archaeology* did not have any reduction in the number of tags for any of the datasets. Furthermore, categories *language_and_linguistics* and *military* did not have any reduction in the number of tags in 2020 and 2021 formal contexts. Moreover in 2023 dataset, for category *language_and_linguistics* category reduction percentage was equal to 0.56%, and for the *military* category, the reduction percentage was 1.01%. Furthermore, for the threshold set to 0.9, the highest reduction percentages were 2.49% in 2020 for the *science_and_technology* category, and 2.48% in 2020 for the *nature_and_environment* category, and for all other formal context reductions, the tag reduction percentage was below 2%. The analysis of reductions with the threshold set to 0.8 showed that the tag reduction percentage was below 12% for all formal contexts, and the lowest reduction percentage was for the *health_and_safety* and *military* categories. As expected, the highest reduction percentages were recorded for the threshold set to 0.7, where, except for the *military* category in 2020, when reduction was equal to 5.94%, for all other formal context reductions, the tag reduction percentage was between 10% and 28%.

Furthermore, from the presented number of nodes in concept lattices, it can be noticed that some of the concept lattices have had significantly different structures between the years. The observed differences can be attributed to the dataset consolidation that was

made on the portal after the 2020 dataset was retrieved and a subsequent increase in the number of datasets within categories, as well as tags, and the combination of tags applied across them.

5. Concept Lattice Usability Analysis—A Categorization Perspective

The sets of concept lattices, e.g., knowledge bases, presented in the previous section of this paper were used for over time reduced concept lattice usability analyses within six test cases. For both portals, knowledge bases created on 2020 datasets were used for the categorization of all new datasets that were available in 2021 (2020–2021 test datasets), that is, datasets from 2021 whose identifiers were not in the 2020 dataset. Furthermore, this knowledge bases were used for the categorization of all new datasets that were available in 2023 (2020–2023 test datasets), that is, datasets from 2023 whose identifiers were not in the 2020 dataset. Lastly, this knowledge bases were used for the categorization of all new datasets that were available in 2025 (2020–2025 test datasets), that is, datasets from 2025 whose identifiers were not in the 2020 dataset. Furthermore, for both portals knowledge bases created on 2021 datasets were used for the categorization of all new datasets that were available in 2023 (2021–2023 test datasets) and 2025 (2021–2025 test datasets). Finally, for both portals, knowledge bases created on 2023 datasets were used for categorization of all new datasets that were available in 2025 (2020–2025 test datasets). Due to the way these datasets were made, both the distribution of the number of instances across the test cases and distribution of test datasets across categories differ.

Since open data portals may have different approaches to dataset categorization in terms of the number of assigned categories, the durability analysis was performed on Ireland OPD as a representative of portals that use one category per dataset approach, and Canada ODP as a representative of portals that assign multiple categories per dataset. Therefore, the analysis is performed per categorization approach and results of the durability analysis are presented in the following part of this research.

5.1. Single-Category Categorization Results

Since datasets available on the Ireland open data portal are assigned to at most one category, whereas the applied categorization algorithm is designed to recommend multiple categories when the relevant criteria are met, the categorization results are divided into three distinct groups:

- FM group—corresponding to fully categorized datasets, meaning that each dataset was assigned the correct category and no additional categories were suggested by the algorithm;
- FME group—corresponding to datasets that were assigned the correct category and at least one additional category;
- MM group—mismatched datasets—datasets that were assigned incorrect category.

However, since both FM and FME groups assign correct category to a dataset, for the purpose of this analysis datasets belonging to these groups will be considered correctly categorized. The total number of instances within test datasets for this analysis and overall categorization results for all reduction thresholds are presented in Table 4.

From the presented results it can be noticed that for 2023–2025 test the percentage of successfully categorized datasets was between 82.55% and 83.02% depending on the reduction threshold. Furthermore in the 2021–2025 test the success rate depending on the reduction threshold was between 76.87% and 77.86%, while in the 2021–2023 test the overall successes depending on the reduction threshold was between 83.65% and 84.68%. In tests performed on the 2020 dataset, e.g., tests 2020–2021, 2020–2023 and 2020–2025, the success rates were lower comparing to other datasets, and were between 46.19% and 59.43%. This

is due to the dataset consolidation that was performed on the portal in that period of time. The chart with summarized categorization results is presented in Figure 2.

Table 4. Ireland ODP—categorization results.

Test	ND	Similarity	FM [%]	FME [%]	MM [%]
2020–2021	7984	0.7	54.01	3.49	42.50
		0.8	42.96	3.23	53.81
		0.9	43.78	3.13	53.09
2020–2023	12,979	0.7	52.25	7.17	40.57
		0.8	45.84	6.97	47.20
		0.9	46.24	6.84	46.92
2020–2025	15,612	0.7	48.86	9.30	41.84
		0.8	45.70	7.05	47.25
		0.9	46.13	6.94	46.93
2021–2023	5541	0.7	70.06	13.59	16.35
		0.8	70.40	14.28	15.32
		0.9	71.20	13.34	15.47
2021–2025	8202	0.7	62.58	14.29	23.13
		0.8	66.00	11.62	22.38
		0.9	66.76	11.09	22.14
2023–2025	3627	0.7	71.22	11.33	17.45
		0.8	74.41	8.13	17.45
		0.9	72.21	10.81	16.98

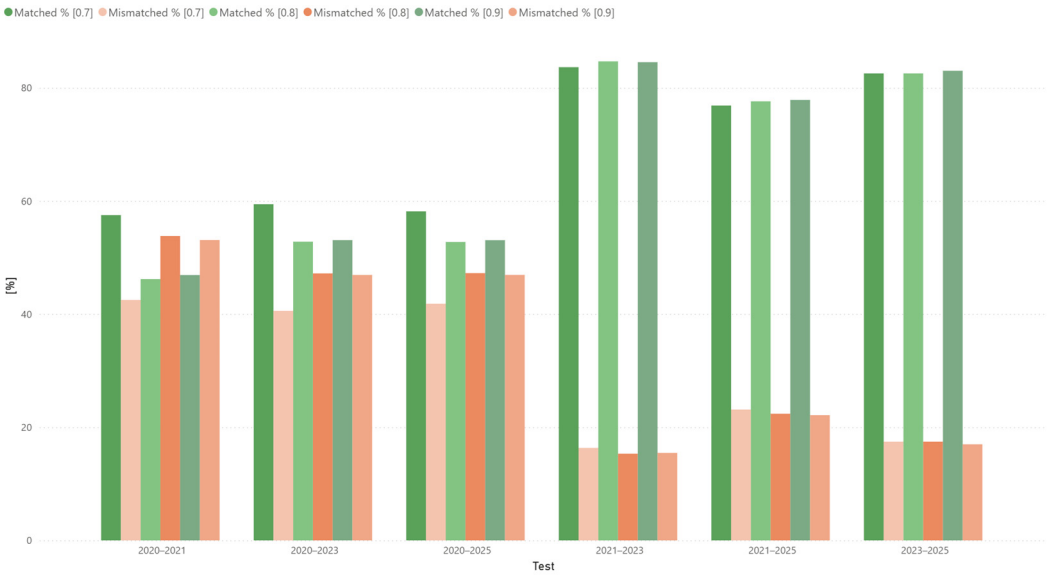


Figure 2. Ireland ODP—categorization summary.

If observed category-wise across all reduction thresholds, categorization results range from exceptional to acceptable. For example, for the *Transport* category exceptional categorization results were obtained, with 99.61% of correctly categorized datasets for all reduction thresholds in the 2020–2021 test, 98.01% of correctly categorized datasets for all reduction thresholds in the 2020–2023 test, and 97.39% of correctly categorized datasets for all reduction thresholds in the 2020–2025 test. Furthermore, for the *Crime* category the success rate was 99.69% for all reduction thresholds in the 2021–2023 test. Moreover, in the 2021–2023 test, there were five categories with over 90% of successfully categorized

datasets for all reduction thresholds, while in the 2023–2025 test there were six categories with success rate over 90%.

To uncover the reduced concept lattices usability to their full extent, in the remaining part of this section we will present the results of a deep analysis of categorization results for each pair of time period and category. The deeper analysis showed that the performance of knowledge bases created on the same dataset but with different thresholds had similar results in tests. The greatest difference in the results depending on the reduction threshold was recorded for the test performed on knowledge bases created on the 2020 dataset. For the 2020–2021 test, the best performance was achieved with a reduction threshold set to 0.7, yielding a success rate of 57.5%. When the threshold was increased to 0.8 and 0.9, the percentage of correctly categorized datasets was similar for both thresholds and more than 10% lower than the result obtained with the 0.7 threshold. For the 2020–2023 test, the knowledge base created with a threshold of 0.7 yielded a success rate of 59.43%. For the thresholds of 0.8 and 0.9, the success rate was worse compared to the 0.7 threshold, by 6.63 and 6.35 percentage points, respectively. For the 2020–2025 test, a threshold of 0.7 with 58.16% of correctly categorized datasets had better results—5.09 and 5.41 percentage points higher compared to thresholds set at 0.9 and 0.8, respectively. However, for tests performed on the 2021 and 2023 datasets, the test results were within 1% for all reduction thresholds.

Furthermore, the category-level analysis of the obtained results showed that for most of the categories, the success rates for different reduction thresholds had a marginal difference. For tests conducted on the 2020 knowledge bases, the *Health* category was the only category in which knowledge bases generated after applying the 0.7 reduction threshold achieved a significantly higher percentage of successful categorizations compared to those generated after applying the 0.8 and 0.9 reduction thresholds. For the knowledge base generated after applying the 0.7 reduction threshold, more that 97% of datasets belonging to the *Health* category were correctly categorized. In contrast, for the 2020–2021 test, for the thresholds of 0.8 and 0.9, these percentages were below 50%, while for the 2021–2023 and 2021–2025 tests these percentages were below 65%. Furthermore, for tests conducted on the 2020 knowledge bases, the *Energy* category was the only category in which knowledge bases generated after applying the 0.7 reduction threshold achieved a lower percentage of successful categorizations, compared to those generated after applying the 0.8 and 0.9 reduction thresholds. For the 2020–2021 test this difference was 22 percentage points, while for the 2020–2021 and 2020–2025 tests this difference was below 14 percentage points. A deeper analysis showed that the success rates of these two categories were largely responsible for the overall difference in the success rates of these reduced knowledge base categorizations.

Next, it was observed that within the 2021–2023 test, only within the *Economy* and *Science* categories the difference between success rates for different thresholds was 10 and 15 percentage points, respectively. For other categories the results were the same for all thresholds, or were within a few percentage points. It was further observed that within the 2021–2025 test, the *Science* and *Transport* categories had success rate differences of up to 13 percentage points for different thresholds, while for other categories, the results were the same or comparable. Lastly, it was noticed that within the 2023–2025 test, the *Economy* category was the only category of up to 10 percentage points' differences in success rate, while for other categories success rates were comparable or the same for different reduction thresholds.

Additionally, it was concluded that in the 2020–2021 test, the *Transport* category had 99.61% of correctly categorized datasets for all reduction thresholds, the *Arts* category had more than 88.24% of correctly categorized datasets for all reduction thresholds, while the *Energy* category had 94.92% of correctly categorized datasets for the thresholds set to

0.8 and 0.9. Within this test, the lowest percentage of successfully categorized datasets was obtained for the *Crime* category where the percentage was 1.1% for all reduction thresholds. Furthermore, for the 2020–2023 test, it was concluded that the *Transport* and *Arts* categories obtained the best categorization results. The *Transport* category had 98.01% of all datasets successfully categorized for all reduction thresholds, while for the *Arts* category this percentage was above 88%. Also in this test, the *Crime* category obtained the worst categorization results, with 0.49% of the test being successfully categorized for all reduction thresholds. In the 2020–2025 test, the *Transport* category had 97.39% of datasets successfully categorized for all reduction thresholds, while the Arts and Environment categories had more than 82% of datasets successfully categorized for all reduction thresholds. Next, in the 2021–2023 test, five categories had all datasets successfully categorized in more than 90% for all reduction thresholds, with one of them being the *Crime* category. Furthermore, in the 2021–2025 test, this number of categories was equal to two for all thresholds, while in the 2023–2025 test there were five such categories and one category with more than 88% of success rate for all thresholds. It is important to emphasize that in the 2020–2025, 2021–2025, and 2023–2025 test datasets there where datasets assigned the *Justice* category. This category was defined on the portal after 2023, and therefore knowledge bases that are being analyzed within this research do not contain the concept lattice corresponding to this category. For that reason, all datasets assigned the *Justice* category were wrongly categorized. Finally, it can be concluded that knowledge bases created using a reduction threshold of 0.9 achieved similar or even better results for some categories comparing to other thresholds, which results in overall very similar results for the complete test.

5.2. Multiple-Category Categorization Results

Since datasets available on the Canada open data portal can be assigned multiple categories, the categorization results are divided into five distinct groups:

- FM group—corresponding to fully categorized datasets, meaning that each dataset was assigned the correct category and no additional categories were suggested by the algorithm;
- FME group—corresponding to datasets that were assigned the correct category and at least one additional category;
- PM group—corresponding to datasets that were not assigned all correct categories and no additional categories were suggested by the algorithm;
- PME group—corresponding to datasets that were not assigned all correct categories and were suggested at least one additional category;
- MM group—mismatched datasets—datasets that were assigned an incorrect category.

However, since both FM and FME groups assign all correct categories to a dataset, and PM and PME assign at least part of the correct categories, for the purpose of this analyses datasets belonging to these groups will be considered successfully categorized. The total number of instances within test datasets for this analyses and overall categorization results for all reduction thresholds are presented in Table 5.

From the presented categorization results it can be noticed that success rate for 2023–2025 tests was from 80.37% to 81.88%. The success rate for tests conducted using the 2021 dataset was above 80% for all tests. More specifically, the success rate ranged from 81.95% to 84.26% for the 2021–2023 tests, and from 80.78% to 83.28% for the 2021–2025 tests. The tests conducted using the 2020 dataset produced the lowest success rates overall. Specifically, the success rate ranged from 75.26% to 76.77% for the 2020–2021 tests, from 77.86% to 79.15% for the 2020–2023 tests, and from 75.52% to 77.36% for the 2020–2025 tests. However, it should be noted that Canada ODP also had the dataset consolidation after 2020, which had a direct impact on the lower categorization success. Furthermore, from the presented

results it can be noticed that on the test level success rate differed by 2.5 percentage points for the corresponding reduced knowledge bases with different reduction thresholds. The category-wise analysis showed that some of the categories, like *nature_and_environment*, had exceptional categorization results for all tests with success rate up to 94.67%. The chart with summarized categorization results is presented in Figure 3.

Table 5. Canada ODP—categorization results.

Test	ND	Similarity	FM [%]	FME [%]	PM [%]	PME [%]	MM [%]
2020–2021	7499	0.7	29.20	26.83	12.40	6.83	24.74
		0.8	29.46	27.38	11.71	6.97	24.48
		0.9	32.59	26.38	11.20	6.60	23.23
2020–2023	17,532	0.7	26.94	25.99	18.99	5.94	22.14
		0.8	28.03	25.21	18.12	5.78	22.86
		0.9	31.03	24.49	18.11	5.52	20.85
2020–2025	20,795	0.7	27.47	25.30	17.79	5.10	24.33
		0.8	28.40	24.74	17.00	5.39	24.48
		0.9	31.80	23.40	16.93	5.23	22.64
2021–2023	10,746	0.7	46.37	20.79	10.07	4.72	18.05
		0.8	46.42	21.94	9.01	4.83	17.80
		0.9	51.29	19.59	8.50	4.89	15.74
2021–2025	13,947	0.7	43.94	20.79	11.44	4.60	19.22
		0.8	43.77	22.98	10.16	4.72	18.38
		0.9	48.62	20.12	9.94	4.60	16.72
2023–2025	3704	0.7	35.99	25.54	14.20	5.24	19.03
		0.8	34.72	26.54	13.96	5.16	19.63
		0.9	39.55	24.08	13.47	4.78	18.12

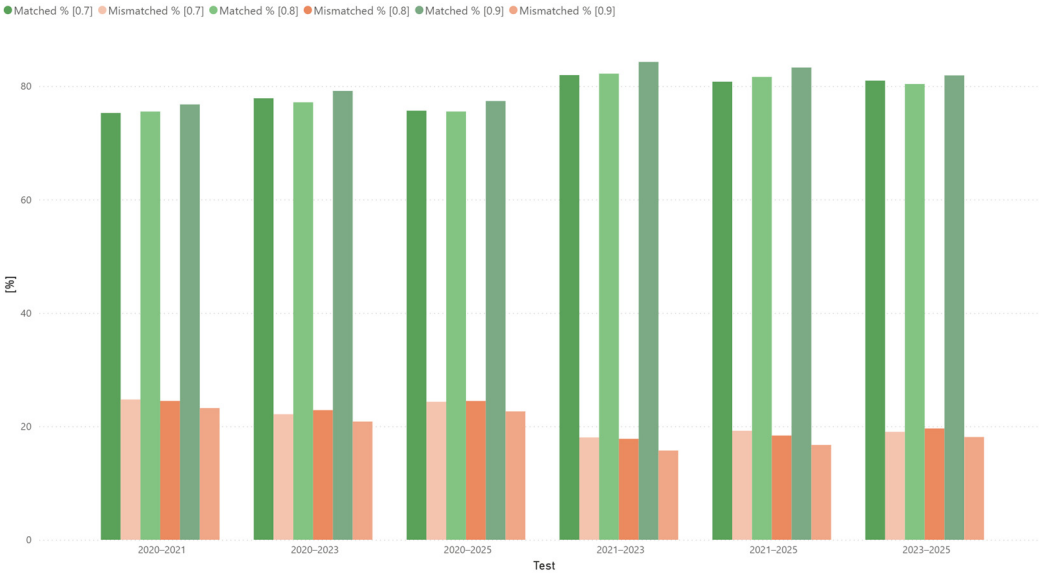


Figure 3. Canada ODP—categorization summary.

To uncover the influence of the reduction threshold on category-wise categorization results, in the remaining part of this section we present the results of a deeper category-level analysis for all tests. Category-level analysis was performed by evaluating all group results together and considering correct categories that were not proposed by the algorithm. This analysis showed that, for most categories within the same test, the categorization results

across different reduction thresholds differed by only a few percentage points. In the 2020–2021 test, only the *agriculture*, *education_and_training* and *law* categories had higher differences in percentage points, whereas knowledge base created with the 0.9 reduction threshold had the highest success rate. In this test, the success rate for the agriculture category was 73.96% at the 0.9 reduction threshold, 64.06% at the 0.8 reduction threshold, and 53.13% at the 0.7 reduction threshold. For the *education_and_training* category, success rates were 57.14%, 70.09%, and 72.88% for the reduction thresholds of 0.7, 0.8, and 0.9, respectively, while for the *law* category, success rates were 65.55%, 70.09%, and 77.26% for the reduction thresholds of 0.7, 0.8, and 0.9, respectively. In the 2020–2023 test, the *language_and_linguistics* and *law* categories were the only categories with higher differences in success rates depending on the reduction threshold, with the *language_and_linguistics* category having success rates of 44.57%, 32.61%, and 60.87% for the reduction thresholds of 0.7, 0.8, and 0.9 respectively, and the *law* category having success rates equal to 46.85%, 50.43%, and 60.17% for the reduction thresholds of 0.7, 0.8, and 0.9 respectively. These categories were also the only categories with higher differences in success rates in the 2020–2025 test. Within this test, the *language_and_linguistics* category had success rates of 42.72%, 31.07%, and 60.68% for the reduction thresholds of 0.7, 0.8, and 0.9, respectively, while the *law* category had success rates equal to 36.99%, 40.53%, and 51.01% for the reduction thresholds of 0.7, 0.8, and 0.9, respectively.

Furthermore, in the 2021–2023 test, the *labour* and *persons* categories were the only categories with greater differences in success rate percentage points, with success rates equal to 46.09%, 65.51%, and 64.93% for the *labour* category, and 60.14%, 67.12%, and 75.23% for the *persons* category for the reduction thresholds of 0.7, 0.8, and 0.9 respectively. These categories were also the only categories with higher differences in success rates in the 2021–2025 test, with success rates equal to 50.62%, 67.04%, and 67.41% for the *labour* category, and 53.62%, 61.67%, and 66.99% for the *persons* category, for the reduction thresholds of 0.7, 0.8, and 0.9 respectively. Across the 2023–2025 tests, four categories stood out with success rate differences exceeding 14 percentage points for different reduction thresholds: *arts_music_literature*, *labour, military*, and *society_and_culture*. Within the *arts_music_literature* category the success rates were 28.57%, 42.86%, and 42.86%, while for the *labour* category success rates were 56.85%, 66.50%, and 72.08% for the reduction thresholds of 0.7, 0.8, and 0.9 respectively. Moreover, for the *military* category the success rates were 50%, 41.67%, and 58.33%, while for the *society_and_culture* category the success rates were 38.70%, 53.95%, and 51.69% for the reduction thresholds of 0.7, 0.8, and 0.9 respectively. Finally, it was observed that across all conducted tests, the knowledge base created using a reduction threshold of 0.9 achieved the highest success rate, or was tied for the highest, in at least 12 categories per test compared to knowledge bases created using other reduction thresholds. In the 2021–2023 test this was the case for all categories within the knowledge base.

Additionally, the results showed that the *nature_and_environment* category consistently achieved the highest success rate across all reduction thresholds in tests conducted on the 2020 datasets, exceeding 80% in the 2020–2021 test, 87% in the 2020–2023 test, and 86% in the 2020–2025 test. Moreover, in the 2021–2023 test, the *health_and_safety* and *nature_and_environment* categories were the categories with the highest success rates, with success rates exceeding 87.95% for the *health_and_safety* category and 91% for the *nature_and_environment* category for all reduction thresholds. These categories achieved the greatest success rates in the 2021–2025 test, with success rates exceeding 88% for the *health_and_safety* category and 87% for the *nature_and_environment* category for all reduction thresholds. In the 2023–2025 test, the *language_and_linguistics* and *nature_and_environment* categories were the ones with the highest success rates, with success rates exceeding 82%

for the *language_and_linguistics* category and 81% for the *nature_and_environment* category for all reduction thresholds.

6. Discussion

The analysis performed in this research was focused on the strength of conceptualization and durability of reduced concept lattices in an evolving metadata environment. The concept of lattice durability was analyzed for the purpose of data categorization on open data portals. Due to different types of categorization approaches that can be applied on one portal, the analysis had two parts: analysis of lattice usability for single-category categorization and analysis of lattice usability for multiple-category categorization.

Based on the presented analysis results, it can be concluded that in the case of regular dataset growth on the portal, concept lattices have shown good long-term usability. This is supported by the results from both performed analyses, where the 2-year-old concept lattices showed a great categorization potential with only 15% to 19% of miscategorized datasets, depending on the test and threshold value. Furthermore, the 4-year-old concept lattices had fewer than 20% of miscategorized datasets in the multiple-category categorization test and fewer than 24% in the single-category categorization test. Lastly, the category-level analysis results, where multiple categories had more than 90% of correctly categorized datasets, support that concept lattices have a good long-term usability in case of regular dataset growth.

From the presented results, it can be concluded that the higher semantic reduction threshold did not lead to better long-term knowledge base usability in the single-category categorization analysis. Moreover, the results indicate that a threshold value of 0.7 achieved better performance than more restrictive thresholds for the 2020 knowledge bases. This suggests that the knowledge base that preserves a broader range of general concepts, due to the lower reduction threshold, demonstrated better categorization performance after data consolidation. Furthermore, the most general knowledge bases had very similar categorization results for knowledge bases created using 2021 and 2023 datasets. Therefore, it can be inferred that the 0.7 threshold provided better long-term knowledge base usability overall. However, in the case of multiple-category categorization analysis, the most restricted analyzed threshold, 0.9, showed the best results in all tests. Furthermore, at the category level, a threshold of 0.9 achieved better performance in most categories, with several exhibiting notably higher results compared to lower thresholds. This indicates that maintaining more specific terms in concept lattices can be valuable for multiple-category categorizations.

In terms of long-term usability, it can be concluded that in both cases, single- and multiple-category categorizations, data consolidations reduce the usability of concept lattices for the categorization task. Therefore, in the case of data consolidation on the portal, the recreation of concept lattices should be performed. Furthermore, based on the results obtained for the case of data consolidation, it can be established that even in the case of a large increase in new data from new sources, an update of a part of the knowledge base affected by data growth should be recreated. Moreover, it can be concluded that category name changes, as well as adding or removing categories, require knowledge base recreation.

If a knowledge base is to be recreated, the performance of reconstruction process is governed by a large number of mutually interdependent parameters, whereas the choice of FCA algorithm should be devoted additional attention. FCA algorithms can be broadly classified into four main categories, each exhibiting distinct characteristics that significantly influence reconstruction efficiency: batch algorithms (NextClosure, Close-by-One (CbO, FCbO), In-Close), incremental algorithms (AddIntent), parallel algorithms (PCbO, PFCbO, FPCbO), distributed algorithms (MRCbO, MRGanter, HyPar-FCA+). The implementations

of these algorithms are available across various libraries and tools, including FCALGS, conexp-clj, FcaKit, concepts, FCA4J, and MRGanter+, primarily developed in Java and Python. Consequently, the selection of programming language and technology stack determines platform compatibility and operating system support, thereby adding further complexity to performance evaluations. Moreover, parallel and distributed implementations require specific hardware resources or system components, where communication latency can substantially affect overall efficiency.

Since computational efficiency is infrastructure-dependent, knowledge base creation practitioners should rely on computational complexity of the algorithm they select but should also be aware of the context characteristics [62]. In terms of worst-case complexity, an algorithm for generating all formal concepts (or the full concept lattice) is considered optimal if it achieves polynomial time delay (between outputting successive concepts) and uses space linear in the total number of concepts (up to a polynomial factor of the input size) [63]. However, contexts that produce an exponentially large number of concepts (the theoretical worst case) are relatively rare in practice. Furthermore, real-world performance is heavily influenced by practical factors such as the typical dimensionality of the context, implementation details, and the operating system, which can prove more important than pure theoretical complexity for evaluating algorithms [62,63].

7. Conclusions

This study demonstrates that FCA-based concept lattices offer a viable and relatively stable approach for open data categorization, even across several years of metadata evolution. The results provide valuable insights for selecting appropriate reduction thresholds and determining when knowledge base reconstruction becomes necessary. The study examines the long-term usability and efficiency of semantically reduced concept lattices for the automatic categorization of datasets in evolving open data portals. Treating concept lattices as knowledge bases constructed via Formal Concept Analysis (FCA), the research evaluated their durability when confronted with temporal changes in metadata.

Two main aspects were addressed: (1) the impact of the semantic reduction threshold on the long-term usability and expressiveness of the generated concept lattices, and (2) the identification of metadata change indicators that necessitate knowledge base reconstruction. The results indicate that semantically reduced concept lattices maintain good long-term usability under regular dataset growth. Four-year-old knowledge bases achieved miscategorization rates below 24% in single-category classification and below 20% in multiple-category classification. At the category level, in multiple cases over 90% of datasets were correctly categorized in both single- and multiple-category scenarios. Furthermore, findings regarding the optimal reduction threshold differed by categorization type. In single-category tests, a moderate threshold of 0.7, which retains a wider range of general concepts, yielded superior long-term performance. In contrast, the highest threshold (0.9), preserving more specific terms, performed best in multiple-category classification. These outcomes suggest that the ideal level of semantic reduction depends on whether the portal employs single or multiple-category assignment.

Overall assessment of the experimental results demonstrate that semantically reduced concept lattices constructed using FCA exhibit significant long-term stability and generalization capability for the categorization of open datasets over time. Semantically reduced concept lattices maintain strong performance when classifying new datasets even across periods spanning three to five years, achieving success rates that generally exceed 80%. These results indicate good temporal robustness of the FCA-based approach. Importantly, the performance remained consistently high for both single-category classification and

multiple-category classification, suggesting that the method scales well across different categorization approaches.

The analysis further revealed that the choice of semantic reduction threshold has only a minor impact on overall categorization success, with performance differences typically not exceeding 2.5%. This indicates the robustness of the method and the possibility of using more reduced lattices without substantial loss in accuracy in most cases, while only in a few categories was this difference noticeable. Category-level evaluation showed particularly strong results in specific domains. The "nature_and_environment" category achieved up to 94.67% correctly categorized datasets across all test periods, demonstrating that FCA-based lattices perform exceptionally well in more structurally stable and less volatile domains.

A notable performance drop occurred when using the 2020 knowledge base. This decline is directly attributed to the major dataset consolidation and structural changes in metadata and category distribution that took place on both portals. These findings highlight that significant structural changes in metadata represent the primary challenge to the long-term usability of concept lattices.

In a subtle manner, this research demonstrates that the optimal timing for knowledge base recreation is not simple to determine as it depends on multiple factors that should be monitored independently, with their relative importance varying according to specific contexts, policies, and open data portal strategies. Our findings provide insights into the following parameters we consider crucial for this decision: structural changes in the data model; category count; data categorization approach (single- vs. multi-label), requiring distinct policies for FM, FME, PM, PME, and MM outcomes (beyond just MM, as new categories may render existing structures inadequate); dataset size and computational costs, influenced by hardware, concept lattice algorithms, and embedding models; estimated data growth, potentially altered by new thematic domains or consolidation; and labeling policy changes, which must be evaluated from both structural and semantic perspectives.

Overall, the findings indicate robust long-term usability of the reduced concept lattices, particularly when metadata evolves gradually, while confirming that major dataset consolidations represent a critical factor necessitating knowledge base reconstruction.

Author Contributions: Conceptualization, M.F.G. and M.B.; methodology, M.F.G. and M.B.; software, D.P. and M.F.G.; formal analysis, M.F.G.; data curation, M.F.G.; writing—original draft M.F.G. and M.B.; writing—review and editing, M.F.G., M.B. and L.S.; supervision, L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [grant number 451-03-34/2026-03/200102].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this research are open data and can be found at the following links: https://open.canada.ca/data/en/ (accessed on 17 April 2026) and https://data.gov.ie/ (accessed on 17 April 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

FCA	Formal Concept Analysis
OGD	open government data
ODP	open data portal

DKAN	Drupal Knowledge Archive Network
CKAN	Comprehensive Knowledge Archive Network
DCAT	Data Catalog vocabulary

References

1. Berends, J.; Carrara, W.; Radu, C. *The Economic Benefits of Open Data*; Analytical Report 9; European Commission: Brussels, Belgium, 2017.
2. Sylikiotis, N.; Boni, M.; Popa, R.; Barekov, N.; Petersen, M.H.; Reda, J.; Tamburrano, D. Proposal for a Directive of the European Parliament and of the Council on the Re-Use of Public Sector Information (Recast). EU Legislation in Progress. 2019. Available online: https://www.europarl.europa.eu/RegData/etudes/BRIE/2018/628312/EPRS_BRI(2018)628312_EN.pdf (accessed on 5 April 2026).
3. Davies, T.; Walker, S.B.; Rubinstei, M.; Perini, F. *The State of Open Data: Histories and Horizons*; African Minds: Cape Town, South Africa; IDRC: Ottawa, ON, Canada, 2019. [CrossRef]
4. Braunschweig, K.; Eberius, J.; Thiele, M.; Lehner, W. *The State of Open Data Limits of Current Open Data Platforms*; The Caribbean Open Institute: Kingston, Jamaica, 2012.
5. Reiche, K.J.; Höfig, E. Implementation of metadata quality metrics and application on public government data. In Proceedings of the 2013 IEEE 37th Annual Computer Software and Applications Conference Workshops, Kyoto, Japan, 22–26 July 2013; pp. 236–241. [CrossRef]
6. Zuiderwijk, A.; Volten, C.; Kroesen, M.; Gill, M. Motivation perspectives on opening up municipality data: Does municipality size matter? *Information* **2018**, *9*, 267. [CrossRef]
7. Gligorijević, M.F.; Bogdanović, M.; Stoimenov, L. Tracking metadata changes in the government open data portals. In *ICIST 2022 Proceedings*; Zdravković, M., Trajanović, M., Konjović, Z., Eds.; Information Society of Serbia-ISOS: Belgrade, Serbia, 2022; pp. 180–184, ISBN 978-86-85525-24-7.
8. Gligorijević, M.F.; Bogdanović, M.; Veljković, N.; Stoimenov, L. Open Data Categorization Based on Formal Concept Analysis. *IEEE Trans. Emerg. Top. Comput.* **2021**, *9*, 571–581. [CrossRef]
9. Ch, A.K. Concept lattice simplification in formal concept analysis using attribute clustering. *J. Ambient. Intell. Humaniz. Comput.* **2019**, *10*, 2327–2343. [CrossRef]
10. Kumar, C.A.; Srinivas, S. Concept lattice reduction using fuzzy K-means clustering. *Expert Syst. Appl.* **2010**, *37*, 2696–2704. [CrossRef]
11. Alwersh, M.; Kovács, L. Survey on attribute and concept reduction methods in formal concept analysis. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *30*, 366–387. [CrossRef]
12. Gligorijević, M.F.; Bogdanović, M.; Stoimenov, L. Determining category metadata in open data portals—An approach based on Formal Concept Analysis. In Proceedings of the 2024 32nd Telecommunications Forum (TELFOR), Belgrade, Serbia, 26–27 November 2024; pp. 1–8. [CrossRef]
13. Attard, J.; Orlandi, F.; Scerri, S.; Auer, S. A systematic review of open government data initiatives. *Gov. Inf. Q.* **2015**, *32*, 399–418. [CrossRef]
14. Henke, N.; Bughin, J.; Chui, M.; Manyika, J.; Saleh, T.; Wiseman, B.; Sethupathy, G. *The Age of Analytics: Competing in a Data-Driven World*; McKinsey & Company: New York, NY, USA, 2016.
15. Neumaier, S.; Umbrich, J.; Polleres, A. Automated quality assessment of metadata across open data portals. *J. Data Inf. Qual.* **2016**, *8*, 1–29. [CrossRef]
16. Milic, P.; Veljkovic, N.; Stoimenov, L. Comparative analysis of metadata models on e-government open data platforms. *IEEE Trans. Emerg. Top. Comput.* **2018**, *9*, 119–130. [CrossRef]
17. Data Catalog Vocabulary Version 3. Available online: https://www.w3.org/TR/vocab-dcat/ (accessed on 17 July 2025).
18. Umbrich, J.; Neumaier, S.; Polleres, A. Quality assessment & evolution of open data portals. In Proceedings of the IEEE International Conference on Open and Big Data, Rome, Italy, 24–26 August 2015; pp. 1–8.
19. Open Government Working Group: 8 Principles of Open Government Data. Available online: https://public.resource.org/8_principles.html (accessed on 5 April 2026).
20. Tygel, A.; Auer, S.; Debattista, J.; Orlandi, F.; Campos, M.L.M. Towards cleaning-up open data portals: A metadata reconciliation approach. In Proceedings of the 2016 IEEE Tenth International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 4–6 February 2016; pp. 71–78.
21. de Castro, B.P.C.; Rodrigues, H.F.; Lopes, G.R.; Campos, M.L.M. Semantic enrichment and exploration of open dataset tags. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (WebMedia '19), Rio de Janeiro, Brazil, 29 October–1 November 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 417–424. [CrossRef]

22. Salminen, J.; Yoganathan, V.; Corporan, J.; Jansen, B.J.; Jung, S.-G. Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *J. Bus. Res.* **2019**, *101*, 203–217. [CrossRef]
23. Rajamae-Soosaar, K.; Nikiforova, A. Exploring Estonia's Open Government Data Development as a Journey towards Excellence: Unveiling the Progress of Local Governments in Open Data Provision. In Proceedings of the 25th Annual International Conference on Digital Government Research, Taipei, Taiwan, 11–14 June 2024.
24. Křemen, P.; Nečaský, M. Improving discoverability of open government data with rich metadata descriptions using semantic government vocabulary. *J. Web Semant.* **2019**, *55*, 1–20. [CrossRef]
25. Bogdanović, M.; Veljković, N.; Gligorijević, M.F.; Puflović, D.; Stoimenov, L. On revealing shared conceptualization among open datasets. *J. Web Semant.* **2021**, *66*, 100624. [CrossRef]
26. Wille, R. Restructuring lattice theory: An approach based on hierarchies of concepts. In *Formal Concept Analysis*; Lecture Notes in Computer Science; Ferré, S., Rudolph, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5548.
27. Ganter, B.; Stumme, G. *Formal Concept Analysis: Methods and Applications in Computer Science*; Technical Report; Otto–von–Guericke–Universitat Magdeburg: Magdeburg, Germany, 2003.
28. Poelmans, J.; Elzinga, P.; Viaene, S.; Dedene, G. Formal concept analysis in knowledge discovery: A survey. In *Proceedings of the International Conference on Conceptual Structures*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 139–153.
29. Poelmans, J.; Kuznetsov, S.O.; Ignatov, D.I.; Dedene, D. Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert Syst. Appl.* **2013**, *40*, 6601–6623. [CrossRef]
30. Chekol, M.W.; Napoli, A. An FCA framework for knowledge discovery in SPARQL query answers. In Proceedings of the 12th International Semantic Web Conference, Sydney, NSW, Australia, 21–25 October 2013.
31. Valtchev, P.; Missaoui, R.; Godin, R. Formal concept analysis for knowledge discovery and data mining: The new challenges. In *Proceedings of the International Conference on Formal Concept Analysis*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 352–371.
32. Alam, M.; Buzmakov, A.; Codocedo, V.; Napoli, A. Mining definitions from RDF annotations using formal concept analysis. In Proceedings of the International Joint Conference in Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
33. Alam, M.; Le, T.N.N.; Napoli, A. Latviz: A new practical tool for performing interactive exploration over concept lattices. In Proceedings of the CLA 2016-Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, 18–22 July 2016.
34. Venter, F.; Kourie, D.G.; Watson, B.W. FCA-based two dimensional pattern matching. In *International Conference on Formal Concept Analysis*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 299–313.
35. Li, G. DeepFCA: Matching biomedical ontologies using formal concept analysis embedding techniques. In Proceedings of the 4th International Conference on Medical and Health Informatics, Kamakura City, Japan, 14–16 August 2020; pp. 259–265.
36. Zhou, B.; Hui, S.C.; Chang, K. A formal concept analysis approach for web usage mining. In *Proceedings of the International Conference on Intelligent Information Processing, Beijing, China, 21–23 October 2004*; Springer: Boston, MA, USA, 2004; pp. 437–441.
37. He, H.; Hai, H.; Rujing, W. FCA-based web user profile mining for topics of interest. In Proceedings of the IEEE International Conference on Integration Technology, Shenzhen, China, 20–24 March 2007; pp. 778–782.
38. Singh, P.K.; Cherukuri, A.K.; Gani, A. A comprehensive survey on formal concept analysis, its research trends and applications. *Int. J. Appl. Math. Comput. Sci.* **2016**, *26*, 495–516. [CrossRef]
39. Azibi, H.; Meddouri, N.; Maddouri, M. *Survey on Formal Concept Analysis Based Supervised Classification Techniques, Machine Learning and Artificial Intelligence*; IOS Press: Amsterdam, The Netherlands, 2020; pp. 21–29. [CrossRef]
40. Krajca, P.; Outrata, J.; Vychodil, V. Parallel recursive algorithm for FCA. In *CLA*; Palacky University: Olomouc, Czechia, 2008; pp. 71–82.
41. Krajcaand, P.; Vychodil, V. Distributed algorithm for computing formal concepts using map-reduce framework. In Proceedings of the International Symposium on Intelligent Data Analysis, Lyon, France, 31 August–2 September 2009; pp. 333–344. [CrossRef]
42. Dean, J.; Ghemawat, S. MapReduce: Simplified data processing on large clusters. In Proceedings of the OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, USA, 6–8 December 2004; pp. 137–150.
43. Chunduri, R.K.; Cherukuri, A.K.; Tamir, M. Concept generation in formal concept analysis using MapReduce framework. In Proceedings of the 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), Chirala, India, 23–25 March 2017; pp. 191–204. [CrossRef]
44. Li, J.; Mei, C.; Lv, Y. A heuristic knowledge-reduction method for decision formal contexts. *Comput. Math. Appl.* **2011**, *61*, 1096–1106. [CrossRef]
45. Medina, J. Relating attribute reduction in formal, object-oriented and property-oriented concept lattices. *Comput. Math. Appl.* **2012**, *64*, 1992–2002. [CrossRef]
46. Wang, H.; Zhang, W.-X. Approaches to knowledge reduction in generalized consistent decision formal context. *Math. Comput. Model.* **2008**, *48*, 1677–1684. [CrossRef]

47. Antoni, L.; Cornejo, M.E.; Medina, J.; Ramírez-Poussa, E. Attribute classification and reduct computation in multi-adjoint concept lattices. *IEEE Trans. Fuzzy Syst.* **2020**, *29*, 1121–1132. [CrossRef]
48. Dias, S.M.; Vieira, N. Reducing the size of concept lattices: The JBOS approach. In Proceedings of the 7th International Conference on Concept Lattices and Their Applications, Sevilla, Spain, 19–21 October 2010; Volume 672, pp. 80–91.
49. Codocedo, V.; Taramasco, C.; Astudillo, H. Cheating to achieve formal concept analysis over a large formal context. In Proceedings of the Eighth International Conference on Concept Lattices and their Applications-CLA 2011, Nancy, France, 17–20 October 2011; pp. 349–362.
50. Kumar, C.A.; Srinivas, S. Mining associations in health care data using formal concept analysis and singular value decomposition. *J. Biol. Syst.* **2010**, *18*, 787–807. [CrossRef]
51. Stumme, G.; Taouil, R.; Bastide, Y.; Pasquier, N.; Lakhal, L. Computing iceberg concept lattices with titanic. *Data Knowl. Eng.* **2002**, *42*, 189–222. [CrossRef]
52. Singh, P.K.; Kumar, C.A. Concept lattice reduction using different subset of attributes as information granules. *Granul. Comput.* **2017**, *2*, 159–173. [CrossRef]
53. Dudyrev, E.; Kuznetsov, S. Decision Concept Lattice vs. Decision Trees and Random Forests. *arXiv* **2021**, arXiv:2106.00387. [CrossRef]
54. Belohlavek, R.; Baets, B.D.; Outrata, J.; Vychodil, V. Inducing decision trees via concept lattices. *Int. J. Gen. Syst.* **2009**, *38*, 455–467. [CrossRef]
55. Kuznetsov, S.O. Machine learning and formal concept analysis. In *Concept Lattices*; Eklund, P., Ed.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 287–312.
56. Kuznetsov, S.O.; Zueva, M. What can FCA do for Artificial Intelligence? In Proceedings of the FCA4AI 2024: The 12th International Workshop, Santiago de Compostela, Spain, 19 October 2024.
57. Dürrschnabel, D.; Hanika, T.; Stubbemann, M. FCA2VEC: Embedding Techniques for Formal Concept Analysis. In *Complex Data Analytics with Formal Concept Analysis*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 47–74, ISBN 9783030932787. [CrossRef]
58. Marquer, E. *LatticeNN–Deep Learning and Formal Concept Analysis*; Artificial Intelligence [cs.AI]; HAL: Lyon, France, 2020.
59. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
60. Rekabsaz, N.; Lupu, M.; Hanbury, A. Exploration of a Threshold for Similarity based on Uncertainty in Word Embedding. In *Advances in Information Retrieval; ECIR 2017. Lecture Notes in Computer Science*; Jose, J., Hauff, C., Sengor Altıngovde, I., Song, D., Albakour, D., Watt, S., Tait, T., Eds.; Springer: Cham, Switzerland, 2017; Volume 10193. [CrossRef]
61. Bogdanović, M.; Gligorijević, M.F.; Veljković, N.; Puflović, D.; Stoimenov, L. Cross-portal metadata alignment–Connecting open data portals through means of formal concept analysis. *Inf. Sci.* **2023**, *637*, 118958. [CrossRef]
62. Strok, F.; Neznanov, A. Comparing and analyzing the computational complexity of FCA algorithms. In Proceedings of the 2010 Annual Conference of the South African Institute of Computer Scientists and Information Technologists, Bela Bela, South Africa, 11–13 October 2010; pp. 417–420. [CrossRef]
63. Kuznetsov, S.; Obiedkov, S. Comparing performance of algorithms for generating concept lattices. *J. Exp. Theor. Artif. Intell.* **2002**, *14*, 189–216. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.