TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2026
42

Abdul Aziz

Promoting Inclusiveness in Open Government Data Portals through Feedback Mechanisms

Director/es
Nogueras Iso, Francisco Javier
López Pellicer, Francisco Javier

ISSN 2254-7606

Prensas de la Universidad
Universidad Zaragoza

Universidad de Zaragoza
Servicio de Publicaciones

ISSN 2254-7606

Universidad Zaragoza
1542

Tesis Doctoral

PROMOTING INCLUSIVENESS IN OPEN GOVERNMENT DATA PORTALS THROUGH FEEDBACK MECHANISMS

Autor

Abdul Aziz

Director/es

Nogueras Iso, Francisco Javier
López Pellicer, Francisco Javier

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

2025

Universidad Zaragoza

1474

Tesis Doctoral

Promoting Inclusiveness in Open Government Data Portals through Feedback Mechanisms

Autor

Abdul Aziz

Directores

Dr. Francisco J. Lopez-Pellicer

Dr. Javier Nogueras-Iso

Programa de Doctorado en Ingeniería de Sistemas e Informática

Escuela de Doctorado

Universidad Zaragoza

Doctoral Thesis

Promoting Inclusiveness in Open Government Data Portals through Feedback Mechanisms

Author

Abdul Aziz

Supervisors

Dr. Francisco J. Lopez-Pellicer

Dr. Javier Nogueras-Iso

Doctorate Program in Systems Engineering and Computer Science
Doctorate School

Abstract

Over the last decade, many Open Government Data (OGD) initiatives have been launched by public administrations to promote transparency, openness, citizen participation and reuse of data. However, their potential is still not being fully realised and they often do not fulfill the expectations of all types of final users, especially those users lacking technical skills. One of the reasons for not achieving this full potential may lie in technical problems derived from the unavailability of adequate metadata describing resources or the lack of appropriate formats and access protocols. Moreover, most OGD portals usually remain supplier-driven, focusing on data publication rather than on engaging users as main actors in the open data ecosystem. Nevertheless, the main reason for not reaching a wider range of final users is due to problems of accessibility to OGD portals, which limits the inclusiveness and the engagement of users. In the context of this research, accessibility refers to the technical ease with which users can locate, understand, and use datasets within open data portals, whereas inclusiveness refers to the extent to which diverse user groups such as citizens, developers, researchers, journalists, and public administrations are able to access, discover, interpret, and interact with open data in ways that meet their needs. To be more precise, accessibility is treated in this thesis as the tangible focus of analysis, while inclusiveness represents its broader implication.

The main objective of this research thesis is to investigate methodologies that enhance the inclusiveness of open data portals by improving data findability and fostering meaningful interaction between multiple stakeholder groups such as citizens, developers, researchers, and government organizations. On the one hand, open data portals should facilitate an easy discovery of the resources of interest to the different stakeholders. On the other hand, open data portals should facilitate the interaction with stakeholders if the discovered resources or facilities do not comply with the needs of the users.

There are four research questions that guide this work: What supplier-driven approaches can be developed to enhance the inclusiveness of open government data portals?; How can the inclusiveness of open government data portals, starting with the back-end design, be improved through the effective curation and annotation of datasets?; What user-driven approaches can support greater inclusiveness in open government data portals?; How can we transform users into the main actors of open data ecosystems?

This research study examines inclusiveness in OGD ecosystems by integrating supplier-driven and user-driven perspectives. The thesis enhances existing knowledge by proposing novel mechanisms and empirically connecting metadata quality, user feedback, and engagement processes within a unified framework. This study employs a design science methodology, supplemented by case study validation, to ensure that each proposed model whether for thematic annotation, feedback conceptualization, or social engagement analysis is theoretically sound and practically relevant. This integration bridges two disparate research areas: the technical enhancement of open data infrastructures and the social dynamics of user engagement. This approach offers a unified perspective on how inclusiveness can be designed, evaluated, and sustained within evolving OGD ecosystems.

In summary, this dissertation contributes novel approaches to open data research by introducing (i) a framework for automated thematic annotation of OGD to improve data findability, (ii) a conceptual model of feedback mechanism that utilises bi-directional interaction between users and data providers and (iii) a method to capture and analyze user feedback through social networks to understand patterns of engagement. These scientific contributions extend the state of the art in metadata enrichment,

feedback modeling, and user engagement analysis. From a societal perspective, the findings of this thesis will provide policymakers, developers, and researchers practical tools to design more accessible and inclusive open data portals, ultimately fostering more user-responsive and effective digital public infrastructures.

Resumen

Durante la última década, numerosas iniciativas de Datos Abiertos Gubernamentales (Open Government Data, OGD) han sido impulsadas por las administraciones públicas con el objetivo de promover la transparencia, la apertura, la participación ciudadana y la reutilización de los datos. Sin embargo, su potencial aún no se ha materializado plenamente y, con frecuencia, no cumplen las expectativas de todos los tipos de usuarios finales, especialmente de aquellos que carecen de competencias técnicas. Una de las razones que explica esta situación puede encontrarse en los problemas técnicos derivados de la falta de metadatos adecuados que describan los recursos o de la ausencia de formatos y protocolos de acceso apropiados. Además, la mayoría de los portales OGD continúan siendo de carácter proveedor, centrados principalmente en la publicación de datos más que en involucrar a los usuarios como actores principales del ecosistema de datos abiertos. No obstante, la principal razón por la que no se alcanza un público más amplio radica en los problemas de accesibilidad de los portales OGD, lo que limita tanto la inclusividad como la implicación de los usuarios. En el contexto de esta investigación, la accesibilidad se entiende como la facilidad técnica con la que los usuarios pueden localizar, comprender y utilizar los conjuntos de datos disponibles en los portales de datos abiertos, mientras que la inclusividad hace referencia al grado en que diversos grupos de usuarios como ciudadanos, desarrolladores, investigadores, periodistas y administraciones públicas pueden acceder, descubrir, interpretar e interactuar con los datos abiertos de acuerdo con sus necesidades. En términos más precisos, la accesibilidad se considera en esta tesis como el foco tangible de análisis, mientras que la inclusividad representa su implicación más amplia.

El objetivo principal de esta tesis es investigar metodologías que contribuyan a mejorar la inclusividad de los portales de datos abiertos mediante el incremento de la capacidad de descubrimiento de los datos y el fomento de una interacción significativa entre distintos grupos de interés, tales como ciudadanos, desarrolladores, investigadores y organizaciones gubernamentales. Por un lado, los portales de datos abiertos deben facilitar el descubrimiento sencillo de los recursos de interés para los diferentes actores. Por otro, deben favorecer la interacción con los usuarios cuando los recursos o servicios disponibles no se ajusten plenamente a sus necesidades.

Este trabajo se guía por cuatro preguntas de investigación: ¿Qué enfoques impulsados por los proveedores pueden desarrollarse para potenciar la inclusividad de los portales de datos abiertos gubernamentales?; ¿Cómo puede mejorarse dicha inclusividad desde el diseño del back-end, mediante una adecuada curación y anotación de los conjuntos de datos?; ¿Qué enfoques impulsados por los usuarios pueden fomentar una mayor inclusividad en los portales de datos abiertos gubernamentales?; ¿Cómo podemos transformar a los usuarios en los principales protagonistas de los ecosistemas de datos abiertos?

Esta investigación analiza la inclusividad en los ecosistemas de OGD integrando las perspectivas impulsadas tanto por los proveedores como por los usuarios. La tesis amplía el conocimiento existente al proponer mecanismos innovadores y conectar empíricamente la calidad de los metadatos, la retroalimentación de los usuarios y los procesos de participación dentro de un marco unificado. El estudio

adopta una metodología de ciencia del diseño, complementada con la validación mediante estudios de caso, para garantizar que cada modelo propuesto ya sea de anotación temática, conceptualización de la retroalimentación o análisis de la participación social sea teóricamente sólido y de relevancia práctica. Esta integración une dos áreas tradicionalmente separadas: la mejora técnica de las infraestructuras de datos abiertos y la dinámica social de la participación de los usuarios. De este modo, se ofrece una visión unificada sobre cómo la inclusividad puede diseñarse, evaluarse y mantenerse en los ecosistemas OGD en constante evolución.

En resumen, esta tesis doctoral aporta nuevos enfoques a la investigación en datos abiertos mediante la introducción de: (i) un marco para la anotación temática automatizada de los OGD orientado a mejorar la capacidad de descubrimiento de los datos; (ii) un modelo conceptual de mecanismo de retroalimentación que promueve la interacción bidireccional entre los usuarios y los proveedores de datos; y (iii) un método para capturar y analizar la retroalimentación de los usuarios a través de las redes sociales con el fin de comprender los patrones de participación. Estas contribuciones científicas amplían el estado del arte en el enriquecimiento de metadatos, el modelado de la retroalimentación y el análisis de la implicación de los usuarios. Desde una perspectiva social, los resultados de esta tesis proporcionan a los responsables políticos, desarrolladores e investigadores herramientas prácticas para diseñar portales de datos abiertos más accesibles e inclusivos, fomentando en última instancia infraestructuras digitales públicas más receptivas y eficaces.

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. Few portions of this thesis were developed with support from AI language models to assist in drafting, editing, and refining the text; all final content is my own work and responsibility.

October 2025

ACKNOWLEDGEMENTS

All praises are due to Allah, the only worthy of worship, the most gracious, and the most merciful. I would like to express my sincere gratitude to my supervisors, Prof. Francisco J. Lopez-Pellicer and Prof. Javier Nogueras-Iso, for their guidance that has allowed me to obtain my Ph.D. in their research group. I am thankful for their continuous support throughout my Ph.D. study for their patience, motivation, and encouraging attitude. I feel very fortunate and blessed to research under the supervision of marvelous human beings on earth with beautiful souls.

I sincerely appreciate the invaluable support from the 7eData team in Zaragoza particularly Jesús Pedro Gerique Molina as well as the team at the University of the Aegean, with special thanks to Charalampos Alexopoulos and Yannis Charalabidis. Their collaboration played a key role in making my time both intellectually rewarding and highly productive. I would also like to extend my heartfelt thanks to my friends Dagoberto Jose Herrera-Murillo and Mohsan Ali. Their unwavering support, expert insights, and contagious enthusiasm significantly contributed to the success of this PhD.

I am fortunate to have friends in IAAA Lab and at Universidad de Zaragoza, for always being around in good and tough times. I am also grateful to all my colleagues from the ODECO project.

My deepest gratitude goes to my father Ghulam Sarwar, my mother Pathani, Razia and my siblings, whose unwavering support has sustained me through various stages of life. Most importantly, I extend my heartfelt thanks to my wife, Parveen Ali, my son, Ammar Aziz for their boundless love, patience, and understanding during difficult times. Their strength and belief in me have been the foundation of all my achievements.

This thesis was supported by the ODECO project. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 955569.

CONTENTS

List of Figures **xv**

List of Tables **xvii**

Nomenclature **xix**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Problem Statement . . . 3
1.3 Challenges, objective and research questions . . . 4
1.4 Contributions . . . 5
1.5 The Structure of the Thesis . . . 7

2 Background **9**
2.1 Open Data and Open Government Data . . . 9
2.2 Open Data Ecosystems . . . 11
2.3 Features of Open Data Portals . . . 12
2.4 Inclusiveness in Open Data Portals . . . 13

3 A thematic annotation framework for Open Government data: a supplier-driven approach for inclusiveness **17**
3.1 Scope . . . 17
3.2 Related Research . . . 20
3.3 Methodology . . . 22
3.3.1 Evaluation of thematic classification correctness . . . 23
3.3.2 Learning to automatically classify based on an annotated corpus . . . 24
3.3.3 Predicting the closest theme of a dataset based on word/sentence embeddings 27
3.4 Experiments and results . . . 29
3.4.1 Corpus description . . . 29
3.4.2 Results of thematic classification correctness evaluation . . . 32
3.4.3 Results of automated supervised classification . . . 32
3.4.4 Results of theme prediction . . . 35

3.5 Discussion . . . 38
3.6 Summary . . . 39

4 A conceptual definition of roundtrip feedback in open government data portals 41
4.1 Related Work . . . 42
4.2 Research Methodology . . . 44
4.2.1 Definition of the Conceptual Scenario of Feedback Mechanisms . . . 44
4.2.2 Instantiation through case studies . . . 44
4.2.3 Extrapolation to other Open Data Portals . . . 45
4.3 Experiments and Key Outcomes . . . 47
4.3.1 Conceptual Scenario of Feedback Mechanisms . . . 47
4.3.2 Instantiation through Case Studies . . . 47
4.3.3 Extrapolation to Broader Open Data Portals . . . 51
4.4 Discussion . . . 56
4.5 Summary . . . 57

5 A method for the analysis of feedback through social networks 59
5.1 Related work . . . 60
5.2 Proposed Framework . . . 62
5.2.1 Snapshot-Based Analysis of OGD Portals . . . 62
5.2.2 Temporal Analysis of OGD Portals . . . 64
5.3 Experiments and Results . . . 68
5.3.1 Results for Snapshot-Based Analysis of OGD Portals . . . 68
5.3.2 Results for Temporal Analysis of OGD Portals . . . 72
5.4 Discussion . . . 79
5.5 Summary . . . 80

6 Conclusions and future work 81
6.1 Summary of contributions . . . 81
6.2 Open Issues . . . 85

7 CONCLUSIONES Y TRABAJO FUTURO 87
7.1 Resumen de contribuciones . . . 87
7.2 Cuestiones abiertas . . . 91

References 95

LIST OF FIGURES

1.1 Overview of the thesis structure, research questions, and contributions within the Open Data Ecosystem. 8

3.1 An excerpt of DCAT-AP metadata model highlighting the free-text elements for describing datasets and its thematic classification (*dcat:theme*). 18
3.2 The workflow for reporting the thematic classification correctness. 23
3.3 The workflow for reporting the thematic classification correctness. 23
3.4 Proposed method for the automatic classification of open datasets based on an annotated corpus 24
3.5 An example of unigrams, bigrams and trigrams for the sentence "Thematic Annotation of Open Government Data" 26
3.6 Proposed method for the prediction of the closest theme of a dataset based on the similarity of word/sentence embeddings 27
3.7 Distribution of datasets across the 13 themes of the European Data Portal. 30
3.8 Distribution of datasets with respect to the 20 most frequent combinations of themes. 30
3.9 The language distribution of the datasets. 31
3.10 Results of thematic classification correctness for a lot size of 29,793 records and LQ of 12.5%. 32
3.11 Confusion matrices for all the themes in experiment 3 of Table 3.1(*core* input, *basic* Normalisation, unigram features, SVM, overall accuracy of 93.65%) 35
3.12 ROC curve for all the themes in experiment 3 of Table 3.1 (*core* input, *basic* Normalisation, unigram features, SVM, overall accuracy of 93.65%) 36

4.1 Proposed Research Process for feedback mechanism 45
4.2 General modeling of feedback interaction 48
4.3 Feedback channels of National Open Data Portals 52
4.4 Average number of times different Feedback channels offered by ODP 53
4.5 Dendrogram of Countries Based on Feedback Channels 54
4.6 Cluster Profiling of Countries Based on Feedback Channels 55
4.7 Input feedback channels at Dataset level and Portal level 55
4.8 Relationship between input and output feedback channels across various countries . . 56

5.1 Proposed Research Methodology for Data Processing 62
5.2 Proposed Methodology for Data Processing 65
5.3 An example of a small SOM trained with a dataset consisting of records describing OGD initiatives 67
5.4 Cluster dendrogram 71
5.5 Cluster profile plot for mean factor score of the clusters obtained with K-means . . . 71
5.6 Dispersion of records in input dataset after applying PCA over a bi-dimensional space 73
5.7 Output SOM with 4 clusters and 8×7 dimension 74
5.8 Cluster dendrogram 75
5.9 Profiling of the Clusters 75
5.10 Dispersion of records in input dataset after applying PCA over a bi-dimensional space, with assigned cluster 77

6.1 A bird's eye view of the contributions 82

7.1 Una visión global de las contribuciones 88

List of Tables

3.1 Experiments and results for automated supervised classification: *core* input. 33
3.2 Experiments and results for automated supervised classification: *extended* input; *basic* and *translation* Normalisation. 34
3.3 Experiments and results for automated supervised classification: *extended* input; *tailored* Normalisation. 35
3.4 Experiments and results for theme prediction 37

4.1 Detection Methods for Various Channels 46
4.2 Comparative Summary of Feedback Interaction Processes across Five European Open Data Portals 50
4.3 Open Data Portals by Country 51

5.1 Description of the variables 63
5.2 Description of the variables collected for each OGD initiative and year 65
5.3 Values of variables for Open Government Data portals of the EU countries and their X (Twitter) activity in 2021 69
5.4 Spearman correlation 70
5.5 Rotated matrix for factor analysis 70
5.6 Cluster membership 72
5.7 Cluster shifting of the countries 78

NOMENCLATURE

Ac	Acceptable Number of Errors
API	Application Programming Interface
AQL	Acceptance Quality Limit
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-Separated Values
DCAT	Data Catalog Vocabulary
DG	Digital Government
DKAN	Drupal-based Knowledge Archive Network
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
GloVe	Global Vectors for Word Representation
GPL	General Public License
ISO	International Organisation for Standardisation
JSON	JavaScript Object Notation
KOS	Knowledge Organisation System
LDA	Latent Dirichlet Allocation
LQ	Limiting Quality
LR	Logistic Regression

LSTM	Long Short-Term Memory
MNB	Multinomial Naïve Bayes
NGOs	Non-Governmental Organizations
NGO	Non-governmental organisation
NLP	Natural Language Processing
ODP	Open Data Portals
OD	Open Data
OGDP	Open Government Data Portal
OGD	Open Government Data
OG	Open Government
OSS	Open-Source Software
OvR	One-vs-Rest Classifier
PDF	Portable Document Format
RDF	Resource Description Framework
ROC	Receiving Operating Characteristic
SOM	Self-Organizing Maps
SVM	Support Vector Machine
TFIDF	Term Frequency-Inverse Document Frequency
W3C	World Wide Web Consortium
WCAG	Web Content Accessibility Guidelines

CHAPTER 1

INTRODUCTION

1.1 Motivation

In the current digital landscape, the Open Data (OD) movement is experiencing rapid growth, driven by the exponential increase in data accessibility through open data portals [1, 2]. The European Commission projects have already estimated that by 2025, the net value of the size of the open data market of the European Union will be almost 200 billion euros, impacting over a million employees in the open data sector [3].

The core principle of open data is to enable free and unrestricted usage, sharing, and access to data in any available format [4]. Governments are progressively initiating OD initiatives and setting up dedicated portals to facilitate the dissemination of open data in reusable formats, as well as to promote transparency, accountability and public engagement [5, 6]. Consequently, a multitude of open data repositories, catalogues, and websites have appeared to serve this purpose.

OD Portals (ODP) play a pivotal role in easing data openness. ODPs serve as online repositories featuring detailed dataset descriptions based on key attributes such as authorship, provenance, and licensing [7]. These catalogues ease the exploration and administration of metadata records that provide valuable insights into datasets that may be accessible for download in various distribution formats. OD initiatives assume that publication of open data through ODPs will increase the demand for high-quality data and enhance the overall quality of ODPs. In addition, the publication of public sector data stands as one of the major drivers behind the prevailing movement to open government data through an open data portal [8].

Governments globally have initiated open data platforms such as the Spanish ODP (datos.gob.es), the French ODP (data.gouv.fr) or the EU ODP (data.europa.eu) to facilitate public access to datasets, including healthcare statistics, agriculture, environment, education, culture and sports. However, since vast amounts of open government data are published through diverse platforms and portals, users often face challenges in locating relevant, actionable information due to fragmented datasets, inconsistent formats, and varying levels of accessibility [9]. According to the International Organisation for Standardisation (ISO) Standard 9241-171:2008, accessibility is defined as the “extent to which

products, systems, services, environments and facilities can be used by people from a population with the widest range of user needs, characteristics and capabilities to achieve identified goals in identified contexts of use" [10, 11]. Some platforms for OGD portals like DKAN or CKAN facilitate plugins for accessibility, but in many cases this is limited to the compliance with web content accessibility guidelines (WCAG) [12], which does not prevent from inconsistent user experiences [13].

Before the rise of open data portals, citizens relied on limited or non-public sources of government information, which hindered informed decision-making and civic participation [14–16]. The absence of accessible and user-friendly data may reinforce inequities, as marginalised groups, including those with limited digital literacy or language barriers, could face difficulties in accessing essential public resources [17]. For instance, according to a 2018 audit conducted by the World Wide Consortium (W3C), only 12% of government portals comply with the basic and fundamental accessibility standards, effectively excluding screen reader users. The fundamental philosophy of OGD, which views data as a public good intended to empower all citizens [18], is clearly at odds with this exclusion. To address these challenges, it is important to note the diverse user groups and their need and hence there is a need for inclusive open data portals that must prioritise accessibility, usability, and diverse stakeholder needs [19, 20]. The thesis contributes to improving inclusiveness and engagement in open data ecosystems, which can directly support and enable democratic practices by fostering equitable participation. In this context, equitable participation refers to the fair and inclusive involvement of all types of users. Either the improvement in the discoverability of data or the integration of feedback channels for allowing both technical and non-technical users to express their data needs facilitate a better access to the data that may be relevant to take decisions in citizen participatory processes.

Before exploring the specifics of inclusiveness and an inclusive open data portal, it is essential to establish the scope of inclusiveness addressed in this Ph.D. For the purpose of this work, inclusiveness in OGD portals refers to the extent to which diverse user groups such as citizens, developers, researchers, journalists, and public administrations are able to access, discover, interpret, and interact with open data in ways that meet their needs [21–23].

In this thesis, open data is not considered a self-contained academic field but as a multidisciplinary research area. A multidisciplinary research area involves experts from various academic disciplines (in our case, it is information systems) collaborating to address a complex problem from their distinct perspectives. This research study is situated at the intersection of established disciplines such as Information Sciences (providing the technical and design perspective), Information Systems (focusing on data curation). This positioning clarifies that the contributions of the thesis are not meant to establish open data as a stand-alone discipline, but rather to advance knowledge within this shared research area by addressing questions of accessibility, inclusiveness, and user engagement in OGD portals.

1.2 Problem Statement

Despite more than a decade of global efforts to promote OGD, the inclusiveness of open data portals remains a largely unresolved challenge. While existing OD initiatives have significantly advanced openness and public access to information, they have paid comparatively little attention to how different user groups actually interact with OGD platforms [5, 6]. Current portal designs often assume a technically proficient audience, resulting in usability and accessibility gaps that marginalize less-experienced users and limit the broader societal value of open data [14, 15]. This thesis addresses this gap by investigating how various mechanisms can enhance inclusiveness in OGD portals through improved metadata quality, dataset annotation, and user feedback integration. The difficulty of ensuring that open data portals are genuinely inclusive has become increasingly apparent as both governments and non-government organisations continue to expand their open data initiatives. Although open data possesses significant potential to enhance transparency, innovation, and public participation, but still numerous open data portals face challenges of underutilisation because of technical obstacles that restrict accessibility for certain user groups. Technical barriers, including poor metadata standards, inconsistent data formats (such as CSV, PDF, and JSON), search features that are not simple, and inadequate support for diverse languages and accessibility requirements are some of the obstacles that hinder the successful data usage within open data portals. Furthermore, users who lack technical expertise may encounter obstacles due to more complicated query interfaces and application programming interfaces (APIs) [24]. Consequently, the application domain plays a crucial role in designing technical solutions for inclusiveness in open data portals. Different user groups including researchers, journalists and citizens interact with open data in distinct ways, requiring adaptive approaches to enhance usability. For example, a data scientist may require API access and well-documented metadata, whereas a policymaker might require user-friendly visualisation tools and plain-language summaries. The lack of personalised search, adaptive content distribution, and intelligent recommendation systems in most of the open data portals amplifies the problem, making data findability and utilisation a complex task for many users [9, 25, 26]. Meanwhile, sectors such as e-commerce and streaming services have utilised sophisticated personalisation strategies to improve customer engagement. Platforms like Netflix, Amazon, and Google use advanced recommendation algorithms to enhance content discoverability, guaranteeing that customers find relevant information quickly. Similarly, in the open data domain, incorporating metadata enrichment, feedback mechanisms, and well-annotated data accessibility can transform the way users interact with data portals, making them more inclusive and user-friendly.

The novelty of this thesis lies in enhancing inclusiveness in OGD portals through the integration of approaches that combine technical innovation with socio-technical considerations. Building on motivation and existing discussions of accessibility and user engagement, the thesis introduces three original contributions. The first is a supplier-driven framework for thematic annotation, which applies automated methods to improve the discoverability and classification of datasets, going beyond prior metadata enhancement techniques. The second is the conceptualization of roundtrip feedback as

a dynamic, bi-directional process, which extends the prevailing one-directional view of user input in OGD portals and highlights the role of trust, transparency, and responsiveness. The third is the development of a method for analyzing feedback through social networks, offering a new perspective on how open data is discussed and appropriated outside portal environments.

1.3 Challenges, objective and research questions

OGD initiatives are often launched with the ambition of promoting transparency, participation, and reuse of public information. These principles define the broader policy rationale for opening government data to society. However, the focus of this thesis is not to evaluate transparency directly, but rather to investigate how accessibility and inclusiveness can be strengthened in OGD portals. Accessibility refers to the concrete, technical, and design-oriented aspects that allow users to discover and interact with datasets, while inclusiveness represents the broader implication of ensuring that diverse user groups are able to benefit from and engage with open data. This Ph.D. research tackles the various challenges for the smooth and inclusive open data portals but here are the two main challenges addressed:

- **Ensuring Easy Discovery of Resources by Diverse Stakeholders:** One of the main challenges in the context of open data portals is the ability to support easy and efficient discovery of relevant resources by a wide range of stakeholders. These stakeholders, including government agencies, researchers, citizens, often have different needs, levels of technical expertise, and contextual requirements. A central issue underlying this challenge is the lack of standardisation in metadata and dataset classification. Various portals employ different metadata schemas, which significantly complicates and even hinders users ability to locate the relevant data effectively. Metadata plays a critical role in numerous aspects of data management, encompassing data integration, transmission, and transformation, as well as in the management of interpretation and search problems [27]. For example, how library users may search by author, topic, or publication date depending on their intent. Without intuitive search functionality, complete metadata, and well-structured categorisation aligned with user expectations, the navigation of vast and heterogeneous datasets becomes a burden. As a result, the utility of open data is diminished, and its potential for transparency, innovation, and informed decision-making is significantly reduced, particularly for non-expert or marginalised users. Hence, data accessibility can be improved by standardised methodologies and the development of complete metadata [5, 28].

- **Facilitating Responsive Interaction When Resources Do Not Meet User Needs:** Equally important is the challenge of enabling effective interaction between users and data providers when the discovered resources fail to meet user requirements. In many open data platforms, the mechanisms for user feedback, issue reporting, or content refinement are underdeveloped or entirely absent. Although open data ecosystems thrive on user contributions and participation, current methods of collecting and integrating feedback remain in their infancy [29, 30]. This

disconnect weakens the ability of portals to evolve in response to user needs. Furthermore, the design of many portals is oriented toward data experts rather than general public users, often excluding communities with limited technical backgrounds. The lack of inclusive design further complicates the ability of users to communicate their needs or articulate issues. Implementing structured feedback mechanisms such as user ratings, comment sections, and discussion forums not only empowers users but also contributes to the overall improvement of data quality and relevance. Facilitating this two-way interaction is essential for creating responsive, user-centered open data environments [31, 32].

The main objective of this research thesis is to investigate methodologies that enhance the inclusiveness of open data portals by improving data findability and fostering meaningful interaction between multiple stakeholder groups such as citizens, developers, researchers, and government organizations. This thesis is situated within the research area of OGD, positioned at the intersection of information sciences and information systems which together provide the conceptual and methodological foundation for the study. To address this objective, there are four research questions guiding this research:

- **RQ1**: What supplier-driven approaches can be developed to enhance the inclusiveness of open government data portals?
- **RQ2**:How can the inclusiveness of open government data portals, starting with the back-end design, be improved through the effective curation and annotation of datasets?
- **RQ3**: What user-driven approaches can support greater inclusiveness in open government data portals?
- **RQ4**: How can we transform users into the main actors of open data ecosystems?

The challenges identified in the domain of OGD directly influenced the development of the research questions. Firstly, the issue of inadequate metadata and restricted dataset findability (Challenge 1) prompted RQ1 and RQ2, which focus on supplier-driven approaches and the potential of dataset curation and thematic annotation to improve accessibility. Secondly, the few ways for users to interact with OGD portals and the lack of integration of user perspectives (Challenge 2) led to RQ3, which looks at user-driven ways to be more inclusive. Lastly, the idea that users are still mostly passive consumers of open data led to RQ4, which looks at how users can be turned into active participants and co-creators in open data ecosystems. Each of these questions is addressed in subsequent chapters through dedicated methodological approaches and empirical studies.

1.4 Contributions

This PhD thesis makes several contributions to the advancement of inclusiveness in open data portals by addressing both supplier-driven and user-driven dimensions. These contributions are designed to

enhance the findability of open government data, promote meaningful user interaction, and create actionable insights through user engagement. The work builds on the dual perspective of supply and demand by proposing structured and innovative solutions to long-standing challenges in data accessibility, annotation, and responsiveness. The core contributions are as follows:

- **A framework for the thematic annotation of open government data:** This is a supplier-driven approach for inclusiveness as an assisted thematic annotation of data during the ingestion process of data in a catalogue should enhance the findability of resources by different user domains. The framework provides and compares the performance of different machine learning classification techniques considering both annotated and not annotated metadata corpora. An assisted thematic annotation of data should enhance the discoverability of resources by different user domains.
- **A conceptual definition of roundtrip feedback in open government data portals:** This is a user-driven approach for inclusiveness as we provide guidelines for the design of ODPs to bridge stakeholder needs, improve data quality, and strengthen user engagement.
- **A method for the analysis of feedback through social networks:** This method allows us to compare the evolution and maturity of user engagement through social networks in different Open Data initiatives.

In the literature of open data ecosystems, relevant studies exist on thematic annotation and feedback mechanisms. However, in the case of thematic annotation we have identified that there is a lack of an integrated framework that unifies in a common place the possibility of evaluating the quality of annotations and suggesting the alternatives for automatic annotation. Our proposed framework for thematic annotation in chapter 3 integrates existing methods to first assess the quality of dataset themes, applying supervised classification when a reliable annotated corpus exists and unsupervised classification otherwise. In the case of works related to feedback mechanisms, existing works lack a coherent conceptualisation of feedback scenarios that capture the diversity of input and output channels and their interactions. This thesis fills that gap by proposing a roundtrip feedback model that visualises feedback as a continuous, bidirectional process between data suppliers and users, clarifying the role and effectiveness of different communication channels.

This thesis conceptualises the OGD ecosystem as a dynamic environment of interaction between data suppliers and diverse user communities including citizens, researchers, journalists, and developers who engage through open data portals. The supplier-driven dimension advances inclusiveness by improving dataset accessibility and thematic discoverability through automated annotation and metadata enrichment. On the other hand, the user-driven dimension investigates feedback mechanisms within portals, encompassing both feedback provision (input channels) and feedback utilisation (output channels). Also, the feedback analysis through social networks extends the inclusiveness framework beyond the boundaries of the portal, enabling the detection and analysis of user engagement patterns in external digital spaces.

1.5 The Structure of the Thesis

The remainder of this Ph.D. thesis is divided into chapters that address the various dimensions of inclusiveness in open data portals, from both supplier-driven and user-driven perspectives. This dissertation is organised as follows:

- **Chapter 2** provides the background necessary to understand the concepts of open data, open government data, open data ecosystems, open data portals, and inclusiveness within open data portals. Building on this foundation, the following chapters present the research contributions of this PhD.

- **Chapter 3** introduces the novel framework for the thematic annotation of Open Government Data. This framework proposes a workflow where the first step is to assess the thematic classification correctness of DCAT metadata using sample-based quality controls. In the case of a properly annotated corpus, we propose a procedure for building automatic classification models with different alternatives for property selection, normalisation, feature representation and supervised classification techniques. In the case of not having a properly annotated corpus, we propose an unsupervised procedure to predict the closest theme based on the similarity of word/sentence embeddings between datasets and themes.

- **Chapter 4** presents complete roundtrip feedback in open data portals with the analysis of input and output channels. To enhance user engagement with OGD portals, it is essential to optimise feedback mechanisms within these portals. This work aims to analyse the current status of feedback mechanisms as a foundation for their future improvement to strengthen user engagement and to advance in accountability and transparency.

- **Chapter 5** introduces the evolution of OGD initiatives and their user engagement along a temporal period. In the methodology, first, a set of variables are collected to describe the main features of Open Data initiatives and their associated social network activity. Then, to analyse these collected data from a multidimensional and temporal perspective, we apply the well-known technique of self-organizing maps to find hidden correlations between the status of different initiatives in the analysed period. Finally, as the number of map nodes is still too big to identify clear levels of maturity, a clustering algorithm is applied to group initiatives with a similar evolution status. The feasibility of this methodology has been tested by analysing 27 European Open Government Data portals between 2017 and 2021.

- **Chapter 6** concludes this dissertation by providing an overview of the results achieved and an outline of future lines for research.

Furthermore, Figure 1.1 presents an integrated overview of the Open Data Ecosystem as conceptualized in this thesis, illustrating how the main research components, questions, and contributions

are interrelated. The ecosystem is composed of open data suppliers and diverse user groups (citizens, researchers, journalists, and developers), who interact through the open data portal and its supporting tools. Within the open data portal, the supplier-driven dimension focuses on improving dataset accessibility and findability through automatic thematic annotation (highlighted in red colour), which addresses research questions RQ1 and RQ2 and is detailed in Chapter 3. This process enriches metadata in the open data catalogue, making datasets more findable and semantically coherent. On the other hand, user-driven dimension focuses on feedback management (highlighted in blue), encompassing both feedback response mechanisms (output channels) and feedback request mechanisms (input channels), corresponding to research question RQ3 and discussed in Chapter 4. Finally, feedback analysis through social networks (highlighted in green) extends the ecosystem beyond the portal, capturing external user engagement and addressing research question RQ4 in Chapter 5. Together, these components create a roundtrip feedback loop where supplier and user activities are continuously connected through data publication, feedback collection, and evaluation. The Figure 1.1 therefore summarizes the overall structure of the thesis, linking its methodological design to the theoretical definitions introduced in Chapter 2 and showing how each contribution collectively enhances accessibility, inclusiveness, and user participation in OGD portals.

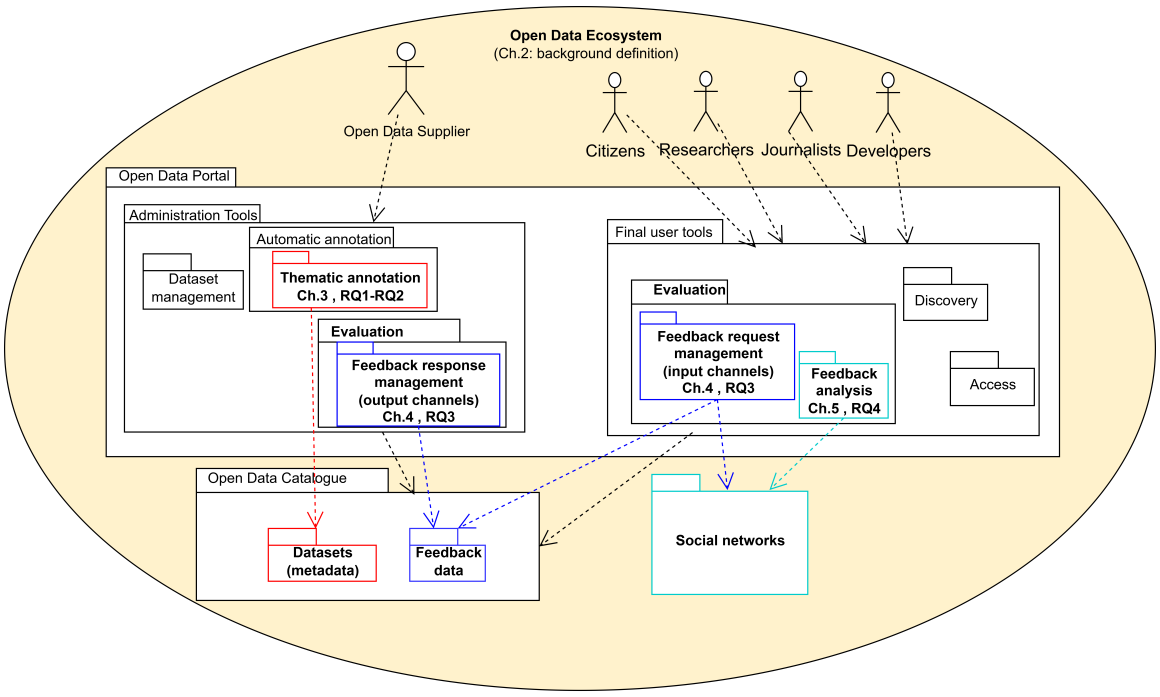

Fig. 1.1 Overview of the thesis structure, research questions, and contributions within the Open Data Ecosystem.

CHAPTER 2

BACKGROUND

This chapter provides the conceptual and theoretical background necessary to contextualize the research study topics introduced in Chapter 1. It reviews key concepts in open data and open government data with the idea of open data ecosystems as a foundation for understanding collaboration and interaction between stakeholders. The definitions presented in this chapter were selected based on their prominence in academic literature [2, 5] and their adoption in influential policy documents [18, 33]. These definitions are widely recognized and provide a consistent foundation for analyzing accessibility and inclusiveness in OGD portals. While alternative definitions exist, particularly in legal or technical communities, the selected ones represent the most authoritative sources for the purposes of this study. This chapter also explores the main features of open data portals as the technological entry points for users, as well as it discusses the notion of inclusiveness in open data portals, which is central theme of this thesis. By mapping these elements, the chapter establishes the context for the research questions, particularly those focused on how supplier- and user-driven approaches can enhance inclusiveness through improved accessibility and interaction.

Furthermore, the literature review presented here followed the strategy of combining the SALSA framework (Search, Appraisal, Synthesis, and Analysis) [34] with PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) [35] inspired documentation to ensure systematic coverage and transparency. The SALSA framework structures the review process and ensures that the literature is comprehensively identified, critically evaluated for quality and relevance, and systematically synthesized to reveal patterns and gaps in existing research. Complementarily, PRISMA provides a standardized protocol for documenting the review process, including the inclusion and exclusion criteria, search scope, and selection flow. Searches were conducted across major academic databases, including Scopus, Web of Science, and Google Scholar, using combinations of keywords such as "open data," "open government data," "open data ecosystem," "open data portals," and "inclusiveness."

2.1 Open Data and Open Government Data

Open Government is defined as "a culture of governance that supports the principles of openness, transparency, integrity, accountability, and stakeholder involvement in order to create democracy

and inclusive growth" [36]. Transparency makes it possible for citizens to scrutinise government and public involvement encourages collaborative decision-making which can lead to meaningful participation.

Open access is defined as "the provision of instant, unrestricted access to scientific work, including journal articles, theses and data sets, without any financial, legal or technological barrier". These works must be available under licenses (e.g. Creative Commons) that allow reuse with proper attribution [37].

Open-Source Software (OSS) defined as "users are granted the freedom to run, study, change, and redistribute the code for any purpose, including commercial use, using software that is released under licenses such as the General Public License (GPL) or the Material Transfer Protocol (MIT)". The Free Software Definition (FSF) and the Open-Source Definition (OSI) both include these ideas as a part of their respective guiding principles [38].

In this digital era, data has become the foundation of modern infrastructure and the driving force behind the success of information-centric economies. There is a need to convert data (raw information) into actionable insights by every sector such as healthcare, education, and government agencies to promote efficiency, competitiveness, and informed decision-making. To sustain growth in this linked world, emphasizing strong data governance, ethical frameworks, and inclusive access is not only favorable but essential [39]. In essence, data is no longer a passive resource but a strategic asset, shaping economies and redefining development in the 21st century [40].

Open data is defined as information that is readily accessible to all individuals and can be used, modified, and shared with minimal restrictions. Open government data (OGD) is a term that refers to data that is generated or requested by government agencies and made accessible to the public. OGD can enhance the effectiveness and efficiency of public services, increase institutional accountability, boost citizen participation, accelerate scientific progress, and foster the creation of other economic and social values [5, 41, 42] by making government datasets publicly accessible. On the other hand, despite the fact that open data is becoming more widespread, the genuine notion of open data is still far from being realised. [43]. The Open Data (OD) movement is seeing phenomenal growth in the current digital world, which is characterised by a rapid speed of change. One of the factors that has enabled the growth of the OD movement is the increased accessibility of data on open data portals [44]. Data portals have emerged as crucial hubs for the dissemination and easy accessibility of massive amounts of open data in today's information-driven world. The European Commission [8] argues that the continual distribution of open data through OGD portals enhances both the need for and the quality of data that is of a higher level. OGD portals are vital to the release of data, and this improves both the demand for and the quality of data. But it is important to make full use of the potential offered by open data through open data portals while simultaneously involving a variety of diverse stakeholders. Because of the disparity in representation, data portals are unable to facilitate the making of decisions that are both fair and driven by data across all industries. As a result, data portals need to identify and solve the technological elements that encourage inclusiveness and also need to suit the diverse expectations of all user domains.

2.2 Open Data Ecosystems

An ecosystem is a system of people, practices, values, and technologies in a particular local environment [45]. Ecosystems consist of many intercommunicating, somewhat tightly linked, and quite highly reliant components. Still, ecosystems vary in certain particular aspects.

According to Zuiderwijk et al. [46], an open data ecosystem is defined as the network of interconnected stakeholders, institutions, infrastructures, and technologies that support the publishing, access, use, and reuse of open data, emphasizing the evolving relationships among data providers, users, and the systems that enable the data flow. In these ecosystems, data is systematically cycled and recycled among providers and users. Participation from users is generally recognised as being crucial to the growth of open data ecosystems. However, suppliers are the primary force behind the emergence of existing ecosystems [47]. By making data openly accessible and useful by many user groups, such as governments, enterprises, academics, and people, an open data ecosystem fosters inclusivity, transparency, and innovation [18]. Several critical elements, such as data suppliers, data infrastructures, governance frameworks, and user interaction methods, are necessary for an open data ecosystem to function well. A number of researchers have characterised OD ecosystems in a technical manner, emphasizing the optimisation of OD supply to enhance data utilisation [46, 48, 49]. The realisation of OD ecosystems may take place at many levels, such as the level of data providers and the level of data users, or between these two levels, at the level of intermediaries [18, 50]. Intermediaries, defined as entities positioned between data producers and consumers are essential in open data ecosystems by fostering linkages among stakeholders and creating additional value.

Moreover, the extent of open data ecosystems may vary among different institutions, nations, regions, globally, and across many fields and domains [51]. Additionally, Kapoor et al. [52] propose that an OD ecosystem encompasses diverse activities, while Mulder et al. [53] enhance the understanding of a data ecosystem by incorporating its participants as well as the political and organisational frameworks that facilitate or engage in those activities. In addition to this, they also provide a comprehensive and theoretical framework to establish an ecosystem that has the potential to mobilise open data as a public resource, which may serve as a basis for the use of open data in a more inclusive manner. On the other hand, when it comes to the efficient and effective development of this idea, which is intended at giving value to all stakeholders, the most important part of the ecosystems metaphor is the recognition that users, technical innovators, government leaders, data managers, and policymakers are all interrelated [49].

A well-functioning open data ecosystem is made up of a number of interconnected components that, when combined, improve the accessibility, usefulness, and impact of the data [54]. The following are a few key components of an OD ecosystem.

a) **Data Providers and Publishers:** Data providers and publishers are the foundation of an open data ecosystem by producing, sustaining, and disseminating datasets across many fields. Government entities are the main sources, disseminating administrative, geographical and socioe-

conomic data via open data portals of national and regional governments such as datos.gob.es and datosabiertos.regiondemurcia.es to improve openness of public services [5]. In addition to governments, international organisations and NGOs such as the World Bank and Open Knowledge Foundation provide essential datasets in public health, climate change, and development, facilitating global research and policymaking [55, 56]. Moreover, academic institutions provide significant research datasets via platforms such as Zenodo but still issues like intellectual property problems and the absence of defined data-sharing protocols remain prevalent [57]. Enhancing legislative frameworks, metadata standards, and cross-sector cooperation may improve the inclusivity and sustainability of open data ecosystems, assuring wider involvement and long-term impact [46].

b) **Open Data Infrastructure:** The backbone of any open data ecosystem is its infrastructure that includes the technical, legal, and organisational frameworks that enable the publishing, accessibility, and the usage of open data. A sophisticated technological infrastructure contains data portals, APIs, metadata standards, and cloud storage solutions that provide efficient data discovery and interoperability [5]. Furthermore, platforms like CKAN, DKAN Socrata, and OpenDataSoft provide essential capabilities for cataloging, querying, and displaying datasets, hence enhancing accessibility for a wide range of various user groups [58]. In section 2.3, the importance of CKAN and DKAN is expalined in detail. Nonetheless, challenges such as inconsistent data formats, and the absence of machine-readable metadata often hinder the smooth integration and reuse [18].

2.3 Features of Open Data Portals

The increased availability of open data over the past decade is a direct consequence of open data portals whose objectives include, among others, improved delivery of government services, greater openness and accountability, more active citizen participation, and increased creation of economic and social value [1, 59]. The goal is to make full use of the potential of open data through open data portals while simultaneously involving a variety of diverse stakeholders.

Open Data portals are a sort of digital library, since they are online catalogues that include descriptions of datasets, known as metadata. These kinds of catalogues make it possible to find and manage metadata records describing datasets that are either accessible online or may be downloaded in a variety of distribution formats. Furthermore, metadata records facilitate the use and reuse of datasets by providing details of authorship, provenance, and license, among other details [7, 60, 61]. In fact, the use and reuse of public sector data is a crucial aspect driving the present trend of opening up government data [8, 62]. OGD portals play a critical role in opening the data, and the constant publishing of open data in OGD portals increases the demand for data of a high quality as well as a higher quality in the portal itself. However, the majority of existing open data ecosystems are not user-driven and thus fail to properly balance supply and demand. Although the importance of users in

shaping open data ecosystems is well acknowledged, existing ecosystems are mainly influenced by service providers [47].

The participation of users is essential to make existing OGD initiatives more user-oriented. Some open data portals already offer specialised forums or online forms where diverse groups of users may report on their experiences reusing the datasets available on these portals. Other initiatives even provide users access with specialised tools for storytelling to narrate their experiences with OGD datasets [63]. However, these feedback mechanisms are, in general, very heterogeneous and the input obtained from users is rarely accessible by the general public to be compared across different OGD initiatives.

Now let us have a look at the fundamental platforms that are regarded as the foundational elements for the emerging open data portals.

a) **CKAN (Comprehensive Knowledge Archive Network)**
In terms of open data publishing and dissemination, the Comprehensive Knowledge Archive Network (CKAN) is among the most popular open-source data management systems (DMS) [64]. In order to efficiently administer their open data portals, governments, academic institutions, and businesses worldwide adopt CKAN, which was first developed by the Open Knowledge Foundation [65]. The modular design and API-driven architecture of data ecosystems are essential for their implementation to make them visible and readily accessible [24]. In spite of the fact that CKAN provides essential tools for managing metadata, discovering datasets, and gaining programmatic access, but still it lacks accessibility and inclusivity and requires additional customisation and plugins to address issues like multilingual support. CKAN supports structured data discovery by implementing international metadata standards such as DCAT (Data Catalog Vocabulary) and Schema.org [66]. This ensures that datasets are interoperable across different platforms and can be easily indexed by search engines and external services.

b) **DKAN (Drupal-based Knowledge Archive Network)**
DKAN is an open-source data portal system developed on Drupal, aimed at improving flexibility, civic involvement, and accessibility [67]. Although functionally comparable to CKAN, DKAN utilises Drupal's comprehensive content management features, making it especially appropriate for accessible and community-oriented open data projects [66]. DKAN has been acknowledged for its inclusive attributes, such as integrated linguistic support and accessibility tools [25]. However, it also faces challenges related to customisation complexity and community adoption, as it has a smaller developer base compared to CKAN.

2.4 Inclusiveness in Open Data Portals

It is important to understand what inclusiveness means before looking into the details of how it works in open data initiatives. In the context of this thesis, inclusiveness in OGD portals refers to the extent to which diverse user groups such as citizens, developers, researchers, journalists, and

public administrations are able to access, discover, and interact with open data in ways that meet their needs [21–23].

Inclusiveness in open data portals is crucial for guaranteeing fair access to information across various user groups. One of the main challenges is making the open data portals accessible to all kinds of users including the non-technical, less-skilled, users facing difficulties while navigating for the datasets. Hence, compliance with international standards like the Web Content Accessibility Guidelines [12] is essential to address these gaps which may hinder the ability to interact with datasets effectively and also ensuring interfaces are navigable and datasets perceivable by all user domains. Additionally, language barriers restrict usability of the portal, as monolingual portals exclude non-native speakers. Providing multilingual metadata, dataset descriptions, and user interfaces can significantly broaden engagement and foster inclusivity, particularly in regions with diverse linguistic populations. Therefore, it is mandatory to understand that to transform an open data portal into an inclusive open data portal, it is essential to identify and address the technical dimensions that support inclusivity [47]. From a software engineering perspective, inclusiveness can be conceptualized as a non-functional property of open data portals, similar to accessibility. As such, it requires explicit characterization and measurable indicators. In this thesis, inclusiveness is operationalized through three complementary dimensions: (i) data discoverability, measured via the accuracy and coverage of thematic annotation and metadata quality indicators; (ii) user interaction, measured via the presence and variety of input and output feedback mechanisms; and (iii) ecosystem engagement, measured via clustering of portals according to their adoption of communication channels and the intensity of user participation. While these indicators do not capture the full societal breadth of inclusiveness, they provide concrete and replicable measures that allow inclusiveness to be assessed in practice within the technical and organizational scope of OGD portals.

Technical solutions are important enablers of improved accessibility, inclusiveness and user participation. For example, adaptive user interfaces allow users to customise features such as managing the personal dashboard layouts to meet their personal preferences and accessibility needs. Additionally, enhancing API usability with simplified documentation and interactive tutorials can make programmatic access to data more accessible. This empowers users who may not have advanced coding skills, making the portals more user-friendly [24].

User-centric design and community-driven strategies are essential for maintaining inclusivity. Integrating crowdsourced data annotations allows users to add contextual insights or metadata modifications, boosting dataset clarity and usefulness for mixed audiences [68]. Engaging with the community through workshops, forums, or participatory design sessions ensures that portals evolve to meet real-world needs, helping to bridge knowledge gaps and build trust with underrepresented groups [69]. At the same time, it is important to recognize that technical solutions alone are not sufficient and can even risk reinforcing exclusion. Digital divides, language barriers, limited digital literacy, and algorithmic biases may prevent certain groups from benefiting equally. For this reason, the thesis adopts a socio-technical perspective, combining technological design with participatory

and inclusive practices to ensure that innovations act as enablers rather than barriers. This balanced approach highlights both the potential and the limitations of technology in fostering inclusiveness.

CHAPTER 3

A THEMATIC ANNOTATION FRAMEWORK FOR OPEN GOVERNMENT DATA: A SUPPLIER-DRIVEN APPROACH FOR INCLUSIVENESS

This chapter proposes a supplier-driven framework aimed at enhancing inclusiveness in open data ecosystems, specifically targeting the improvement of dataset accessibility through effective curation and annotation. This framework builds on existing work in metadata quality, dataset documentation, and semantic enrichment, and positions suppliers as active enablers of discoverability in open data portals. To ground the framework, a review of relevant literature on metadata quality, dataset documentation, and semantic enrichment was conducted. The literature review followed the same structured strategy of combining the SALSA framework with PRISMA as explained in the introductory part of chapter 2. However the search terms for this chapter are different and searches were carried out in Scopus, Web of Science, and Google Scholar using combinations of keywords such as "metadata quality," "dataset curation," "automated annotation," "open data discoverability," and "open government data portals". Moreover, this chapter addresses research questions RQ1 and RQ2.

3.1 Scope

The increased deployment of open data catalogues and portals has enabled the distribution and straightforward retrieval of substantial quantities of open data in the information-driven society and has leveraged the growth of the open data movement [70–74]. However, the challenge lies in making sense of this enormous resource. OGD portals are organised on the basis of pyramidal structures: national open data portals are aggregators of the contents harvested from open data portals maintained by governments in charge of administrative areas with a narrower scope. This means that open data catalogues must usually integrate heterogeneous metadata records describing datasets that have been harvested from different catalogues with a more local scope. Therefore, often open data catalogues struggle with findability due to limited and inconsistent metadata [24].

Taking into account these scenarios, where OGD portals at national or cross-national level must integrate the metadata contents from different sources, it is essential to provide users with classification tools in order to locate easily the information of interest [56]. Information classification schemes

generally fall into two main types: exact and ambiguous [75]. Exact organisational schemes precisely divide information into clear and mutually exclusive sections, using methods such as alphabetical, chronological or geographical sorting. Ambiguous organisational schemes, on the other hand, delimit information into categories that resist precise definition, often due to linguistic ambiguities and human subjectivity. This classification encompasses thematic organisation schemas along with other variants such as task-oriented and metaphor-oriented schemas. Thematic organisation, in particular, prioritises the grouping of related items, promoting associative learning and facilitating adaptive information retrieval strategies.

This work is focused on the study of the thematic annotation of datasets included in metadata of OGD catalogues. Rich metadata is one of the core tools to fulfill the FAIR principles [57] and improve the findability, accessibility, interoperability, and reuse of digital assets. To address the findability aspect, one of the almost mandatory recommendations is to include keywords and themes within metadata. This is typically checked by metadata quality evaluation methodologies [60, 76, 77]. Although national or even cross-national catalogues like *data.europa.eu*, the official European Data Portal, combine metadata compliant with different metadata vocabularies, most of them are derived from DCAT [78, 79]. DCAT is the W3C's Data Catalog vocabulary for describing open data [80]. The advantage of using DCAT derived vocabularies is that the thematic annotation of datasets is encouraged thanks to the inclusion of a specific property called *dcat:theme*. Figure 3.1 shows an excerpt of DCAT-AP [81], one of the main application profiles derived from DCAT for the description of public sector information, where we can observe some of the main metadata properties for describing catalogues, datasets and their distributions. Within the metadata, themes provide a higher degree of semantic structure that goes beyond individual keywords and descriptions. Users may conveniently find important information by classifying datasets according to their applicable themes, independently of the use of exact terminology.

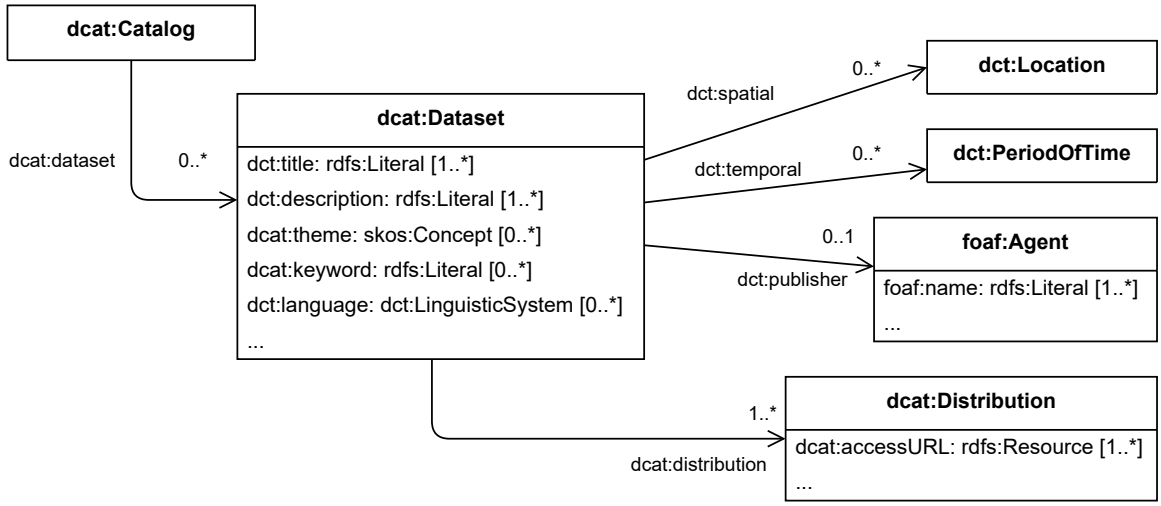

Fig. 3.1 An excerpt of DCAT-AP metadata model highlighting the free-text elements for describing datasets and its thematic classification (*dcat:theme*).

Nevertheless, just the use of metadata vocabularies including a property for thematic annotation is not enough. A missing or wrong thematic annotation hinders the findability. Therefore, we must ensure that the content of this property is accurate. Recognizing the potential of thematic annotation, several researchers [82, 83] have investigated several techniques for annotation, which involves giving significant labels to data. Manual annotation is precise and thorough but is limited in its capacity to scale up [84]. Automatic annotation exploits machine learning and natural language processing (NLP) to automatically classify information thematically [85]. Using automated annotation may have several advantages. It has the ability to significantly reduce the amount of human work needed, enabling the rapid processing of large datasets. Moreover, it may promote consistency and establish criteria in the annotation process, hence improving the findability and availability of the annotated resources [86].

While there has been substantial progress in publishing OGD, many OGD portals still lack a consistent and scalable mechanism for the thematic annotation of datasets coming from heterogeneous sources, making it difficult for users to discover relevant datasets efficiently. Existing approaches often rely on manual categorisation and are inconsistent across portals. This need for an automated mechanism establishes a research niche, which is covered in this work with the design of a framework that encompasses a structured and systematic set of concepts, methods, and software components to guide the process of thematic annotation of OGD. The design of this framework aims to address three main research questions:

1. What is the current state of thematic annotation in open government datasets, and how accurately do these annotations align with the content of the datasets? To answer this question, we have proposed an implementation of the method proposed in

2. Assuming that our collection of datasets (corpus) is properly annotated with themes, to what extent can new datasets be automatically classified? To answer this question, we have tested different machine learning algorithms and preprocessing strategies on an annotated metadata corpus.

3. In the case of not having an annotated corpus, or in the case our corpus is not properly annotated, how can relevant themes be assigned to a dataset from free text metadata? To answer this question, we have tested different strategies based on word-embeddings and sentence-embeddings of metadata to identify the closest theme from a predefined list of themes based on their definitions.

As stated in the introduction chapter, this framework is a supplier-driven approach because thematic annotation can be integrated during the ingestion process of metadata records describing the datasets provided by an OGD catalogue. The rest of this chapter is structured as follows. Section 3.2 provides a literature review about the thematic annotation of OGD. Then Section 3.3 presents our proposed framework for the automatic thematic annotation of OGD, which includes the evaluation of the thematic classification correctness in the case of having an existent annotated corpus. Section 3.4 presents the results after applying the proposed methodology to a corpus of metadata records from

the European Data portal (*data.europa.eu*), which are discussed in Section 3.5. Finally, Section 3.6 concludes with a summary of the contributions.

3.2 Related Research

Metadata is a critical component in numerous facets of data management, encompassing the integration, transmission, and transformation of data, among others [27]. As highlighted by frameworks for assessing the quality of open data portals [60, 76, 77, 87], missing or incomplete metadata hinders the findability of data. A wrongfully assigned theme which does not accurately represent the content of the dataset might induce low discoverability of the data, as well as low recall of search results to the interested stakeholder aiming to acquire the data, a situation that gives impetus to the quest for efficient and accurate thematic annotation.

Given the importance of providing correct metadata without the burden of accomplishing this task manually, different strategies have been suggested and developed over the years with the aim of achieving a fully or partially automated thematic annotation of resources. Although each strategy emphasises a unique set of conceptual areas of knowledge and experience, artificial intelligence techniques like machine learning are acquiring an increasing role of portal curators [88, 89].

As Semantic Web technologies are widely used as a mechanism to publish and reuse open data [90] and metadata is the core of the Semantic Web [91], many research works on automatic annotation are close related to the use of these technologies. For instance, Pavia et al. applied ensemble methods to classify Web-scale datasets through their metadata using a hybrid Recurrent Neural Network composed of LSTM and Bi-directional LSTM units and Naïve Bayes models at a second phase. In a more specific context and regarding bibliographic data [92], Carducci et al. worked on text categorisation for automatic metadata annotation in order to annotate records, separating between philosophical documents and other disciplines [93]. To facilitate this binary classification purpose, they employed NLP and other ensemble learning techniques, integrating domain knowledge and information gained through semantic networks (BabelNet) to decide whether a given document (e.g., thesis) is within the philosophical domain or not. The annotated data is then used to train the chosen supervised learning algorithms and automatically classify the metadata according to the thematic subject of the examined record. Likewise, Verberne et al. investigated the processing and classification of electoral manifestos [94]. After optimizing different parameters including passage segmentation, OCR or formatting, the results showed that the classifier matches human experts in accuracy and recall.

There are also recent studies focused on OGD portals highlighting the role of automated keyword extraction in enhancing thematic organisation and improving findability in open data portals. For instance, Ahmed et al. propose BRYT, an automated keyword extraction tool designed specifically for open datasets, enabling more accurate and scalable thematic categorisation [95]. Similarly, Kliimask and Nikiforova introduced TAGIFY, a language model-powered tagging interface aimed at improving data discoverability through enriched metadata in OGD portals [96]. These approaches underscore how

NLP and semantic tagging can mitigate linguistic ambiguities inherent in ambiguous organisational schemes, thereby supporting more effective associative learning and adaptive retrieval strategies. Huseynov et al. also emphasises the power of NLP to propose a recommender system for datasets [97]. Using the Word2Vec word-embedding technique to encode the free text content of different metadata properties in a vector space, their system provides the users with the possibility of selecting an input dataset and discovering the recommended datasets with a closer embedding in the vector space.

Several attempts for improved annotation services using semantic approaches have also been made in specific data domains such as the biomedical domain. Sasse et al. conducted a literature review on existing semantic metadata annotation services and identified their software requirements in accordance with the FAIR principles: availability as open code; compatibility with common data formats; use of FAIR terminologies; possibility of terminology search; suggestion of annotations; availability of interfaces to external terminologies; and extension of terminologies [98]. Although they concluded that there are no metadata annotation tools that meet all the requirements, this study highlights the importance of annotation tools and the availability of functionalities for suggesting annotations. In a more specific context about the psychiatric and psychological domain, Hudon et al. analysed the literature on the potential of machine learning to assist in the thematic annotation and classification of text in a psycho-therapeutic context [99]. Their findings demonstrated that, although the existing literature on this specific topic is limited, some algorithms such as Support Vector Machine classifiers achieved sufficient accuracy in the performed text classifications, and that this type of classifiers is consistently used for classification in the context of medical or clinical text data [100].

Automatic annotation has also been attempted for environmental science metadata. Tuarob et al. aimed to alleviate the problem of environmental metadata harvesting from various and disparate sources with varying levels of metadata quality and curation [101]. They gathered datasets from 4 different archives, selecting for each of them a subset of 1000 annotated documents, and the textual content and attributes of the documents were pre-processed (removal of stop-words, stemming) to obtain a Term Frequency-Inverse Document Frequency (TFIDF) representation. In order to rank automatically candidate themes for the dataset they used different similarity measures based on cosine similarity and Latent Dirichlet Allocation (LDA). Focusing on the processing of images, Ellen et al. targeted plankton image classification using context metadata (such as perimeter, symmetry, temporal and geographic information) to improve the performance of feature-based classifiers [102]. They demonstrated that the inclusion of context metadata might be of substantial gain for classification accuracy in deep learning models, mainly Convolutional Neural Networks. Likewise, Peng et al. proposed a unique biological data classification feature selection method to enhance feature categorisation [103]. The technique uses filter and wrapper approaches: it pre-selects feature subsets to improve search efficiency and utils receiving operating characteristic (ROC) curves to assess feature and subset performance. Furthermore, on the viticulture domain, Mylonas et al. proposed a platform for data annotation that includes a thesauri manager for the obtainment of Linked Data Vocabularies [104]. These vocabularies are used in the platform for both manual and automatic annotation based on NLP

techniques and supervised learning models such as k-nearest neighbours and linear and random forest regression.

When domain-specific research is being carried out, the specificities and domain-sensitive requirements need to be considered, to prevent or be aware of advanceed algorithmic biases and limitations. Wu et al. presented the status for automated metadata annotation in the cultural heritage (CH) domain and discussed the potential of machine learning applications supporting the curating processes of digital artifacts [105]. They provided a summary of recommendations to improve these aspects of automated metadata annotation by leveraging already existing text and images of high quality, utilising inference of meaning for classification from simple object recognition to tackle metaphoric and symbolic representations in the digital realm and providing quality indicators on the results to tackle non-uniform and non-consistent automated indexing. Similarly, Ibáñez et al. provided a quantitative analysis of Linked Data in accessible government datasets throughout Europe [106]. They examined the popularity of RDF as a publication format, the accuracy of connected datasets, and the prevalence of established terminologies. Furthermore, the negative effect of poor metadata description on the discoverability of digital cultural heritage artifacts was also addressed by Kaldeli et al. who proposed CrowdHeritage, an ecosystem supportive of end-to-end improvement of metadata utilising crowdsourcing, machine and human intelligence, semantic, and aggregation techniques [107].

The framework for thematic annotation proposed in this work integrates the existing knowledge in the state of the art of this field. First, the initial assessment of thematic classification correctness adapts the methodology proposed by Nogueras-Iso et al. to establish quality controls on dataset themes [76]. Second, the supervised classification techniques applied for new datasets in case of having a previously annotated corpus are similar to other works in the literature [92, 99]. In addition, the needs for preprocessing and feature representation are similar to other works using free text metadata as input [101]. Third, the unsupervised classification techniques applied in our framework also share some similarities with respect to the works of Ahmed et al. [95], Kliimask et al. [96] and Huseynov et al. [97], as they also exploit the benefits of using word embeddings and language models. Our proposal compiles all these alternatives within a unified framework, which allows the comparison of the suggested thematic annotations for new datasets in two different scenarios: the existence of a properly annotated corpus; or the unavailability of a properly annotated corpus.

3.3 Methodology

This section outlines our proposed methodology of the thematic annotation of OGD. Figure 3.2 shows the general workflow envisioned in this framework. In the case of counting on an annotated corpus, we first need to evaluate the thematic classification correctness before building a machine learning model for the classification of datasets. In contrast, if there is not an available annotated corpus or its classification correctness is not acceptable, we opt for predicting the closest theme measuring the similarity between the word/sentence embeddings of datasets and themes.

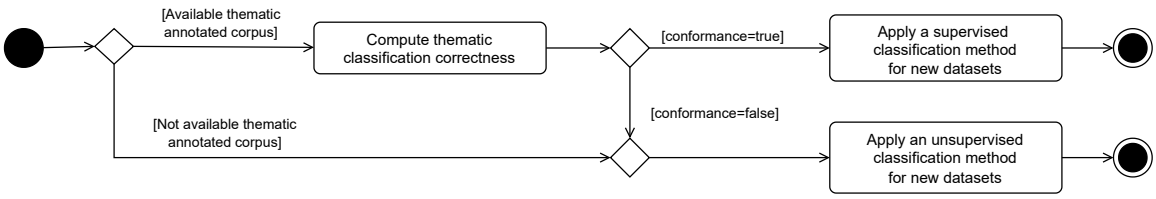

Fig. 3.2 The workflow for reporting the thematic classification correctness.

3.3.1 Evaluation of thematic classification correctness

To evaluate the thematic classification correctness of an annotated corpus, we propose to follow the method proposed by [76]. This method adapts ISO 19157 standard for geographic information quality to the context of open data metadata, and proposes a series of quality controls on six quality categories: completeness, logical consistency, temporal quality, thematic accuracy, positional correctness and quality of free text. In particular, the thematic accuracy category includes a quality element focused on thematic classification correctness, i.e. the correctness of the thematic keywords and categories included in the metadata with respect to a universe of discourse.

As we need a manual annotated corpus for the ulterior development of automatic classification model, the assessment of the thematic correctness must be made also manually using a sample-based inspection and a *Limiting Quality* index, which determines the sample size (n) according to the corpus size and the maximum number of errors (Ac) that can be accepted to assure a statically equivalent percentage of errors (*Acceptance Quality Limit* or *AQL*) if the full corpus were evaluated. Therefore, the computation of the thematic classification correctness requires the compilation of two associated results: a quantitative result and a conformance result. The quantitative result consists in obtaining a numerical value for the ISO 19157 D.63 measure, which is defined as the number of incorrectly classified records. The conformance result verifies whether the number of errors in the quantitative result surpasses or not the acceptable number of errors (Ac) for the considered sample size.

Figure 3.3 shows the workflow that must be followed to compute the thematic classification correctness. In general, the workflow of the assessment starts by considering the whole corpus of

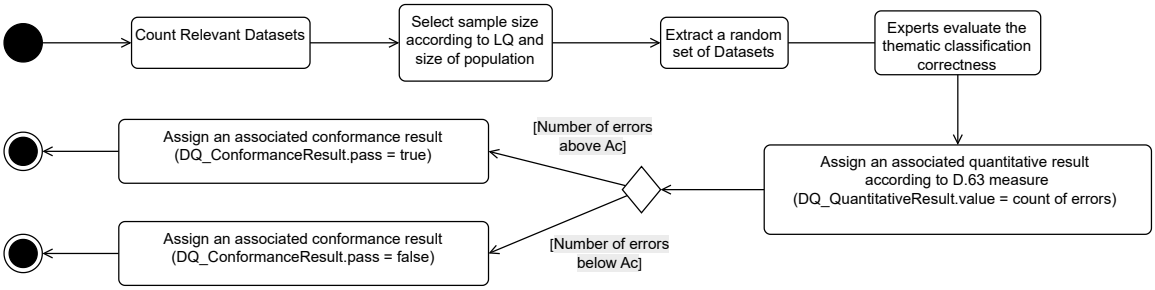

Fig. 3.3 The workflow for reporting the thematic classification correctness.

datasets of our case, including all the valid and relevant samples. Then, a selected sample size proportionate to the initial population is included to undergo the pre-assessment process, and a random

sample set is chosen using a random number generator. Afterwards, the process of manual assessment of erroneous instances from the selected sample is initiated. In order to implement this assessment, we decided that the random sample had to be evaluated by different experts and that this evaluation implied the inspection of the resources associated with the datasets. Then, the experts should assign to them between one and three related themes according to the perceived content of the dataset, its title, its description, and the associated keywords. Then, a consensus should be reached by the experts to consider as correct a theme classification if at least one of the assigned themes by the experts corresponded to the initial theme assigned to the dataset. The cases where the initially assigned theme of the dataset do not correspond to the themes assigned by the experts should be annotated as errors. Finally, the associated quantitative and conformance results are assigned.

3.3.2 Learning to automatically classify based on an annotated corpus

Assuming that we count on an annotated corpus of datasets where each dataset has been properly annotated with themes, this component of our annotation framework is focused on building models for the automatic thematic annotation of datasets. For this purpose, we have tested different machine learning algorithms that are typically applied for automatic classification problems in supervised scenarios. Figure 3.4 shows the workflow followed to build a model for the thematic annotation of OGD thematic annotation. The proposed steps in this workflow are the selection of metadata properties, the normalisation of the input text to extract terms, the transformation of the terms into an appropriate feature representation, and the generation of the classification models.

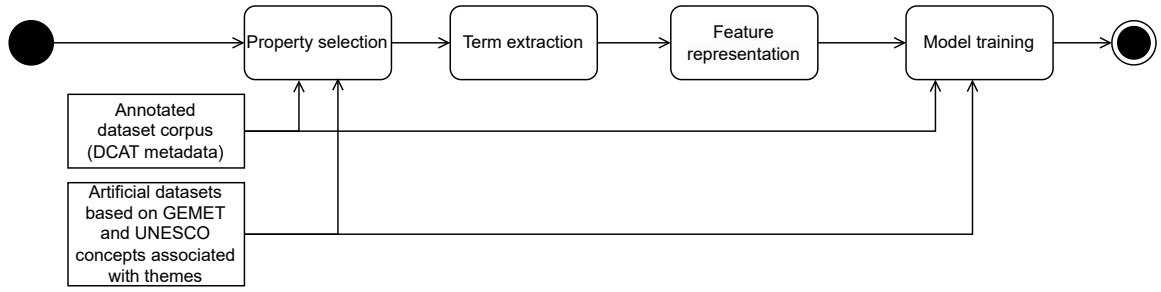

Fig. 3.4 Proposed method for the automatic classification of open datasets based on an annotated corpus

Property selection

In this framework we assume that metadata is compliant with a metadata vocabulary derived from DCAT. As shown in Figure 3.1, this type of vocabularies includes free-text properties for describing a dataset in terms of a title (*dct:title*), a general description (*dct:description*) and several keywords (*dcat:keyword*). In addition, datasets have also an associated theme thanks to the *dcat:theme* property, whose range is a concept from a well-known Knowledge Organisation System (KOS) expressed in SKOS format [108].

Therefore, we decided to use the combination of the text provided in *dct:title*, *dct:description* and *dcat:keyword* as input text for the classification. With respect to the themes or categories to be assigned after the classification process, we assumed in the experiments that the list of themes belonged to the KOS proposed by the European data portal [109], but this is interchangeable with any other KOS if a different corpus of metadata must be classified.

In addition, we also considered the possibility of generating some artificial datasets for each theme to reinforce the classification models to be built. For this research study, we selected GEMET[1] and UNESCO[2] thesauri due to their thematic breadth and multilingual coverage. These thesauri have been widely used for cataloguing purposes along the years to facilitate a harmonised thematic classification of datasets and reduce the gap between the vocabulary of data users and data publishers [110, 111]. Using the GEMET and the UNESCO thesauri, we aligned the European data themes with the main themes in GEMET and UNESCO (a theme is defined as a micro-thesaurus in UNESCO). Using this alignment, we extracted the preferred labels of the concepts associated with each theme in GEMET and UNESCO thesauri. Dividing the list of associated concepts for each theme in groups of a fixed number of concepts, we converted each group of concepts into an artificial dataset classified with a European data theme and described with the preferred labels of these concepts.

Term extraction

The next step in the workflow is the tokenisation of the input text describing each dataset and the extraction of terms. For the transformation of tokens into final terms, we have considered a mandatory *basic* level of Normalisation and two optional processes of Normalisation called *translation* and *tailored* Normalisation.

The mandatory *basic* Normalisation level incorporates the following processes: stop word removal (i.e., removing common words like ‘a’, ‘an’, ‘the’), special character removal (i.e., removing characters like ’$’, ’%’, ’&’), link removal (i.e., removing hyperlinks), lowercasing (i.e., converting all text to lowercase) and stemming (i.e., reducing words to their root form).

In addition, we observed that although metadata from OGD catalogues can be downloaded in RDF format and the language of metadata properties can be restricted to a common language such as English, the free-text content frequently appears in other languages. To address this issue, we explored the use of a *translation* Normalisation approach that employs an API to detect the most likely source language in the free text values and translate them into English, thereby improving consistency.

Last, we also considered a *tailored* Normalisation to remove noise in the free text derived frequently from spelling mistakes and the use of non-common English words such as acronyms, the names of data provider organisations or other technical terms which only make sense within the context of the data provider organisation. For this purpose, there are resources like PyEnchant,[3] which provides access to a dictionary of the English dialects spoken in different regions of the world such

[1]https://www.eionet.europa.eu/gemet/en/themes/
[2]https://vocabularies.unesco.org/browser/thesaurus/en/groups
[3]https://pypi.org/project/pyenchant/

as American English, British English, or Australian English and can be used to discard terms not contained in this dictionary.

Feature representation

Feature representation is an essential step in our workflow since our objective is to convert unprocessed text input into a vector representation acceptable for our machine learning models. Here, we will explore key features commonly used in text processing, specifically unigram, bigrams, and trigrams (also called N-grams) for all of our experiments. Unigrams are a typical bag of words vector representation where each word is a distinct dimension. Bigrams are vector representations where each dimension are the biwords found in the input text. Trigrams are vector representations where each dimension is a distinct trigram. Figure 3.5 illustrates an example of the unigrams, bigrams, and trigrams that can be generated from an input text.

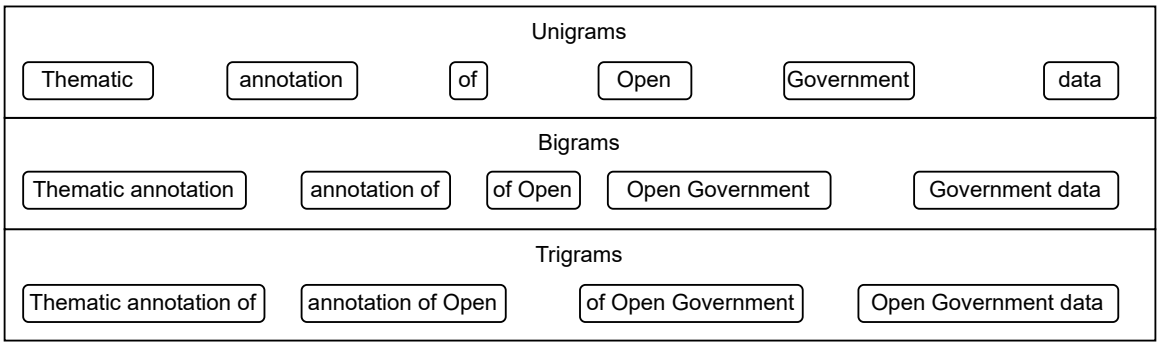

Fig. 3.5 An example of unigrams, bigrams and trigrams for the sentence “Thematic Annotation of Open Government Data”

N-grams do not require any typical type of calculation performed using equations or formulas. Basically, N-gram features are constructed by calculating the word sequences in the corpus. These number of N-grams affect the unique number of features fed into the final model training. For example, unigram features would be fewer in number than if we combined unigram with bigram features. As the number of N-grams increases, the vector length (sparsity) increases, which increases the space and time complexity of the model training. On the other hand, there is not a standard way to decide what value of N for the N-grams will work optimally.

Model training

The critical step of the workflow is the training of models where the system learns from the labelled cases (datasets annotated with themes). The models recognise patterns and properties that divide various classes and this allows them to generalise the problem and classify new, unlabelled datasets.

We have used the One-vs-Rest (OvR) classifier with three machine learning techniques: Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Support Vector Machines (SVM). The OvR

classifier is used in machine learning for situations involving multiple class classifications, because it partitions multi-class problems with more than two classes into a series of binary classification tasks. Each class has its own binary classifier that has been trained to distinguish it from the others. However, data class imbalances may hurt its performance on under-represented groups [112].

Our underlying problem is to classify the open datasets not just into multi-class but also into multiple multi-class themes. For instance, a dataset in the European data portal could be classified into more than one theme of the 13 proposed themes. The OvR classifier can help us to train MNB, SVM, and LR for multiple, multi-class classification.

3.3.3 Predicting the closest theme of a dataset based on word/sentence embeddings

The objective of this component of the thematic framework is to predict the correct theme when an annotated corpus is unavailable or the datasets in this corpus are not properly annotated. Figure 3.6 shows the proposed method for the prediction of the closest theme of a dataset based on the similarity between the word/sentence embeddings representing a dataset and its potential associated themes.

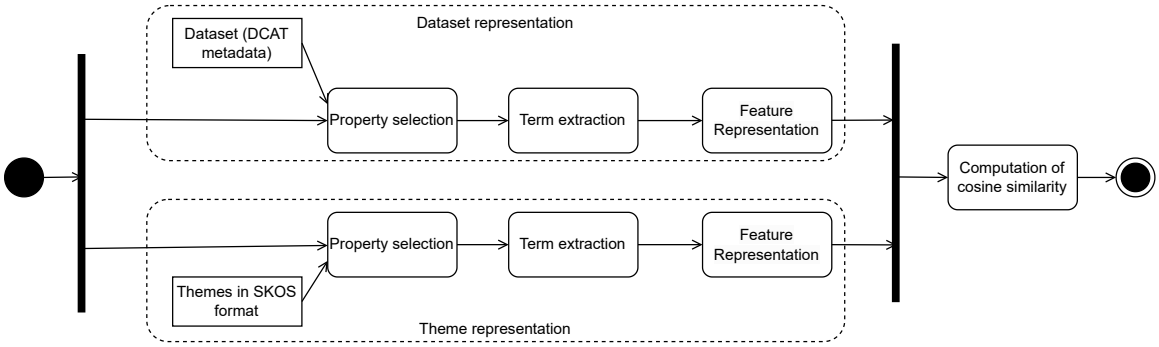

Fig. 3.6 Proposed method for the prediction of the closest theme of a dataset based on the similarity of word/sentence embeddings

We propose the use of word and sentence embeddings for representing datasets and themes instead of a bag-of-words representation because this allows us to represent similar input texts as close points in a vector space with a number of dimensions much lower than the number of dimensions needed in vector spaces generated when using bag-of-words representations. Furthermore, the training data and neural networks used to generate these embeddings allow us to find similarities between two texts even in the case of not having any lexical matching between the compared texts.

Next subsections explain the process proposed for the property selection, term extraction, feature representation of both datasets and themes. In addition, we describe how we have computed the cosine similarity between the dataset and theme embeddings.

Property selection

In the case of datasets, the property selection is similar to the one proposed for an annotated corpus of datasets in section 3.3.2. We assume that metadata is compliant with a DCAT vocabulary and we select the content of *dct:title*, *dct:description* and *dcat:keyword*. In addition, we considered two possibilities for generating the input text of a dataset: the concatenation of the three properties, and just the concatenation of title and description. We considered the second possibility because some sentence embeddings may not work properly if we create sentences including disconnected keywords.

In the case of themes, we assume that the list of themes is provided in SKOS format and that each theme is represented with an SKOS concept having an associated preferred label (*skos:prefLabel*) and a definition (*skos:definition*). Therefore, the input text to be processed for each theme is the concatenation of its preferred label and its definition in English.

Term extraction

The next step in the proposed method is the extraction of tokens from the input texts for datasets and themes and the generation of terms. In this case we considered a basic Normalisation consisting in the removal of special characters and the transformation of text to lower case. It must be taken into account that the word/sentence embedding representation avoids implicitly the appearance of non-common English words. In addition, in some cases we also considered the removal of stop words.

Feature representation

For the representation of datasets and themes, we considered different possibilities of embeddings:

- Sum of GloVe word embeddings: The terms extracted from the input text of each dataset and theme are converted into a word embedding according to the Global Vectors for Word Representation (GloVe)[4] using vectors of 200 dimensions. To represent the complete input text, this alternative computes the sum of the word embeddings.
- Average of GloVe word embeddings: This alternative is similar to the previous one, but in this case the complete input text is represented with the average of the word embeddings.
- BERT sentence embeddings: This alternative transforms the input text into a vector representation of the sentence by applying the pretrained Bidirectional Encoder Representations from Transformers (BERT) [113].
- HuggingFace sentence embeddings: This alternative transforms the input text into a sentence embedding thanks to HuggingFace representation [114, 115].

[4]https://nlp.stanford.edu/projects/glove/

Cosine similarity

To find the closest themes that can be associated with a dataset, we propose the use of the cosine similarity distance, which is typically applied to compute the ranking of results in information retrieval systems using a vector space model for representing documents and queries. Equation 3.1 shows the customisation of this cosine distance to our context: the similarity between a theme T and a dataset D is equivalent to the cosine of the angle formed by the vectors $\vec{T}$ and $\vec{D}$ corresponding to their word/sentence embeddings. The similarity is therefore a real value between 0 (least similarity) and 1 (most similarity), which is computed dividing the scalar product of the embedding vectors by the product of their norms.

$$Similarity(T,D) = Cosine(\vec{T},\vec{D}) = \frac{\vec{T}\cdot\vec{D}}{\|\vec{T}\|\|\vec{D}\|} \tag{3.1}$$

As the similarity is computed for all datasets that require annotation and all the candidate themes, the output of this step is a matrix where each row represents a dataset and the similarity of each theme is provided in the columns. This way we can generate a rank of associated themes for each dataset, and select, for instance, the top 3 themes.

3.4 Experiments and results

This section describes the applicability of the thematic annotation framework to a corpus of metadata records downloaded from *data.europa.eu*, the official portal for European OGD. The implementation of the thematic framework (Python programs and notebooks), together with the data and the associated results, are available in a GitHub repository.[5]

3.4.1 Corpus description

The metadata used in our experiments came from *data.europa.eu*. This portal serves as a centralised access point to open data published by both European Union institutions and member states. The metadata describing the datasets is compliant with the DCAT-AP vocabulary [81] and can be queried through an SPARQL end-point.[6] In July 2022 we developed a harvester program to create a corpus of 29,793 metadata records in RDF format containing title (*dct:title*), description (*dct:description*), theme (*dcat:theme*) and keyword (*dcat:keyword*) properties. Although we downloaded the full list of records but one of the constraints applied to filter the corpus was to have metadata records with at least one associated theme from the list of themes proposed by the European data portal. We also restricted the download to the metadata records declaring the use of English as language, and having at least one title and one description in English.

[5]https://github.com/IAAA-Lab/Thematic_Annotation_of_Government_Data/
[6]https://data.europa.eu/sparql

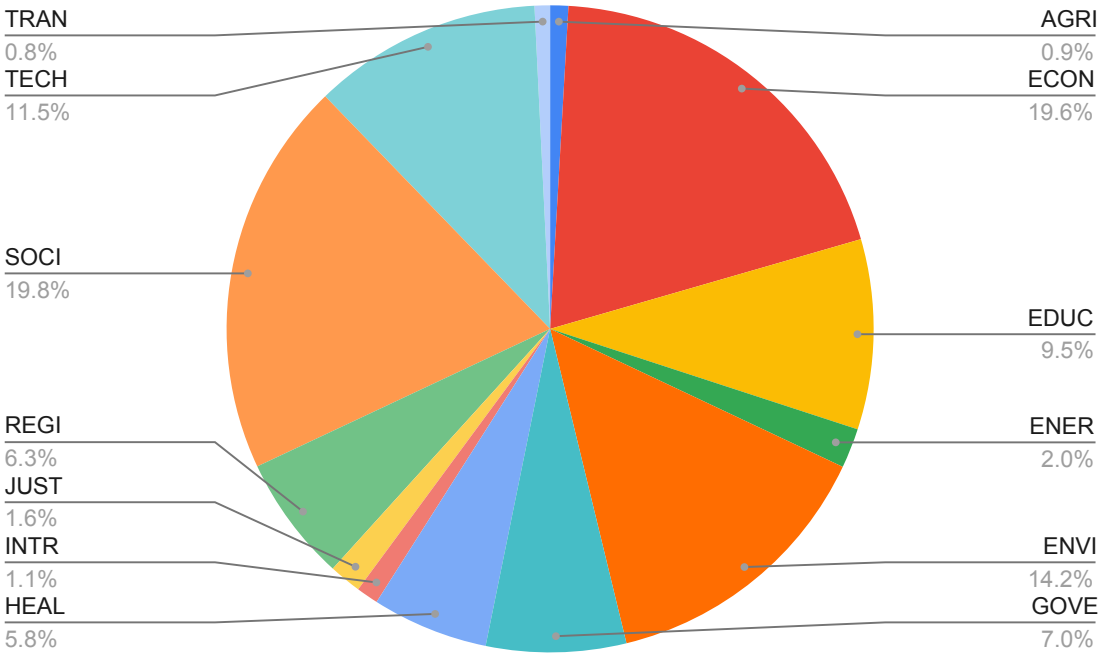

Fig. 3.7 Distribution of datasets across the 13 themes of the European Data Portal.

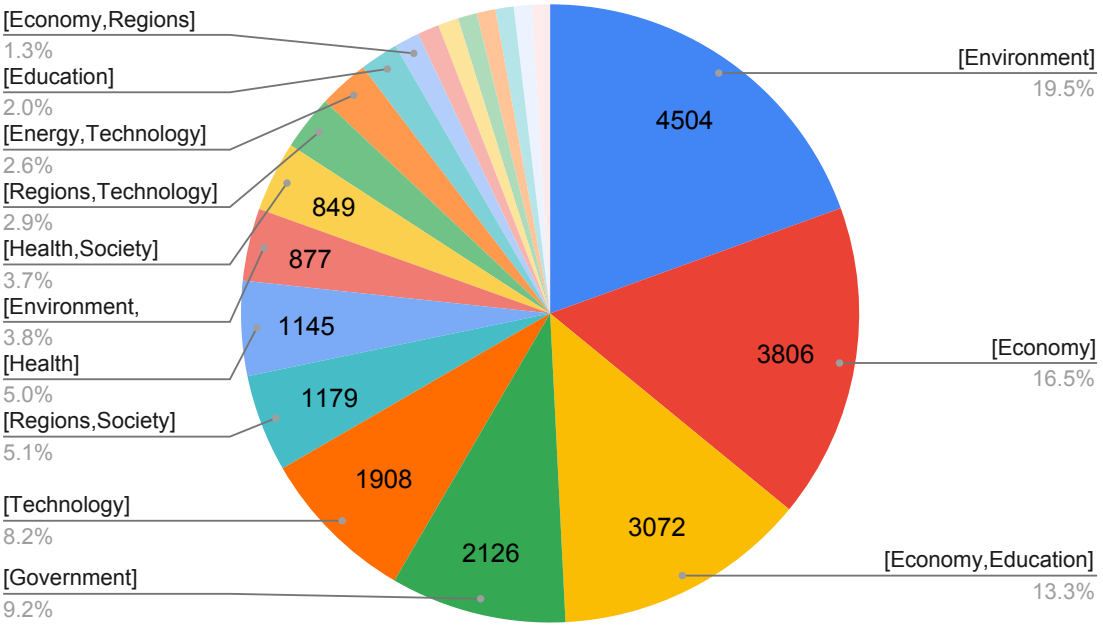

Fig. 3.8 Distribution of datasets with respect to the 20 most frequent combinations of themes.

Figure 3.7 shows the distribution of the datasets in the corpus among the thirteen thematic

categories of the European Data Portal: 'Agriculture, fisheries, forestry and food' (AGRI), 'Economy and finance' (ECON); 'Education, culture and sport' (EDUC), 'Energy' (ENER), 'Environment' (ENVI), 'Government and public sector' (GOVE), 'Health' (HEAL), 'International issues' (INTR), 'Justice, legal system and public safety' (JUST), 'Regions and cities' (REGI), 'Population and society' (SOCI), 'Science and technology' (TECH), and 'Transport' (TRAN). In addition to this, as the datasets may be associated with more than one theme, the pie chart shown in Figure 3.8 illustrates the distribution of the datasets according to the 20 most frequent combinations of themes.

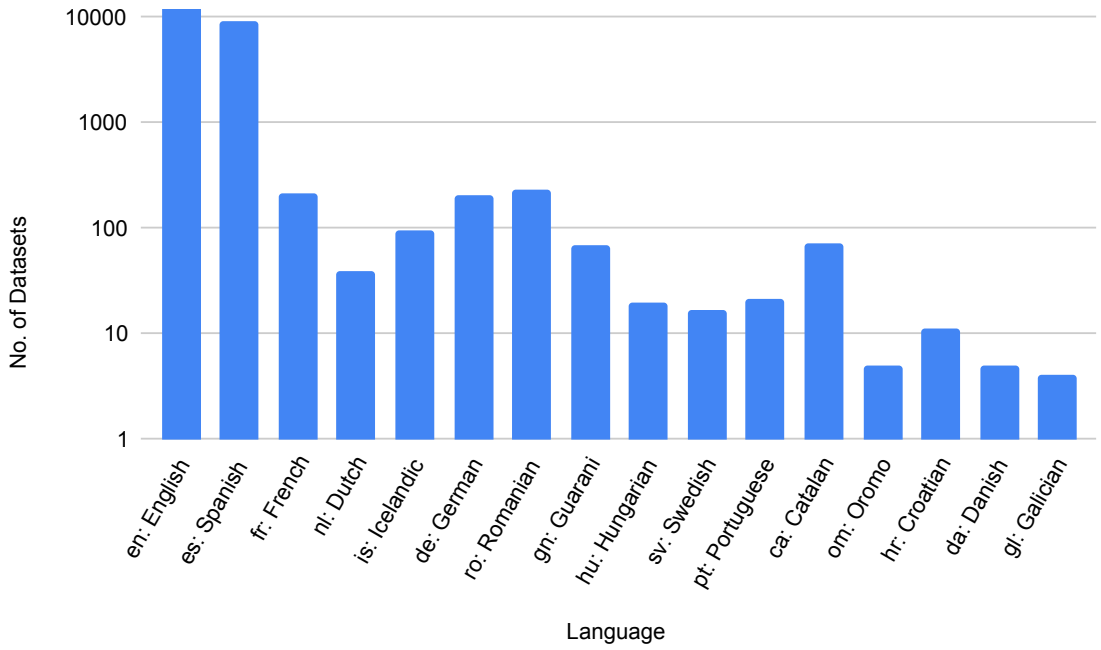

Fig. 3.9 The language distribution of the datasets.

Furthermore, after a manual inspection of the records we realised that a significant number of metadata records had metadata properties with text content in a different language from English. Although metadata records were specifically retrieved declaring the use of English as the language attribute for string literal values, this was not the case for many of the records. Using an API to detect the most likely source language, Figure 3.9 shows the distribution of languages employed in the corpus. This circumstance motivated the translation of the input text into English as a Normalisation process during term extraction for some experiments in Section 3.4.3. In a similar way, it was noticed that 18,633 words found in the input text of the corpus were not recognised as common English words, and this motivated a tailored Normalisation level for some experiments in Section 3.4.3 to remove these noise words.

3.4.2 Results of thematic classification correctness evaluation

Considering an AQL of 5% of errors in the thematic classification of corpus datasets, the Limiting Quality (LQ) that must be applied to a corpus that is manually inspected is thrice the AQL. As the table of ISO 2859-2 standard [116] defines the relationship between the lot size of the corpus and the selected LQ does not provide a value for 15%, the most approximate value of 12.5% must be selected.

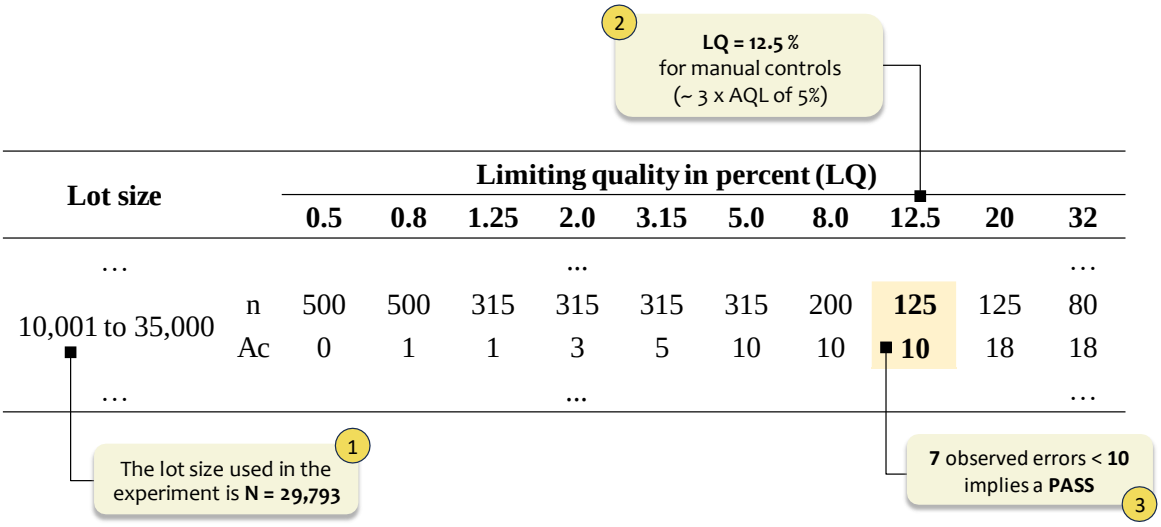

Fig. 3.10 Results of thematic classification correctness for a lot size of 29,793 records and LQ of 12.5%.

Figure 3.10 describes the process followed to identify the size (n) of the sample that must be evaluated and the maximum number of errors (Ac) that can be accepted in Table A of ISO 2859-2 taking into account that our corpus consists of 29,793 datasets and we want an LQ of 12.5%. Following this, a sample consisting of 125 records was randomly selected. The sample was then evaluated by two experts, who manually visited the dataset resources and assigned to them between one and three related themes according to the perceived content of the dataset and the text contained in title, description and keyword properties. As indicated in section 3.3.1, the cases where none of the initially assigned dataset themes matches with one of the themes assigned by the experts were considered as errors. Upon this criterion, only 7 cases of incorrect classification were counted. As the number of errors was below the Ac threshold of 10 items, the quality control was passed and the thematic classification of the corpus was considered correct.

3.4.3 Results of automated supervised classification

This section presents the experiments performed to build models for the automatic thematic annotation of datasets using the proposed approach in section 3.3.2 for supervised classification. Tables 3.1, 3.2 and 3.3 show the description of the 54 experiments that were performed considering different variants for input datasets, term extraction, feature representation and use of machine learning techniques:

- The *Input* column indicates the alternatives for used for the input records. The default alternative is the use of the annotated corpus of 29,793 records (denoted as *core*). A second alternative, as proposed in section 3.3.2, was the incorporation of artificial datasets generated from associated themes in GEMET and UNESCO thesauri. Following this approach, we generated 686 additional records and an extended corpus of 30,479 (denoted as *extended*).

- The *Term extraction* column indicates the alternatives for term extraction: the *basic*, *translation* and *tailored* Normalisation levels explained in section 3.3.2.

- The *Feature representation* column indicates the alternatives for feature representation: the use of unigrams (*uni*); the combined use of unigrams and bigrams (*uni+bi*); and the combined use of unigrams, bigrams and trigrams (*uni+bi+tri*). The number of dimensions in the vector representation of each alternative is shown in the tables within parentheses.

- The *Classification technique* column indicates the alternatives for machine learning classification techniques (*LR*, *MNB*, or *SVM*). As indicated in section 3.3.2, we used the OvR classifier to solve our multi-class classification problem. When making a prediction, all available binary classifiers are applied to the input data until one produces a confidence score high enough to be considered trustworthy. This method simplifies complex multi-class problems into binary decisions, and it enhances classification performance by focusing on differences between classes [112].

Table 3.1 Experiments and results for automated supervised classification: *core* input.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
1	core	basic	uni	LR	0.8881
2			(35,891)	MNB	0.7710
3				SVM	0.9365
4	core	basic	uni+bi	LR	0.84114
5			(290,600)	MNB	0.57377
6				SVM	0.8995
7	core	basic	uni+bi+tri	LR	0.8073
8			(659,880)	MNB	0.5137
9				SVM	0.8503
10	core	basic	uni	LR	0.8854
11		+ translation	(25,622)	MNB	0.7817
12				SVM	0.9355
13	core	basic	uni+bi	LR	0.8417
14		+ translation	(265,399)	MNB	0.5472
15				SVM	0.8920
16	core	basic	uni+bi+tri	LR	0.8082
17		+ translation	(619,227)	MNB	0.4969
18				SVM	0.8394

Tables 3.1, 3.2 and 3.3 also include a column with the accuracy obtained for each experiment. This accuracy is computed according to equation 3.2 considering the number of true positives (TP),

false positives (FP), true negatives (TN), and false negatives (FN).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.2}$$

Table 3.2 Experiments and results for automated supervised classification: *extended* input; *basic* and *translation* Normalisation.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
19	extended	basic	uni	LR	0.8751
20			(30,443)	MNB	0.7654
21				SVM	0.9226
22	extended	basic	uni+bi	LR	0.8288
23			(280,834)	MNB	0.5656
24				SVM	0.8827
25	extended	basic	uni+bi+tri	LR	0.7959
26			(645,748)	MNB	0.4992
27				SVM	0.8325
28	extended	basic	uni	LR	0.8721
29		+ translation	(26,497)	MNB	0.7623
30				SVM	0.9217
31	extended	basic	uni+bi	LR	0.8298
32		+ translation	(273,849)	MNB	0.5346
33				SVM	0.8754
34	extended	basic	uni+bi+tri	LR	0.7935
35		+ translation	(638,605)	MNB	0.4769
36				SVM	0.8254

SVM is the machine learning technique that performed best for all the variants incorporated in the experiments related to the input, term extraction and feature representation. We also computed the confusion matrices for each theme. For instance, Figure 3.11 shows the confusion matrices for each individual theme in the best experiment, i.e. experiment 3 in Table 3.1.

Figure 3.11 also includes the precision, recall and F1 evaluation metrics according to formulas in equation (3.3):

$$Precision = \frac{TP}{TP+FP}; Recall = \frac{TP}{TP+FN}; F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.3}$$

In addition, Figure 3.12 shows the curve known as the receiver operating characteristic (ROC) for experiment 3. The ROC curve is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR). Whereas TPR reflects the percentage of cases that were properly labelled as positive, FPR reflects the proportion of instances that were incorrectly classified as positive (see formulas in equation 3.4). It can be observed that the area under the curve (AUC) of the ROC curve is close to the maximum value for practically all of the themes, which demonstrates that the configuration of the SVM experiment has a high probability to assign correctly the theme of a dataset.

$$TPR = \frac{TP}{TP+FN}; FPR = \frac{FP}{FP+TN} \tag{3.4}$$

Table 3.3 Experiments and results for automated supervised classification: *extended* input; *tailored* Normalisation.

#	Input	Term extraction	Feature representation	Classification technique	Accuracy
37	extended	basic	uni	LR	0.8715
38		+ tailored	(17,371)	MNB	0.7737
39				SVM	0.9152
40	extended	basic	uni+bi	LR	0.8248
41		+ tailored	(222,559)	MNB	0.5242
42				SVM	0.8720
43	extended	basic	uni+bi+tri	LR	0.7909
44		+ tailored	(529,092)	MNB	0.4647
45				SVM	0.8201
46	extended	basic	uni	LR	0.8677
47		+ translation	(12,906)	MNB	0.7696
48		+ tailored		SVM	0.9142
49	extended	basic	uni+bi	LR	0.8236
50		+ translation	(215,844)	MNB	0.4971
51		+ tailored		SVM	0.8632
52	extended	basic	uni+bi+tri	LR	0.7886
53		+ translation	(524,646)	MNB	0.4433
54		+ tailored		SVM	0.8090

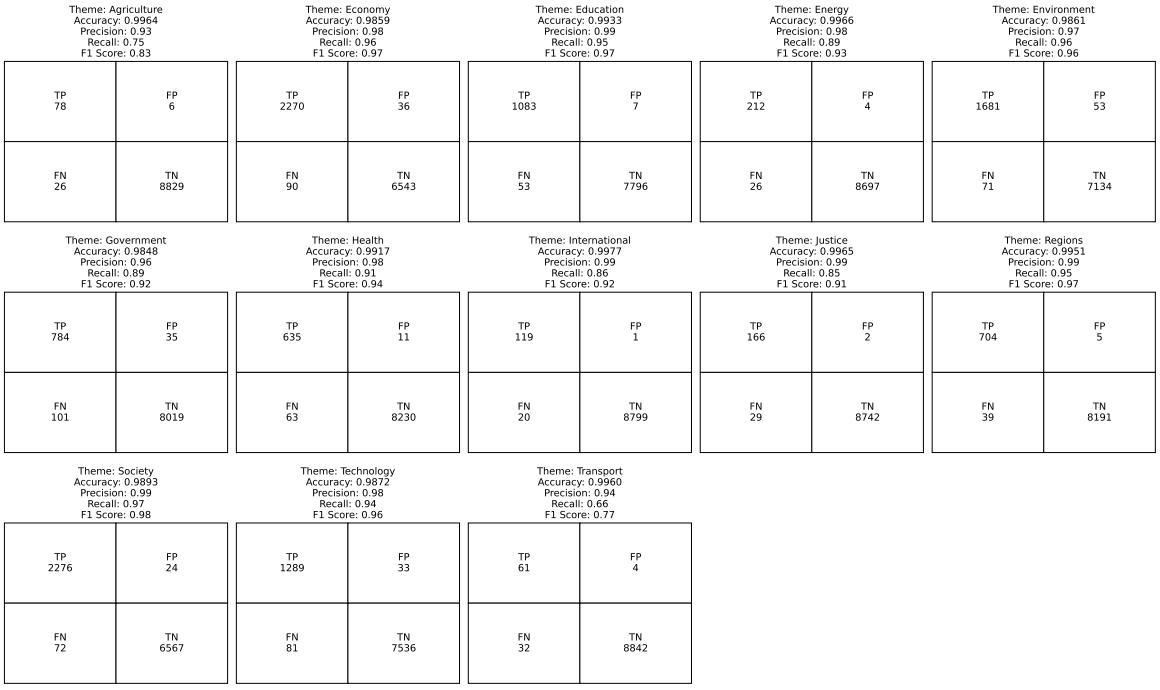

Fig. 3.11 Confusion matrices for all the themes in experiment 3 of Table 3.1(*core* input, *basic* Normalisation, unigram features, SVM, overall accuracy of 93.65%)

3.4.4 Results of theme prediction

This section presents the results of the approach proposed in section 3.3.3 to predict automatically the closest theme according to the similarity between the word/sentence embeddings of the metadata content and the definition of the European Data themes. It is an unsupervised approach to predict

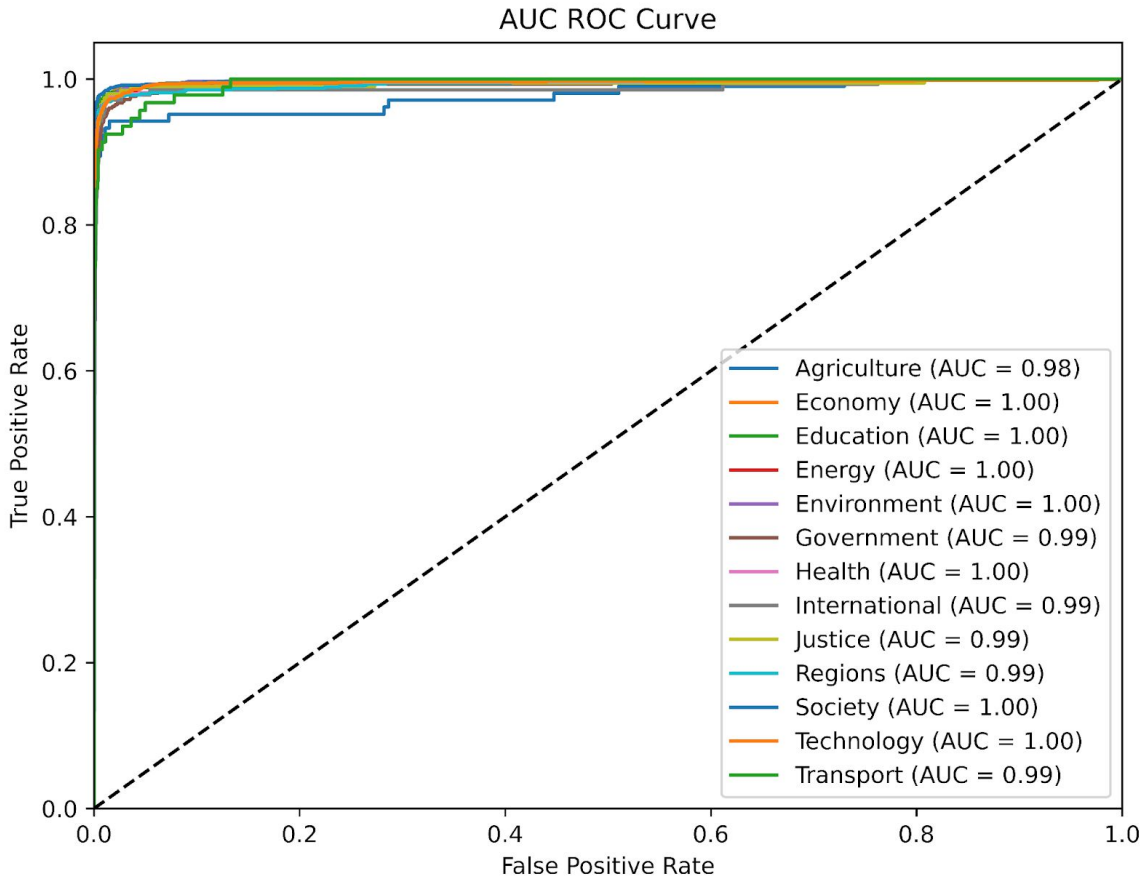

Fig. 3.12 ROC curve for all the themes in experiment 3 of Table 3.1 (*core* input, *basic* Normalisation, unigram features, SVM, overall accuracy of 93.65%)

themes. Table 3.4 shows the description of the 7 experiments that were performed to assign automatically themes to the datasets in our corpus considering different variants for property selection, term extraction and feature representation:

- The *Dataset property selection* column indicates the alternatives for the selection of metadata properties describing the datasets as proposed in section 3.3.3: the concatenation of three properties (*title+description+keywords*) or just the concatenation of title and description (*title+description*). It must be noted that the property selection for themes is not detailed because it is maintained in all the experiments: the input text is the concatenation of the preferred label and definition of each theme in English.

- The *Term extraction* column indicates the alternatives for term extraction explained in section 3.3.3: *basic* Normalisation and the additional process of *stop word removal* in some cases.

- The *Feature representation* column shows the alternatives that have been used for feature

representation as proposed in section 3.3.3: *GloVe sum* indicates the use of GloVe word embeddings and the representation of the full text as the sum of the embeddings of each word in the text; *GloVe average* indicates the use of GloVe word embeddings the representation of the full text as the sum of the embeddings of each word in the text; *BERT* indicates the use of BERT sentence embeddings; and *HuggingFace* indicates the use of HuggingFace transformers for sentence embeddings.

Table 3.4 Experiments and results for theme prediction

#	**Dataset property selection**	**Term Extraction**	**Feature Representation**	**Agreement score**		
				Top 1	**Top 2**	**Top 3**
1	title + description + keywords	basic	GloVe sum	0.4127	0.5394	0.6536
2	title + description + keywords	basic	GloVe average	0.4397	0.5770	0.6865
3	title + description + keywords	basic + stopword removal	GloVe average	0.4680	0.6186	0.7253
4	title + description	basic	BERT	0.2480	0.3360	0.3920
5	title + description + keywords	basic	BERT	0.1908	0.3132	0.4109
6	title + description	basic	HuggingFace	0.3920	0.6320	0.7120
7	title + description + keywords	basic	HuggingFace	0.5023	0.6734	0.7456

In order to have an orientation about the appropriateness of the predicted themes by the different experiments, we compared the top three themes (ranked by decreasing cosine similarity distance) with the original dataset themes assigned in the annotated corpus. Table 3.4 includes an agreement score for the top 1, top 2 and top 3 themes. This agreement score measures the proportion of matches between the top 1/2/3 themes and the assigned themes in the corpus. For instance, if a dataset was originally annotated with the "society" theme, and "society" is the third more relevant theme assigned, this is a match for the top 3 agreement score. Equation 3.5 shows the formula for computing the agreement score of a *corpus* and top n predicted themes: $themes(d)$ stands for the function that returns the themes assigned to a dataset d in the *corpus*; $predicted\_themes(d,n)$ stands for the function that returns the top n predicted themes of a dataset d; and $|corpus|$ is the number of datasets in the corpus.

$$Agreement\_score(corpus,n) = \frac{\sum_{d \in corpus} \begin{cases} 1, & \text{if } themes(d) \cap predicted\_themes(d,n) \neq \emptyset \\ 0, & \text{otherwise} \end{cases}}{|corpus|} \tag{3.5}$$

It can be observed that the best agreement score is obtained with the HuggingFace Transformer for sentence embeddings in experiment 7 and comparing the 3 best ranked themes with the original

dataset themes. Although the obtained agreement score of 0.7456 is lower than the accuracy obtained with the best experiment for classification models in section 3.4.3 (0.9365 in experiment 3 with SVM), these numbers are not comparable. In the component for thematic prediction of our framework for thematic annotation, we are assuming that the initial thematic annotation is not perfect (or not existent) and we try to identify the closest theme according to the similarities of the language models used to generate the embeddings of dataset metadata and theme descriptions. In some cases the definition proposed by the Publication Office of the European Union for a theme [109] consists of a reduced number of words (sentences) and may not encompass all the possible aspects that the datasets associated with this theme may cover. For instance, the 'Health' theme is defined with just three sentences: *"This concept identifies datasets covering the domain of health. Health is a state of physical, mental and social well-being in which disease and infirmity are absent. Dataset examples: COVID-19 Coronavirus data; European Cancer Information System."*. Larger texts would generate an embedding vector representation with a better alignment with all the aspects covered by a dataset theme. The experts that evaluated the thematic classification correctness reported in section 3.4.2 assessed that there were not more than 5% of errors in the classification, but their manual annotation of themes was not constrained by the short definitions of themes.

3.5 Discussion

After assessing that the thematic classification correctness of the annotated corpus had an acceptable quality with less than 5% errors, we analysed the feasibility of different machine learning techniques to build models for automatic thematic annotation of datasets. In the experiments we cleaned the DCAT-based metadata text by applying several text processing techniques, which also included the translation of false English text and the removal of non-common English terms. With respect to feature representation, we also tested the combined use of unigrams, bigrams and trigrams. Supervised classification techniques such as Logistic Regression and Naive Bayes, as well as Support Vector Machines (SVM), showed effectiveness in classifying datasets with themes using titles, descriptions and keywords, being SVM the technique having the highest accuracy of 93.65%.

Our thematic annotation framework also considers the possibility of not having an annotated corpus or not counting on a perfect annotated corpus with themes. In this case we propose a representation of texts derived from metadata and theme descriptions in terms of word or sentence embeddings. To predict the themes closer to a dataset, we compute the cosine distance between the embedding representations of the dataset and the candidate themes. After doing the experiments with the same sample of datasets extracted from *data.europa.eu* and different techniques for word embeddings (GloVe) and sentence embeddings (BERT and HuggingFace Transformers), we concluded that HuggingFace Transformers seem the best approach. The predicted themes have a high agreement score (74.56%) with respect to the original themes assigned in the European data portal. This is not comparable with the accuracy obtained with SVM and an annotated corpus, but it must be noted that the texts defining the European data themes are short (just 3 sentences in some cases). Despite using

language models for obtaining the embeddings, short texts may not reach a vector representation encompassing all the dataset topics that the theme may cover. The use of a larger text for the definition of themes would probably lead to a higher agreement with the original themes assigned to the datasets in the corpus.

3.6 Summary

This chapter has presented a supplier-driven approach for inclusiveness by means of a framework for the thematic annotation of OGD, which has been tested against a representative sample of 29,793 datasets from *data.europa.eu*, a portal that aggregates datasets (and their associated metadata) harvested from both the member states of the European Union and the European institutions. Our proposed framework has theoretical and practical implications for the field of OGD by providing a systematic tool for thematic annotation. Theoretically, it enhances the basic understanding of the thematic annotation topic and how structured thematic annotations can support data findability. Practically, the framework can be adapted as a plugin for well-known open data software platforms such as CKAN, DKAN or Socrata. Such integration would primarily imply the definition of the themes that should be used for classification in each portal.

This study is also subject to several limitations. Although *data.europa.eu* stands as one of the largest open data government portals and serves as a hub for the national OGD portals of the European countries, it is crucial to recognise that the categorisation of datasets may heavily reflect the biases of the entities responsible for publishing them. Exploring the relationship and compatibility of thematic classification schemes employed in OGD portals across other regions could enhance the representativeness and generalisation of automated thematic classification algorithms. Furthermore, the quality of metadata presents another significant constraint. The prevalence of datasets nominally labelled in English but containing text in other languages exemplifies the noise inherent in the training data. Consequently, sensitivity to such noise emerges as a pertinent consideration in the algorithmic approach to the thematic classification of datasets.

In addition, it must be observed that our framework has not been integrated and tested within the scope of an open data portal with end users. Our framework is not aimed at being directly executed by end users interacting with open data portals, but to be integrated during the ingestion process of datasets in a data portal. In order to simulate the thematic annotation during this ingestion process, this work reports experiments whose results have been evaluated in terms of relevance measures, which are employed in the information retrieval discipline to estimate user satisfaction. However, we acknowledge that techniques like A/B testing [117] could be used to verify with end users if an open data portal incorporating this innovation during the ingestion process is better accepted than the portal without the innovation. For instance, we could compare the number of clicks on the first hits returned by both portals with thematic searches.

Finally, it must be noted that the metadata text used for automatic annotation is usually short, and sometimes not accurate. Therefore, as future work we would like to explore if the information related

to the application schema of the different distributions of datasets can help us to improve the automatic thematic classification of datasets. Available distributions in machine readable formats such as CSV or RDF can provide in some cases meaningful names of thematic attributes of the dataset content. Even in the case of RDF (graph data), these attributes are usually selected from well-known vocabularies, and this may be used to infer links with the themes that can be assigned automatically. Moreover, the impact of data policies on thematic annotation practices and the user experience in accessing and utilizing annotated data could be more deeply investigated to understand how regulations influence the effectiveness of open data ecosystems.

CHAPTER 4

A CONCEPTUAL DEFINITION OF ROUNDTRIP FEEDBACK IN OPEN GOVERNMENT DATA PORTALS

An increasing number of countries across the world have started making Open Government Data (OGD) accessible to the general public freely and in various formats via a variety of open data portals [118, 119]. The main objective is to enhance the transparency of government operations and actions and to promote the notion of generating value from OGD [120, 121]. In order for users to be successful in achieving this objective, users must be able to grasp the narratives behind the published data and extract knowledge from them [122]. Data literacy and experience are essential for extracting insights from data [123, 124], which means we still have a long way to go before we achieve the goal of user engagement within open data portals. Therefore, one possible way to bridge the gap in user engagement and improve the data quality is to offer the heterogeneous feedback mechanisms for the open data portal (ODP).

An open data ecosystem is defined as the network of interconnected stakeholders, institutions, infrastructures, and technologies that support the publishing, access, use, and reuse of open data, emphasizing the evolving relationships among data providers, users, and the systems that enable the data flow [46]. The feedback mechanism is a crucial part of the open data ecosystem, as it allows users to provide their opinion in the form of feedback on the data and also for the portal itself. This feedback can be used to improve the quality of the data and the portal, and to make the data more accessible and usable for users [48]. Similarly, it is also important for the open data portal administrators to know that what are the most effective methods for providing feedback to enhance the quality of data [125].

Feedback mechanisms have the potential to enhance government operations and monitoring methods, thereby advancing openness, accountability, and citizen engagement [36]. Existing feedback methods in open data ecosystems are not meeting the different demands of the stakeholders, preventing optimum usage and reuse of data and also there is a gap in knowing how feedback is handled and integrated into open data systems to encourage data reuse and improve the data quality [24, 54, 126–128]. Hence, it is important to note that there is a need to identify the interests of various stakeholders to modernise these feedback mechanisms in open data portals. Addressing these concerns is essential for

improving open data accessibility, usefulness, and relevance and boosting data-driven decision-making in numerous fields.

The main motivation for this work is to understand the current challenges of providing feedback. For this analysis, we propose a methodology based on three main stages. First, we define a conceptual framework for feedback scenarios in open data portals where two main stakeholders interact: governments in their role of data publishers, and final users with a generic profile. This framework is designed to define the feedback mechanism, how it operates, facilitating continuous improvement and adaptability in open data initiatives. Our aim is to provide a structured understanding of feedback scenarios, addressing both the needs of data publishers and users. Second, this conceptual framework is then improved through the analysis of 5 case studies using structured questionnaires. By employing a mixed-methods approach and incorporating the data from the case-studies of 5 open data portals, we examine how feedback scenarios are implemented and maintained across the portals. This validation strengthens the theoretical foundations of our feedback model. Thirdly, focusing on the identification of input and output channels for feedback, we have extrapolated the analysis of feedback mechanisms across the 29 open data portals, including 26 European open data portals. This broader analysis allows us to generalise the feedback scenario for a wider range of open data initiatives, highlighting commonalities and divergences across different portals.

The methodology adopted in this chapter follows a Design Science Research (DSR) approach [129], which is particularly suited to developing and evaluating artifacts such as frameworks and models in information systems research. The proposed framework is therefore constructed and validated through case studies, providing a structured process for creating a solution-oriented contribution. Case study research is effective for exploring "how" and "why" concerns, since it provides an in-depth examination of research processes across many contexts [130–132]. Moreover, this contribution aims to provide a user-driven approach for inclusiveness, i.e. transforms the users into the main actors of open data ecosystems. Furthermore, this chapter addresses research question RQ3. The rest of the chapter is structured as follows. Section 4.1 provides a review of the research background and relevant literature. The methodology that includes the framework of the working model is presented in Section 4.2. Section 4.3 presents the findings and results of this research, which are discussed in Section 4.4. Finally, Section 4.5 provides some concluding remarks, together with a description of the limitations of this research work and some ideas for future work.

4.1 Related Work

This section compiles some relevant works related to feedback and its role in Open Data ecosystems. Although this chapter does not aim to provide a systematic review, we followed some of the steps in traditional methodologies for systematic reviews such as SALSA and PRISMA (as explained in the introductory part of chapter 2) to find relevant resources. These steps included: identifying databases to search (Web of Science, Scopus, ScienceDirect, and Google Scholar); defining the search strategy (using terms like “feedback loop”, “open data feedback”, “data portal feedback” and “feedback

channels"); establishing inclusion/exclusion criteria; and finally synthesizing the contributions of the work with respect to our research. After applying this process, we were able to classify the contributions into three categories: (1) works justifying the need for and importance of feedback; (2) works identifying different channels for providing feedback; and (3) works identifying the issues subject to get feedback.

Due to the expansion of OGD initiatives and the efforts of governments to make open data ecosystems more user-centric [5, 29, 133], feedback mechanisms have been a longstanding topic in the domain of OGD. Since the middle of the last decade several researchers have highlighted that open data infrastructures should have a feedback mechanism in place to facilitate the communication between OGD providers and users [29, 30] and enhance their transparency [119, 134]. In general, the availability of feedback channels is considered as a relevant factor influencing the increase in the quality of OGD portals [60]. Moreover, these feedback channels should be bidirectional. According to Janssen et al. to engage stakeholders in an effective way in public discussions around the OGD, policy-making, and creation of application services, it is essential to have a communication and interaction channel establishing feedback loops between data publishers and data users [32]. This allows consumers to rate open data and the providers to respond to their comments. Emphasizing this bidirectional character of feedback, Purwanto et al. mention feedback in open data initiatives as a main citizen engagement factor with the availability of feedback channels and means of communication among data users and publishers, including follow-up communication [31]. In addition, it is acknowledged that open data ecosystems are systems in constant evolution as user needs and preferences change over time. In order to support a sustainable development of the infrastructure associated with these ecosystems, feedback processes are essential to track this change of user behaviour [135].

With respect to the diversity of feedback channels, Máchová et al. have proposed a framework for evaluating the usability of open data portals where the availability of feedback channels is considered a relevant criterion [13]. Among the options available for users to submit feedback they identify discussion forums, contact forms, user ratings, dataset comments and social media. In particular, social media is acquiring an increasing relevance in the last few years to increase the communication between data publishers and users [136, 137]. Although social activity is mostly originated by data publishers and creators for dissemination purposes [138], social networks are an independent forum where any stakeholder may share their perspectives. In addition, more dedicated channels can be employed to address specific objectives. For instance, the European Data Portal performed surveys and interviews to assess the user experience with the portal [139]. Whereas survey questionnaires allow a quick response on a specific issue from a broad audience, interviews with individuals can help to obtain detailed insights. For instance, Zhang et al. used interviews as a methodology for identifying incentive mechanisms in the implementation of OGD [140]. About the issues communicated by users through feedback channels, Zuiderwijk et al. have investigated the design of OGD infrastructures where interaction mechanisms are a key element [141]. By designing these interaction mechanisms, they identify the main topics of the issues reported by users: requests for new datasets, requests on existing datasets (e.g., errors or lack of the desired formats), feedback to policy makers, information

related to a dataset (e.g., publications and applications based on the dataset), data use cases, or technical experienced with the use of the infrastructure. Máchová et al. and Nikiforova have also confirm these topics in the results of their assessment of open data portals [142, 143]. The only remarkable difference is that apart from the reporting of use cases about the exploitation of data, they also consider simple dataset ratings or view/download statistics.

4.2 Research Methodology

In this section, we present the research methodology that we used to conduct this chapter for the thesis. Figure 4.1 highlights the details of our proposed methodology. Our methodology is organised into three separate stages, conceptual definition of the feedback scenario; instantiation through case studies; and extrapolation to other open data portals. Each stage builds on the one that came before it to develop an in-depth comprehension of the feedback mechanisms that are present in open data portals. The following subsections explain these stages.

4.2.1 Definition of the Conceptual Scenario of Feedback Mechanisms

Initially, a literature study was conducted with the objective of creating a basic and conceptual model of feedback processes within the framework of open data. This conceptual model was refined during the second phase through case study analysis. This way of proceeding is inspired by the design science research methodology. According to Venable et al., [144], this methodology is the "research that invents a new purposeful artefact to address a generalised type of problem and evaluates its utility for solving problems of that type". Within the context of our work, the conceptual scenario acts as the artefact whose utility is evaluated through the instantiation of case studies and the extrapolation to other open data portals.

4.2.2 Instantiation through case studies

In this phase, we designed questionnaires for the administrators of the selected national open data portals, aligning them with our conceptual scenario of the feedback mechanism. Apart from preparing the questionnaire we need to define the target audience of our survey, i.e., which national data portals will receive our survey and how we can contact them. During this phase, the objective was to study the functioning of feedback mechanisms in practice and to discover the elements that influence the efficiency of certain feedback mechanisms. Our assumption is that we can improve the original conceptual model by incorporating the insights that we gained from these real-world implementations. This enabled us to update the model so that it accurately reflected the feedback implementation and user engagement methods and systems in open data contexts.

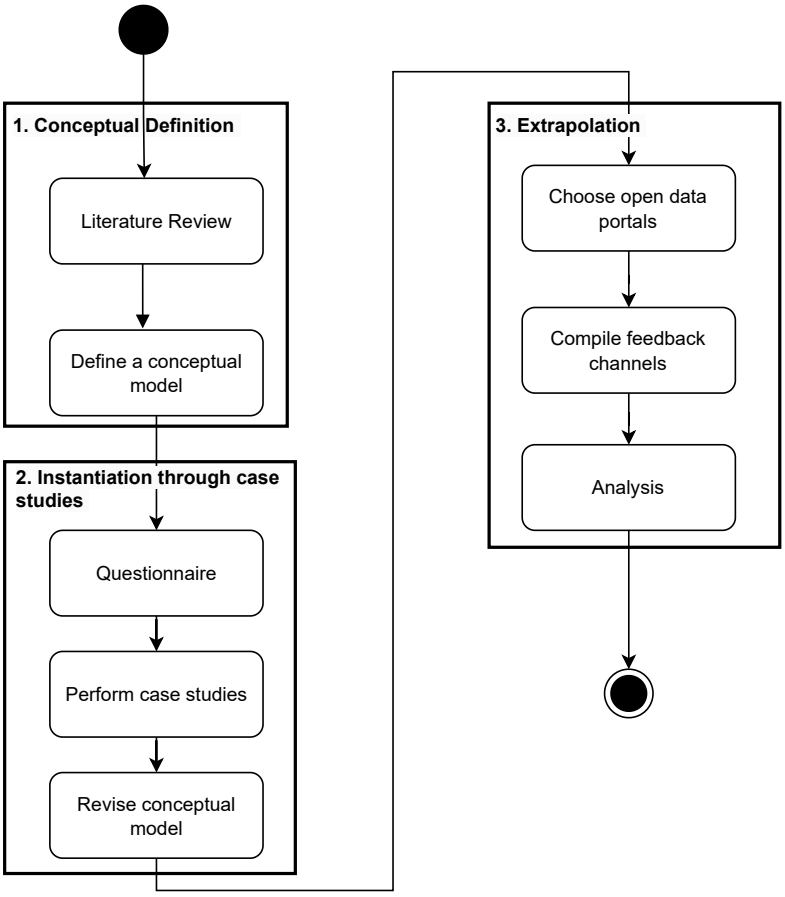

Fig. 4.1 Proposed Research Process for feedback mechanism

4.2.3 Extrapolation to other Open Data Portals

Desk-based research must be conducted as part of the third and final phase of this study, with the aim of extrapolating the results from the case studies to a wider variety of open data portals. For this purpose, it is necessary to analyse the portal features related to feedback that are accessible to the public through various open data portals. The goal is to determine whether the modelling of feedback scenarios could be applied more generally across different contexts. In addition, it must be noted that as it is impossible to investigate all the steps of a feedback scenario without the direct information obtained by the administration staff of open data portals, this third stage is focused on analysing the feedback flows between data users and providers through the identification of input channels (feedback requests sent by users to data publishers) and output channels (feedback replies returned by data publishers to users). This stage of the methodology consists of three steps: selection of open data portals, compilation of feedback channels and analysis. Although the selection of open data portals is obviously a personal decision of the experts performing the study, the compilation and analysis steps

can be guided by a partially automated procedure. For the compilation, we must identify different types of input/output channels as shown in Table 4.1, which presents the feedback channels (input, output, and both) identified in this research study for open data portals, along with the technique used to determine them. Input feedback channels are the identified feedback channels through which users can actively provide direct feedback on datasets, services, or portal features. These channels are designed to capture user perspectives, suggestions, and ratings that help in assessing user needs and portal performance. Output feedback channels are primarily used by the portal team to communicate updates, information, and announcements to the users. While users may not provide direct feedback through these channels, they offer indirect insights into user engagement based on how users interact with the shared content. Last, input/output feedback channels serve as both feedback collection points and information-sharing platforms. Users can engage with the platform by both receiving updates and providing feedback through comments or discussions

Table 4.1 Detection Methods for Various Channels

Channel	**Input**	**Output**	**Detection Method**
E-mail	X		Automated: via regex pattern ‘\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z]‘
Feedback form	X		Automated: Identified `<form>` tags with keywords like "feedback" in labels, buttons, or URLs.
Survey	X		Manual
Interviews	X		Manual
Dataset			
Like/Rate/Fav	X	X	Manual
User/Discussion Forum	X	X	Manual
Social Media	X	X	Automated: Detected links for Twitter or Facebook
Blog/News		X	Manual
Newsletter		X	Manual

With the purpose of automating this compilation step, we have prepared a Python-based script web-scraper, available in a GitHub repository [1], that explores the web site of OGD portals and generates a draft list of potential channels, which must be manually verified later by experts.

Finally, we propose the development of a series of artefacts in order to analyze and compare the status of various open data portals. First, we suggest a comparison of OGD portals in terms of the number and diversity of employed channels through a bar chart. Second, it is interesting to compare the frequency of employment of each channel in an OGD portal with an area chart. Third, we can provide a cluster analysis of OGD portals by grouping the portals upon their similarity in the employment of different channels. Fourth, we can provide a fine-grain analysis to investigate if the feedback channels are provided for reporting issues on specific datasets or the portal as a whole by means of a specific bar chart. Last, we can analyse the potential flows connecting input and ouput channels co-occurring in an OGD portal with a Sankey diagram

[1] https://github.com/IAAA-Lab/ODP-Feedback-Mechanisms/

4.3 Experiments and Key Outcomes

This section presents the findings of our research, highlighting key outcomes and trends observed throughout the study. The results are arranged following our methodology, beginning with a comprehensive overview of the conceptual definition of the feedback mechanism and followed by detailed analyses through case studies that offer insights into the conceptual model. This approach is designed to address each research objective. Finally, we conducted an extrapolation to more extensive open data portals to analyse the feedback mechanisms across these portals.

4.3.1 Conceptual Scenario of Feedback Mechanisms

As described in the proposed methodology that we first dig deep the conceptual model of feedback scenarios from the literature and the overall results of this research study. Sieber et al. [17] investigated the importance of civic feedback in open data ecosystems and argued that successful portals need to complete the feedback loop by converting user input into improvements that can be implemented directly. Moreover, the authors provide a criticism of the one-way transmission of data and argued in favor of communication procedures that are bidirectional. Likewise, Veljković et al. [145] proposed a criterion for assessing the feedback scenarios where the authors stress to include the main participatory feedback elements like dataset comments and dataset issues tracker with proper responses.

For this research study, Figure 4.2 presents the general modelling of feedback interaction and depicts the feedback interaction process between the data user and the data publisher inside the portal. The process starts with the data user defining a feedback message, which is then sent through an input channel. Depending on the selected input channel, the message is also broadcasted to other users. The procedure thereafter diverges according to the policies established by different data publishers. If transparency is promoted by data publishers, the received messages are directly published. Otherwise, the feedback undergoes classification, where it is assessed to determine if it is an incident related to the portal or the dataset. For portal incidents, an action is taken directly on the portal, while dataset-related incidents trigger actions on the dataset itself. Additional steps, such as statistical tracking and reporting on use cases, follow as appropriate. Finally, the process concludes with the feedback loop, where an output channel reports back to the original user who initiated the feedback, providing them with updates or outcomes resulting from their input. This feedback interaction scenario shows the foundational working of a feedback mechanism throughout an open data portal. It offers a paradigm for evaluating the integration of feedback into portal operations.

4.3.2 Instantiation through Case Studies

Having the conceptual model of feedback scenario ready, we wanted to see how the various national open data portals are dealing with the feedback and to measure their impact of a feedback mechanism. Therefore, we prepared research questions for the portal administrators of open data portals. This allowed us to get more comprehensive insights and propose data-driven modifications in the feedback

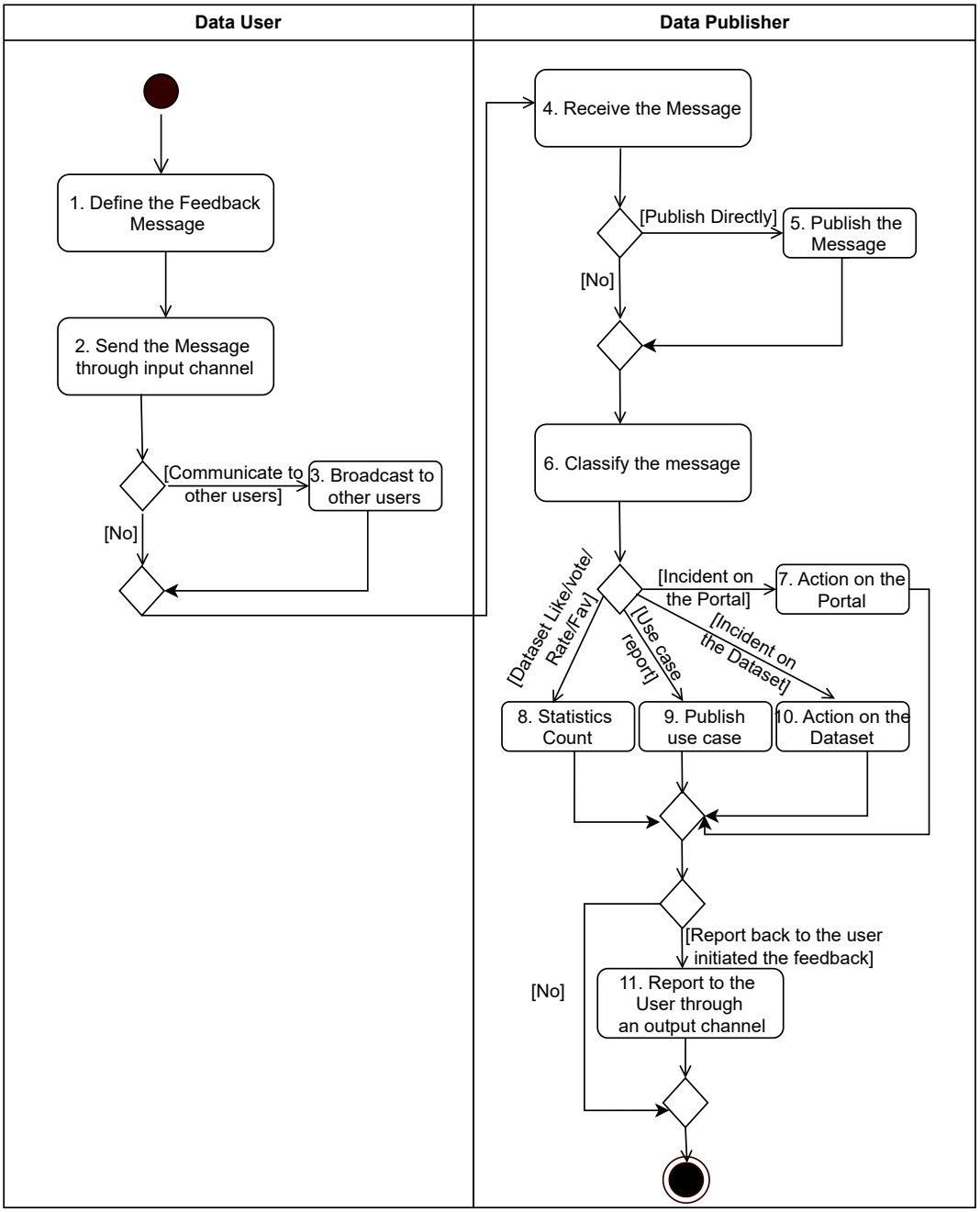

Fig. 4.2 General modeling of feedback interaction

mechanisms for the ODPs. To communicate our inquiry to the administrators of the open data portals, we compiled the official email addresses of the administrative teams and sent them an email with the following questions:

- **Q1**: Which input channels are facilitated to users for providing feedback?
- **Q2**: How does your team review, classify, and prioritise feedback from various input channels?
- **Q3**: How does your team handle feedback classified as an incident related to the portal or datasets? What specific actions are taken based on these classifications?
- **Q4**: How are users informed about actions taken in response to their feedback, i.e. what are the output channels for feedback?

Although we sent our questionnaire to the administrative teams of 29 OGD national data portals (the portals listed in Table 4.3), We got responses from the administrative teams of 5 European open data portals: Estonia, Sweden, Luxembourg, Poland and Spain, totaling five successful replies. We have summarised the answers we received and mapped the responses according to our conceptual model (presented in Figure 4.2) in Table 4.2.

In the case of Estonia, its open data portal offers a straightforward yet underdeveloped system, where feedback is limited in volume and structured mostly around direct dataset concerns. Feedback primarily comes in the form of comments and direct emails from users, addressing dataset-specific errors or accessibility issues. The feedback mechanism lacks broadcasting and categorisation features, limiting how feedback can be structured or classified. Despite this simplicity, the portal shows a commitment to user engagement, enabling users to comment on data quality while allowing portal administrators to assess and respond accordingly.

In the case of Luxembourg, its open data portal lacks a systematic process, handling feedback on a case-by-case basis without formal analysis. Feedback is received through various channels, allowing users to comment on data format, accessibility, and other technical issues, which are forwarded to data producers or portal administrators for response.

Poland open data portal provides a comprehensive platform that integrates various feedback options to enhance user interaction. This portal has developed a robust feedback infrastructure that encompasses multiple input channels and emphasises collaborative feedback review by portal and technical teams. User feedback, gathered through dataset comments, dedicated emails, and new data requests, informs actions aimed at improving data quality and addressing user needs. Although explicit feedback assessment techniques are not highlighted, the portal employs diverse input strategies, benchmarking against European standards to continuously adapt and improve. This approach reflects a comprehensive commitment to understanding user requirements and implementing solutions, demonstrating a forward-looking stance on open data engagement.

The Swedish portal emphasises community engagement through a structured, interactive feedback dashboard, where users can participate in discussions categorised into distinct topics. Eleven themes structure the discussions, and the portal tracks engagement through metrics like participant profiles and post counts, providing insights into active areas, such as data access and API requests. The portal's hierarchical feedback organisation allows administrators to better understand and address user needs across various levels.

The open data portal of Spain stands out as a highly integrated and structured feedback system within the open data landscape. Through multiple channels, including data requests, application sharing, and comments on datasets, users can influence content, and the portal administrators incorporate this feedback into catalogue expansions and continuous improvements. With a variety of specialised support mechanisms and tools, Spain's open data portal not only encourages user engagement but also systematically translates user suggestions into platform updates. This comprehensive feedback system, combined with proactive input collection from both the public and institutional users, enables data.gob.es to dynamically adapt to user needs and maintain relevance in the open data ecosystem.

Table 4.2 Comparative Summary of Feedback Interaction Processes across Five European Open Data Portals

Step No.	**Step Description**	**Estonia**	**Luxembourg**	**Poland**	**Sweden**	**Spain**
1	Define the Feedback Message	Specific questions or dataset concerns	Case-by-case; format, access, updates	Dataset comments, new data requests	Data requests, API issues	Dataset comments, new initiatives
2	Sending the Message	Email, forum, rating	Email, forums, favorite/reuse marking	Surveys, feedback forms, email	Interactive dashboard, comment threads	Feedback/request forms
3	Broadcast to other users	Not applicable	Forum only	Not applicable	Public community forum	Public comments, social media
4	Read the Message	Admins assess issues	Technical team collaborates with data producers	Team and officers assess feedback	Team and community members read and engage	Admins and responsible bodies review
5	Publish the Message	Not usually published	Managed internally, not published	Not public	Immediately published in threads	Visible on dataset pages
6	Classify the Message	By theme (e.g., errors)	Not formally classified	No specific methods	Eleven predefined topics	Categorised (e.g., incidents, data requests, use cases)
7	Action on the Portal	Focus on portal usability	Mostly accessibility, scalability improvements	Steps to improve data quality	Some feedback leads to portal improvements	Incident handling, platform improvements
8	Statistics Count	User star ratings updated	Favorite/reuse counts updated	Not applicable	Tracks posts & participants by topic	Likes updated and shown
9	Publish Use Case	Sometimes use cases are generated	Not indicated	Benchmarking with other portals	Not explicitly published	Successful use cases shown publicly
10	Action on the Dataset	Dataset corrections based on feedback	Mostly format and update issues	Data quality improvements	Dataset actions arise from discussion	Dataset requests relayed to providers
11	Report to the User	Private reports, ratings are public	Not systematically reported	No mention of response	Transparent, real-time discussions	Public responses or private follow-ups depending on type

4.3.3 Extrapolation to Broader Open Data Portals

Once we analysed in detail the feedback within 5 specific case studies, the objective of this third phase of the study was to extrapolate the analysis into a wider variety of open data portals. We have chosen 26 European national open data portals, along with the Australian, US, and Canadian open data portals as shown in Table 4.3. As in this extrapolation exercise it is not possible to obtain detailed information for all the steps in the feedback scenario, we have focused on analysing the feedback flows between data users and providers through the identification of input channels and output channels.

Table 4.3 Open Data Portals by Country

Country	Acronym	Open Data Portal URL
Australia	AU	data.gov.au
Austria	AT	data.gv.at
Belgium	BE	data.gov.be
Bulgaria	BG	data.egov.bg
Canada	CA	open.canada.ca
Croatia	HR	data.gov.hr
Cyprus	CY	data.gov.cy
Czech Republic	CZ	data.gov.cz
Denmark	DK	opendata.dk
Estonia	EE	avaandmed.eesti.ee
Finland	FI	avoindata.fi
France	FR	data.gouv.fr
Germany	DE	govdata.de
Greece	GR	data.gov.gr
Ireland	IE	data.gov.ie
Italy	IT	dati.gov.it
Latvia	LV	data.gov.lv
Lithuania	LT	data.gov.lt
Luxembourg	LU	data.public.lu
Malta	MT	open.data.gov.mt
Netherlands	NL	data.overheid.nl
Poland	PL	dane.gov.pl
Portugal	PT	dados.gov.pt
Romania	RO	data.gov.ro
Slovakia	SK	data.gov.sk
Slovenia	SI	podatki.gov.si
Spain	ES	datos.gob.es
Sweden	SE	dataportal.se
United States	US	data.gov

To accomplish this objective, it was necessary to investigate the open data portals to find the way users interact with the data and provide feedback. Using the web-scraping method described in Section 4.2.3, we generated an initial list of channels for each OGD portal. Furthermore, input feedback channels are the identified feedback channels through which users can actively provide direct feedback on datasets, services, or portal features. These channels are designed to capture

user perspectives, suggestions, and ratings that help in assessing user needs and portal performance. Moreover, output feedback channels are primarily used by the portal team to communicate updates, information, and announcements to the users. While users may not provide direct feedback through these channels, they offer indirect insights into user engagement based on how users interact with the shared content, whereas input/output feedback channels serve as both feedback collection points and information-sharing platforms. Users can engage with the platform by both receiving updates and providing feedback through comments or discussions.

Figure 4.3 shows how many feedback channels each respective national open data portal has, and it provides insights into the diversity and extent of feedback channels available for users to engage with the open data portals.

The height of each bar in the chart stands for the number of feedback channels that are linked with a particular open data portal. France and Poland have the biggest number of feedback channels on their respective open data portals.

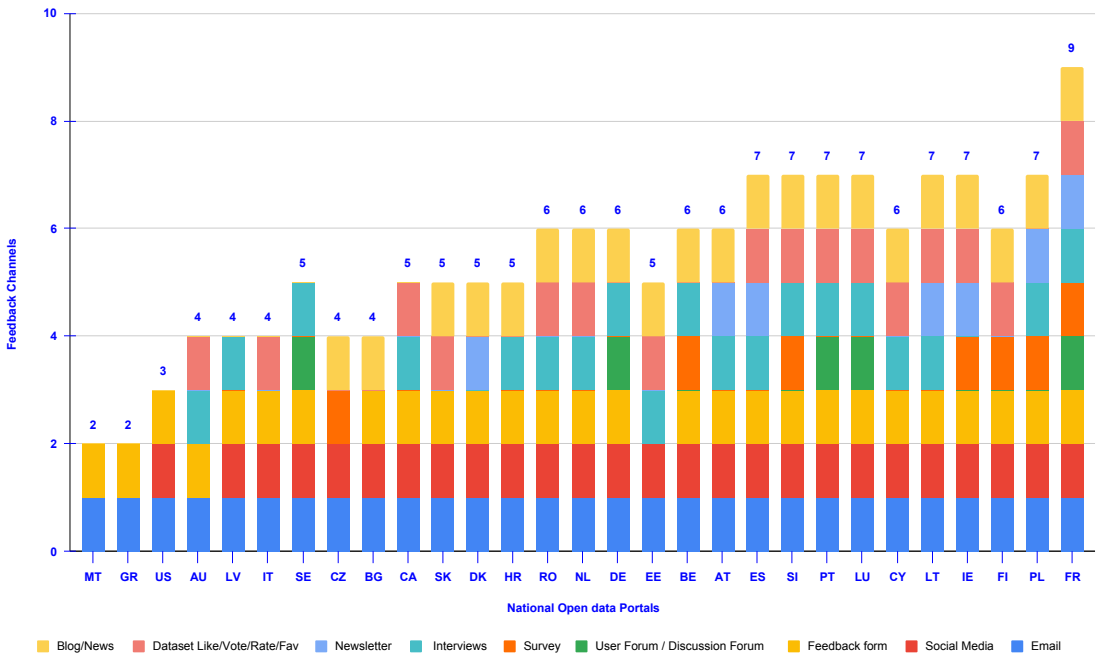

Fig. 4.3 Feedback channels of National Open Data Portals

On the other hand, we can observe that the open data portals in Malta and Greece have the fewest number of feedback channels. As a result, Figure 4.3 highlights the more feedback channels an open data portal has the more chances of user involvement to connect with the open data portal. To enhance user engagement, it is essential to ensure that feedback mechanisms within the open data portal are utilised to their fullest potential. Likewise, Figure 4.4 illustrates the probability of including one of the nine feedback channels identified in Table 4.1.

When the number on the y-axis is higher, such as when it is 0.9, it indicates that the linked

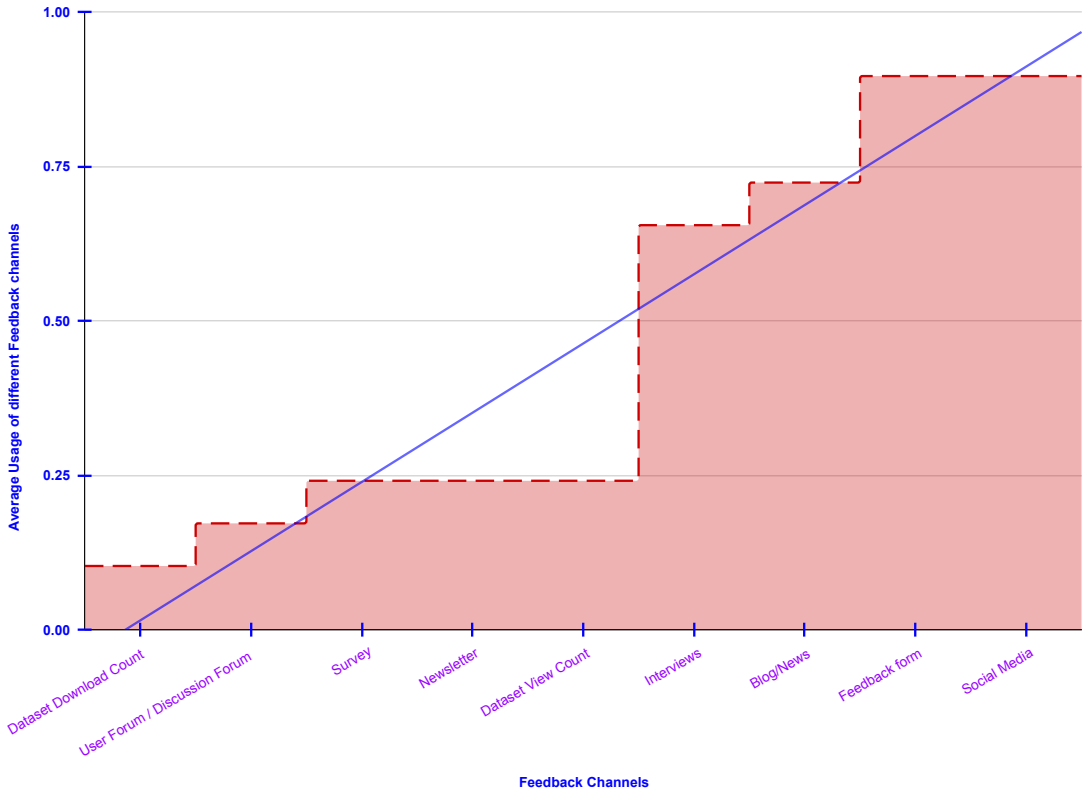

Fig. 4.4 Average number of times different Feedback channels offered by ODP

feedback channel, which includes Social Media and Feedback Form, is utilised to a significant degree across the majority of national open data portals. However, the lower values, such as 0.1 and 0.15, for channels such as Discussion Forum and Survey correspondingly, suggest that these channels are employed less frequently among the open data portals. This is the case since these channels are less frequently used. Furthermore, by utilizing this visual depiction of Figure 4.4, we can ascertain, at a glance, the number of individuals who are utilizing the various feedback channels and the level of popularity that they possess.

Moving further, Figure 4.5 depicts the dendrogram resulting from a hierarchical clustering analysis of countries based on their feedback channels. Hierarchical clustering was employed to group countries with similar feedback setups, using Ward's linkage method to minimise the variance within each cluster.

In the dendrogram, each leaf represents a country, and the branches illustrate the merging process of the clustering algorithm. The height of the branches corresponds to the dissimilarity (Euclidean distance) between countries or clusters. The dendrogram provides a visual understanding of the clustering process. Countries with shorter branch lengths between them share similar feedback channel patterns. The cutoff line at a specific height [3.0] was used to segment the data into distinct

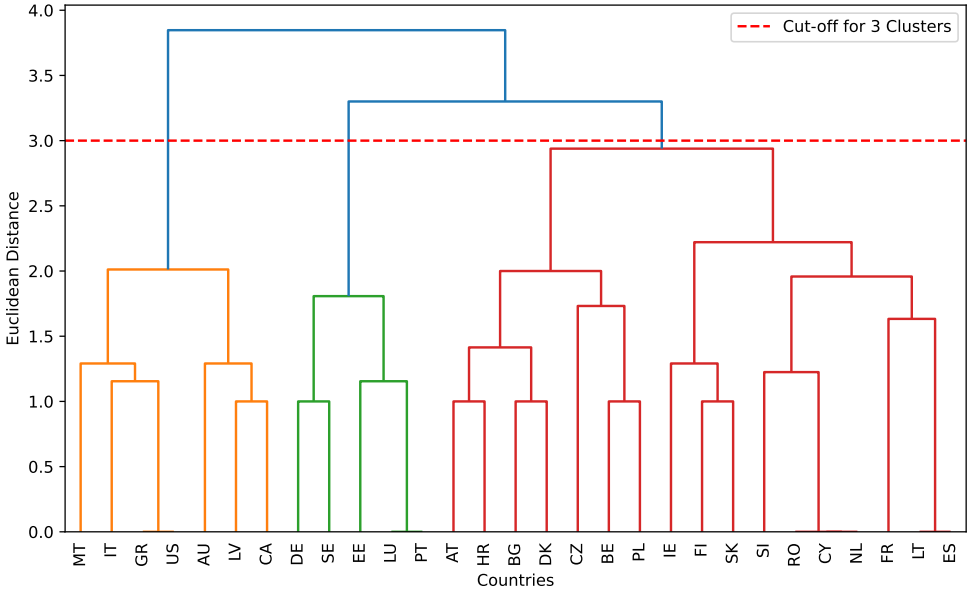

Fig. 4.5 Dendrogram of Countries Based on Feedback Channels

clusters, facilitating the interpretation of the results. These clusters reflect the varying strategies and preferences of countries in utilizing feedback mechanisms, such as Email, Surveys, and Social Media, for input and output communication. Moreover, it is necessary to see the cluster profiling as it provides a deeper understanding of the formed clusters and their distinguishing characteristics. This step not only helps to interpret the results but also reveals actionable insights by highlighting key differences and similarities across clusters. Hence, Figure 4.6 shows the cluster profiling of countries based on feedback channels.

Furthermore, Figure 4.7 shows the details and usage of the feedback channels (email and feedback form) for both; at the dataset level and at the portal level for chosen open data portals.It is possible to have a better understanding of the considerable distinctions that exist across the various national open data portals by reading this graphical representation. In particular, the United States of America, Canada, the Netherlands, and Australia stand out as having a complete feedback infrastructure that includes all four feedback channels. These channels are referred to as “Email at Portal Level", “Feedback Form Portal Level", “Email at Dataset Level", and “Feedback Form Dataset Level". Additionally, several countries like the Czech Republic, Estonia, Germany, and others have a feedback landscape that is more restricted than others. This is because the national open data portals of these countries only support two of the feedback channels that were outlined.

Moreover, Figure 4.8 illustrates illustrates the potential connections between input feedback channels and output feedback channels considering the co-occurrence of the channels in the different countries. The thickness of the connecting flows represents the number of countries having such a pair of input and output channels. Before making the relationship, we have thoroughly examined the input

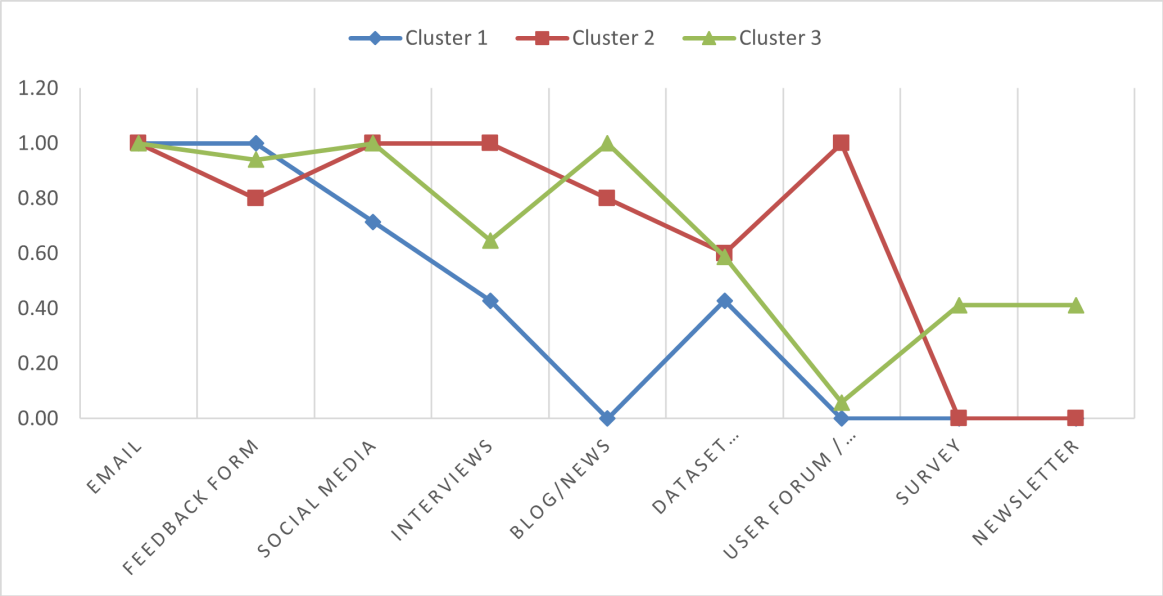

Fig. 4.6 Cluster Profiling of Countries Based on Feedback Channels

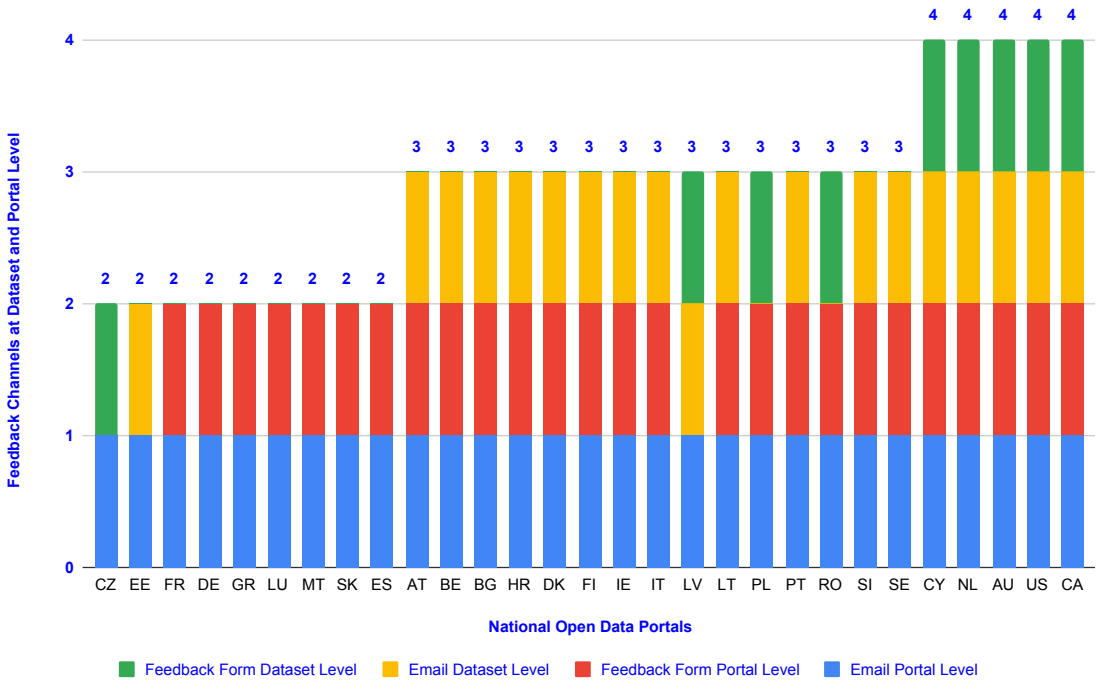

Fig. 4.7 Input feedback channels at Dataset level and Portal level

and output feedback channels to assess not only the dependency of the input and output feedback channels, but also which channels should be considered as input, and which should be considered the output based on the nature of the feedback channel. For instance, in Figure 8 it can be seen that "Dataset Like/Vote/Rate/Fav" can only be the output of "Dataset Like/Vote/Rate/Fav". That is to say,

once a user confirms a like for a dataset, other users can see how many users have marked that dataset as liked/rated/favourite, but this channel cannot be the output of a previous Email input.

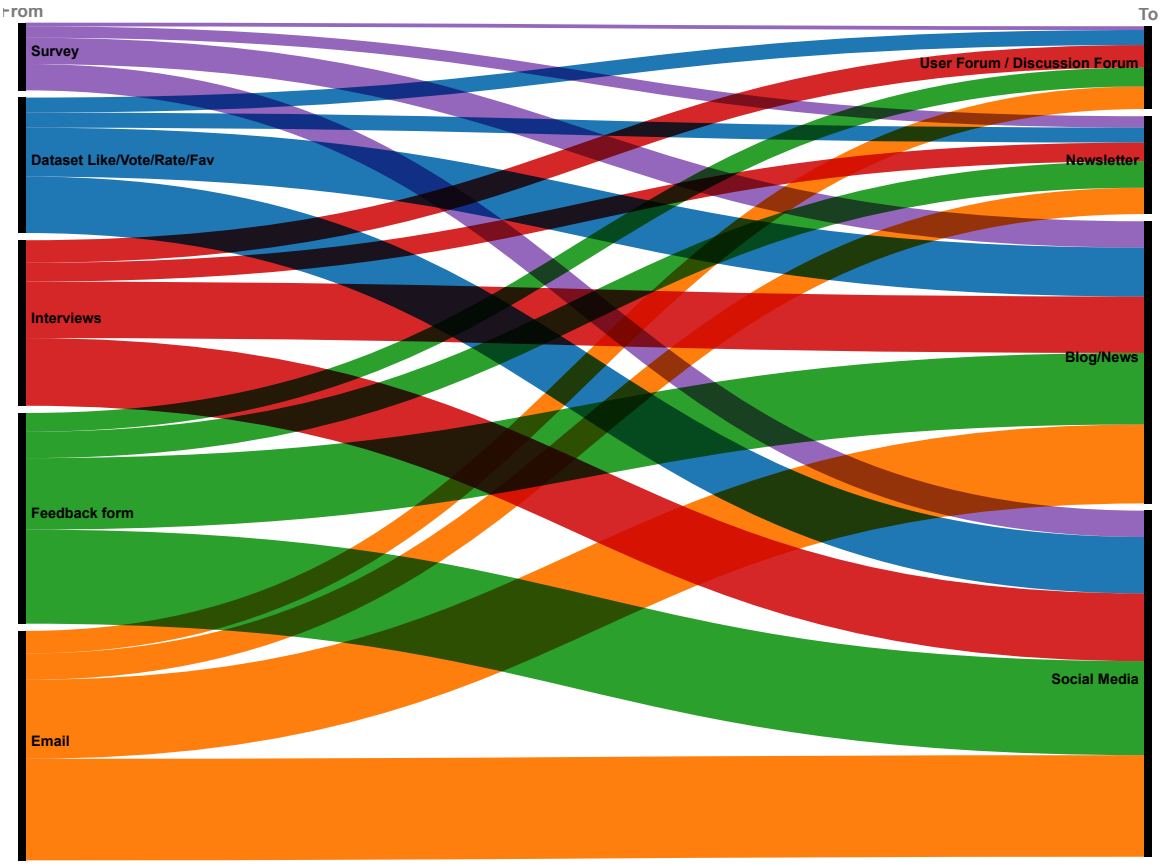

Fig. 4.8 Relationship between input and output feedback channels across various countries

In general, Figure 8 highlights how certain input channels, such as Email, feedback forms and social media have thick flows connecting them to multiple output channels, indicating their widespread adoption and versatility in feedback mechanisms. In contrast, channels like surveys and User Forum/Discussion Forum features exhibit more targeted and less widespread connections, suggesting specialised usage. The diversity in flow thickness reflects the probability of the occurring flows.

4.4 Discussion

The findings reveal that open data portals employ diverse feedback channels, each serving unique interaction purposes. Input feedback channels such as email, feedback forms, comments, and dataset ratings provide users with direct means to express their specific needs, suggestions, and concerns. These channels are invaluable in capturing granular user perspectives that inform data quality and portal improvements. Output feedback channels, including blogs and newsletters, are primarily one-

way communication tools designed to disseminate updates and announcements. While they play a critical role in maintaining user awareness, they do not facilitate interactive engagement. Input/output channels like social media and discussion forums, on the other hand, serve as bi-directional platforms that enable real-time interactions and foster collaborative contributions from users. In addition, it must be noted that, except for the specific five cases reported in section 4.2, the flows in Figure 8 only reflect potential connections between the feedback received through input channels and the responses addressing these inputs that can be conveyed through the available output channels of an OGD portal. In order to have a real evidence of the internal process for managing feedback, we should have a direct communication with the administrative teams of every portal.

Anyway, our analysis highlights the varying levels of sophistication in feedback classification and prioritisation mechanisms across portals. Structured methods, such as those used by the Swedish Open Data Portal, categorise feedback into specific topics and track user engagement metrics like the number of posts and participants. This systematic approach facilitates the prioritisation of feedback and enhances the ability of administrators to act on user concerns effectively. In contrast, portals such as those in Estonia and Luxembourg often rely on ad hoc feedback management processes, which may lead to inefficiencies in addressing user needs. The lack of formal classification frameworks in these cases limits the scalability and effectiveness of feedback processing. Moreover, feedback is essential for facilitating improvements at both the portal and dataset levels. For example, portals in Spain and Poland integrate user recommendations into new data efforts, apps, and upgrades. These acts illustrate the capacity of feedback to improve the relevance and accessibility of data. Nonetheless, the research indicates that several open data portals fail to consistently communicate the results of user input. Concluding this feedback loop is vital for establishing user trust and promoting user involvement. By offering updates or replies to user comments, portals may illustrate the significance attributed to user contributions and build a stronger relationship with their user base.

4.5 Summary

Understanding the feedback mechanisms in open data platforms is crucial for evaluating their effectiveness in promoting the quality of data and user participation. We tested the feasibility of our methodology by designing a comprehensive conceptual definition of feedback mechanism through the literature review, a refinement through specific case studies, and the extrapolation of the feedback analysis through the automated study of input and output channels offered by 29 open data portals, which included 26 OGD initiatives from Europe. This study analysed feedback interaction processes focusing on mechanisms for collecting, categorizing, and utilizing user feedback to improve data quality, data accessibility, usability, and overall portal effectiveness.

The findings of this study have a direct and practical implication on the management of OGD initiatives. OGD portals of countries incorporated in the extrapolation phase, or in general OGD portals with similar features, can compare their status with respect to other countries and decide to change their user engagement policies and enhance their feedback mechanisms. With the guideline of

the input and output channels investigated in this study, the administrative teams of these initiatives may identify channels that are not exploited and initiate communication flows with final users.

Regarding the limitations of our study, it must be acknowledged that one of the limitations of our study was that our research questionnaire was only completed by 17% of the national portals that were contacted. Part of our future work will be devoted to conducting additional surveys with administrators from the 29 analysed open data portals. These surveys would facilitate the validation of the specified input and output channels and evaluate the success of the connecting flows between them as well as the processing and implementation of user feedback. This step is crucial to ensure that the current understanding of feedback systems aligns with actual practices and operational realities.

Another limitation of our study is that we have not analysed whether we can distinguish the interactions by different types of user groups. Within the scope of this work, we have just considered the general needs for feedback of final users without considering whether students, journalists, non-governmental organisations, private companies, or open data intermediaries, among other different user groups, interact with an open data ecosystem in a different way. This profiling of users with respect to feedback should be investigated as future research.

Another research line is to explore whether it is possible to establish reliable metrics for evaluating the success of feedback mechanisms. Metrics like feedback response from data publishers, initiating discussion at dedicated discussion forums, and post-submission user satisfaction can provide actionable insights into the effectiveness of these feedback systems. Additionally, community engagement strategies should be prioritised to raise awareness about the availability and importance of feedback channels. Outreach campaigns, co-creation initiatives with users, and targeted efforts to build trust in feedback systems are essential to improving adoption of the feedback channels.

CHAPTER 5

A METHOD FOR THE ANALYSIS OF FEEDBACK THROUGH SOCIAL NETWORKS

The idea behind Open Data openness is that users should be able to freely access, utilise, and share the data in whatever manner they want [146]. Open Data portals are a sort of digital library since they are online catalogues that include descriptions of datasets, known as metadata. These kinds of catalogues make it possible to find and manage metadata records describing datasets that are either accessible online or may be downloaded in a variety of distribution formats. Furthermore, metadata records facilitate the use and reuse of datasets by providing details of authorship, provenance, and license, among other details [7, 60, 61]. Indeed, the use and reuse of data from the public sector is a crucial aspect that is driving the present trend of opening up government data [8, 62]. Open Government Data (OGD) portals play a critical role in opening the data, and the constant publishing of open data in OGD portals increases the demand for data of a high quality as well as a higher quality in the portal itself. However, the majority of existing open data ecosystems are not user-driven and thus fail to properly balance supply and demand. Although the importance of users in shaping open data ecosystems is well acknowledged, existing ecosystems are mostly influenced by service providers [47].

The participation of users is essential to make existing OGD initiatives more user-oriented. Some Open Data portals already offer specialised forums or online forms where diverse groups of users may report on their experiences reusing the datasets available on these portals. Other initiatives even provide users access with specialised tools for storytelling to narrate their experiences with OGD datasets [63]. However, these feedback mechanisms are, in general, very heterogeneous and the input obtained from users is rarely accessible by the general public to be compared across different OGD initiatives. Given this lack of mature feedback mechanisms, this chapter proposes employing social networks as one of the main sources to investigate user involvement in OGD initiatives. Social networks function as an open forum in which a variety of stakeholders may share their perspectives about any kind of activity or organisation. In addition, social media platforms have the potential to enhance visibility by driving visitors, engaging them via the presentation of data and portal functions, and motivating them to return [137]. With respect to the selection of the social network that better depicts the involvement of users in Open Data portals, *X* (formerly Twitter) seems to be one of the most

practical sources. Apart from being used to discussing subjects ranging from personal to professional interests, there is a growing trend to share academic content and knowledge [147]. Furthermore, according to studies performed within the context of European Union [137], *X* is the most extensively used social media channel for OGD initiatives.

Furthermore, this chapter addresses the research question RQ4. The purpose of this chapter is to propose a methodology for analysing OGD initiatives, and in particular, their user engagement. As a first step of the methodology, we propose to define a set of variables compiled along a time-period frame that characterise both the main features of the Open Data initiatives and the activity related to these initiatives that has been reported in the *X* social network. Then, to analyse the situation of OGD initiatives from a multidimensional and temporal perspective, we propose a combined use of self-organizing maps (SOM) and clustering techniques. On the one hand, SOM allows the distribution of OGD initiatives over a two-dimensional map with a reduced number of nodes. Each node represents a neuron of the SOM neural network, which has identified hidden partial correlations among the data, characterizing the initiatives classified within this node for a particular date. On the other hand, we propose to apply an agglomerative clustering algorithm over SOM neurons to identified uniform areas in the SOM map, which represent initiatives with a similar status of development and user engagement. The classification of initiatives into different clusters in the analysed time period allows us to establish trajectories of development and detect which types of initiatives are more prone to evolve into a more mature status. The feasibility of our proposed methodology has been tested by conducting an in-depth study of 27 European OGD portals during the period of 2017 to 2021, collecting variable data at a yearly rate.

5.1 Related work

There are several research works in the literature that have proposed frameworks for monitoring the quality of Open Data portals. For instance, Kubler et al. [60] proposed a framework of 21 metrics to evaluate the metadata of Open Data portals in five quality dimensions: existence of properties describing key aspects of datasets such as the access, discovery, contact, rights, preservation, or temporal/spatial coverage; conformance of the content of some properties (e.g., URLs, e-mail, formats); retrievability of datasets and resources; accuracy of format and file size; and an Open Data dimension assuring the existence of open and machine readable formats. Nogueras-Iso et al. [76] proposed a framework consisting of different quality controls on Open Data Metadata with quality elements and measures inspired by the ISO 19157 standard for geographic information quality. Apart from completeness and consistency, their approach reviews exhaustively the correctness of temporal, positional, and attribute information. After testing this approach on the Spanish OGD initiative, the quality indicators revealed that accuracy and correctness of metadata should be improved. Furthermore, Máchová and Lněnička [65] also proposed a framework to assess the quality of Open Data portals on a nationwide basis in the Czech Republic. Their results indicate that there is a need for quality standards and that Open Data portals differ in the number of provided datasets as well as in the level

of sophistication of the services offered. More focused on transparency aspects, Lourenço et al. [148] proposed a set of criteria that Open Data portals should meet. This work concludes that entity coverage, information types, information seeking strategies, and data quality features are significant factors to ensure transparency and accountability. In addition, there are also works that have investigated the influence of transparency as a design concept for Open Data portals [73, 119]. By correlating certain literary features with various phases of the transparency cycle, Open Data portals should be able to fulfil the transparency requirements.

The previous works are relevant to have an overall perspective of the current status of Open Data initiatives, their maturity or their commitment to FAIR principles [57, 149]. However, they do not consider any insight into the direct opinion of user engagement. Moving forward to the analysis of the user perspective with respect to the interaction with prevalent Open Data platforms (e.g., CKAN, DKAN, Socrata), there are several works that have examined the technical commons, approaches, features, and methodologies provided by each platform, as well as their visualisations tools [150, 151]. In addition, they explored the question of why these platforms are significant to users like providers, curators, and end-users, as well as the question of what the most important publishing alternatives are accessible on these platforms.

With a higher emphasis on the analysis of user interactions in Open Data portals, Begany and Gil-Garcia [43] monitored the levels of user engagement by analysing web analytic behavioural data taken from the New York State open health data portal. In addition, they emphasised the actual use of open data and more specifically how users of Open Data portals interact with open datasets. Relying on a more manual and qualitative approach, Nikiforova et al. [152] proposed a survey to analyse and compare the various contexts regarding the employment of OGD portals by users and emphasising the most often disregarded user-centred aspects. This work has resulted in the validation of a paradigm for the usability analysis of OGD portals, as well as the identification of the strengths and flaws of portal usability that are similar across settings. In the same line, Zhu and Freeman [153] evaluated various approaches to user interactions with OGD Initiatives. They developed a user interaction framework, in which they evaluated the United States Municipal Open Data portals and provided the findings regarding user understanding and engagement with the data portals.

Concerning the evaluation of Open Data portals in the context of the European Union, it is worth noting the existence of the Open Data Maturity Report released by the Publications Office of the European Union [137] on a yearly basis. This report mentions four dimensions for the analysis of initiatives: policy, impact, portal, and quality. In the portal dimension, it includes a sustainability variable that identifies actions applied to promote the visibility of the portal, including social media presence. According to this report, *X* (formerly Twitter) is the most widely used social media channel for this purpose.

Although there are numerous works using social media as the main source for investigating the impact of public and private organisations [154, 155], the influence of users [156], or the dissemination of scientific publications [157], there are relatively few works using social media for studying the impact of Open Data portals on the user community. Most of the existing works focus on the

dissemination of datasets. For instance, Khan et al. [138] explored data citation and reuse practices in 43,802 openly available biodiversity datasets. The altmetrics sourced from blogs, *X*, Facebook, and Wikipedia suggest that social activity is driven by data publishers and data creators. Authors made a hypothesis that such activities are promotion-related and may lead to more reuse of open datasets. Likewise, Hou et al. [158] conducted a study that investigates the distribution of datasets on *X* among academics and the general public. After an analysis of 2,464 datasets from Altmetric.com, they identified viral and diverse dispersion patterns within one or two diffusion levels in social networks.

5.2 Proposed Framework

This study takes a quantitative approach to the analysis of a variety of indicators about the open data portals that are maintained at national level by EU member nations. We propose two different research methods: the first approach is to examine a snapshot of the current status of the portals, and the second method is to study the temporal evolution of the open data portals. Both of these methods are discussed below in detail.

5.2.1 Snapshot-Based Analysis of OGD Portals

This section presents how to make a snapshot-based evaluation of OGD portals, focusing on the data of a specific year. The analysis applies a factor analysis technique. Figure 5.1 highlights the details of our proposed methodology. Our methodology consists of 6 steps, which are described below:

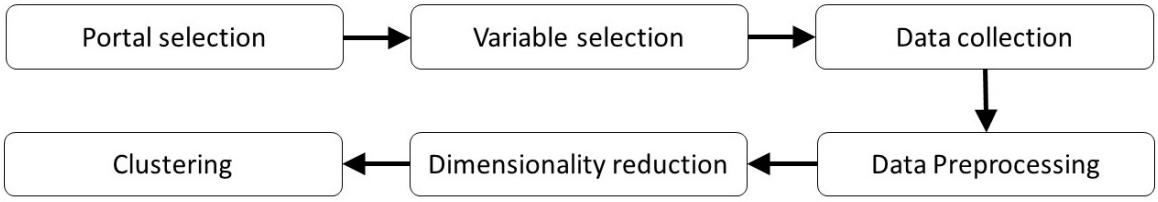

Fig. 5.1 Proposed Research Methodology for Data Processing

Portal Selection: The first step of the methodology is to find appropriate resources to identify the location of portals and the documents describing their features. The list of platforms studied comes from two sources: the national catalogues of the European Data Portal (EDP) euwebsite, and the compilation made by Juana-Espinosa and Lujan-Mora [1]. Both sources showed a high degree of concordance.

Variable Selection: This refers to the process of choosing relevant variables describing the features of the Open Data portal to include in our model. The relevant variables and their sources are shown in Table 5.1. In our experimental design, we make a choice of variables taking advantage of the sources of information available through the national portals under observation and through the European Data Portal. Therefore, some of the most representative operational attributes are gathered

from the EDP (ND, ODM, MQA, URL), some of them are X (Twitter) activity metrics (NT, TFP, UT, NI), and there are other variables (NU, GS) that help us understand the magnitude of data reuse around each portal. It must be noted that the variables present in Table 5.1 are the final selection of variables: our experiment included some other variables and combinations of them, but they were discarded due to their negative effect on the feasibility of indicators obtained during experiment in the last two phases of the methodology (Dimensionality Reduction and Clustering). Variables present in Table 5.1 are by no means intended as a complete and exhaustive list. In fact, later steps help us to explore the underlying structure that may be useful for refining the variable selection in the future.

Table 5.1 Description of the variables

Variable	Description	Source
ND	Number of datasets available for consultation	Automatic from data.europa.eu
ODM	Open Data Maturity score (0-100)	Manual from data.europa.eu
MQA	Metadata Quality Assurance rating (0-405)	Automatic from data.europa.eu
URL	% of accessible URLs	Automatic from data.europa.eu
NU	Number of data use cases listed in the portal	Manual from portals
GS	Number of items in Google Scholar citing the portal	Manual from Google Scholar
NT	Number of relevant Tweets	Automatic, derived from X (Twitter) API
TFP	Number of Tweets by portal account	Automatic X (Twitter) API
UT	Number of users posting Tweets	Automatic X (Twitter) API
NI	Number of interactions generated by Tweets. This corresponds to the sum of retweets, replies, quotes and likes.	Automatic X (Twitter) API

Data Collection: This refers to the process of gathering data from reliable sources mentioned in Table 5.1 that guarantee the reproducibility of the measurements. We assume that the values obtained are valid and representative as they are gathered from recognised sources such as the European Data Portal and the academic X (Twitter) API. The variables representative of the EDP can be collected through the EDP API (MQA, ND, URL) [159] or manually (ODM). The variables measuring the conversation on X (Twitter) related to portals for the year 2021 (NT, TFP, UT, NI) can be collected using the X (Twitter) API for Academic Research [160]. This API allows the retrieval of tweets whose text mentions the URL of portals or their X (Twitter) accounts. Finally, the number of use cases listed

in a data portal (NU) and the number of mentions in Google Scholar (GS) must be collected manually for each data portal.

Data Processing: This consists of preparing the raw data and making it suitable for the analytical models. First, we must compute the correlations between the metrics using the Spearman coefficient. This coefficient can range from -1 to 1, with -1 or 1 indicating a perfect monotonic relationship: when the value of one variable increases, the other variable value also increases or decreases. After that, we must normalise the variables by removing the mean and scaling them to unit variance.

Dimensionality Reduction: This step involves exploring the underlying variable structure and reducing the data to a smaller number of explainable factors. For this purpose, we propose the use of factor analysis to reduce the dimensions of the original dataset [161]. Likewise, Bartlett test ($X^2 = 166.56$, $p < 0.0001$) and the Kaiser–Mayer–Olkin test ($KMO = 0.67$) are employed to verify the feasibility of the overall factor analysis. We take into account the Kaiser-Guttman criterion ($eigenvalue > 1.0$) to decide the optimal number of factors and, for each factor, only variables with loading greater than 0.4 after applying Varimax rotation are considered to influence the factor.

Clustering: Clustering consists of grouping portals into groups based on the dimensions that describe them. For this step, we propose to apply three common clustering methods: hierarchical clustering, K-means clustering, and K-medians clustering [162] (less sensitive to outliers). Combining these clustering techniques is a common way to improve the robustness of the final results [163]. The ideal number of clusters for K-means is defined by plotting the explained variation as a function of the number of clusters and identifying the inflection point at which adding another cluster does not improve much better intra-cluster variation, a procedure also known as the “elbow method”.

5.2.2 Temporal Analysis of OGD Portals

This section describes how to make a longitudinal analysis of OGD portals over a period of years, using the five-year period from 2017 to 2021 as an illustrative example. By applying Self-Organizing Maps (SoM) technique, we aim to uncover temporal trends in portal maturity, metadata quality, and user engagement. Figure 5.2 show our second proposed research technique consists of five stages: selection of portals; selection of variables characterizing each portal; collection of data for each variable; application of the SOM technique to reduce the dimensionality of variables; and the clustering of SOM results.

The first stage which is portal selection is the same as of described in section 5.2.1 and the rest of the stages are described below:

Variable Selection: This refers to the process of selecting variables that adequately characterise the properties of the Open Data portal for inclusion in our methodology. Table 5.2 displays the pertinent variables together with the sources from which they can be derived. The selection of variables was driven by both theoretical considerations and data availability. From a theoretical perspective, we have replicated the methodological approach followed by other authors [1, 164] who have benchmarked open data initiatives using a combination of technical indicators about portal performance and few

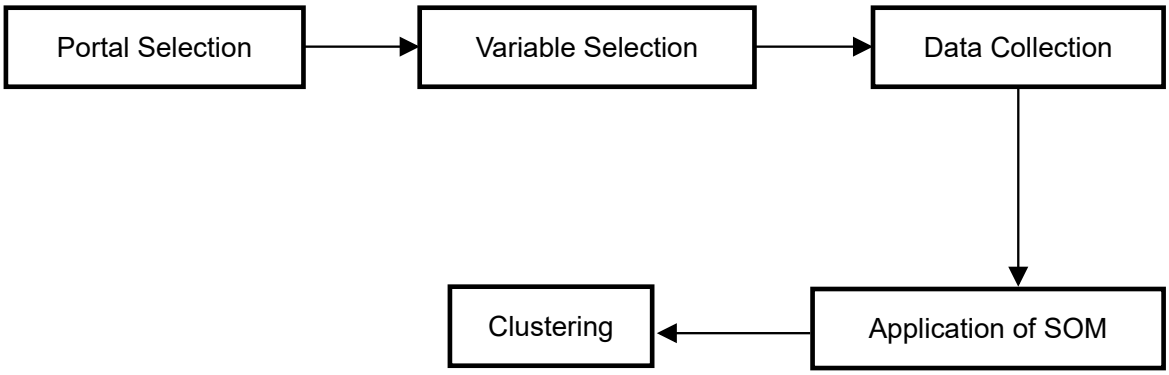

Fig. 5.2 Proposed Methodology for Data Processing

other metrics related to impact dimensions that have not already been covered in current literature. As a result, some of the most representative operational characteristics are taken from the European Data Portal (Number of datasets - ND, Open Data Maturity score - ODM). These variables are commonly used in other similar studies [1, 164]. For introducing the new social media impact dimension, we included metrics of social activity on X (Number of Tweets - NT, Tweets from Portal - TFP, User Tweets - UT, and Number of Interactions - NI). Additionally, we added a metric of academic impact for exploratory purposes (Google Scholar - GS).

Table 5.2 Description of the variables collected for each OGD initiative and year

Variable	**Description**	**Source**
ND	Number of datasets available for consultation / log of population	European Data Portal SPARQL API and statistics REST API
ODM	Open Data Maturity score (0-100)	Reports at European Data Portal
GS	Number of items in Google Scholar citing the portal / log of population	Google Scholar
NT	Number of relevant Tweets / log of population	*X* API
TFP	Number of Tweets from portal account / log of population	*X* API
UT	Number of users posting Tweets / log of population	*X* API
NI	Number of interactions generated by Tweets (sum of retweets, replies, quotes and likes) / log of population	*X* API

In addition, it must be noted that the raw value of some of the selected variables (e.g. number of datasets or number of users in social networks) is clearly proportional to the size of the country behind the OGD initiative. Therefore, we decided to normalise the values of these variables dividing by the log of the population of the country at each analysed year. The only exception is the ODM variable,

as this refers to a qualitative measure of maturity reported by experts that take into consideration the whole context of the initiative.

Data Collection: This step refers to the process of collecting data of the selected variables for each OGD initiative and year in the analysed period. Regarding the variables that were gathered manually, it is important to point out that the ODM variable was extracted from the reports published by the European Commission [137]. In the case of GS variable, a manual search in Google Scholar for the number of publications citing the homepage URL of each OGD initiative was carried out. In addition, it must be noted that this search was performed for each year in the analysed period by adding a temporal filter on the citing publications.

With respect to the values collected automatically, the collection of values associated with the ND variable was not an easy task. Although the European Data Portal facilitates an SPARQL endpoint to query the Open Data collected from the different national initiatives [159], the temporal information contained in metadata records is not a completely reliable source. The *dcat:Dataset* entity of metadata records (compliant with the DCAT-AP vocabulary) includes *dcat:created*, *dct:modified* and *dct:issued* properties to inform about the creation date, the modification date and the publication date of a dataset. However, either this information is sometimes missing or it does not explicitly imply that a dataset was directly published on a national data portal. A most reliable source also available at the European Data Portal are the statistics compiled as a result of the harvesting processes performed along time from the different catalogues of the OGD initiatives.[1] This statistical information can be queried through a specific API.[2] The problem is that this information is only available from 2019 onwards. Therefore, in order to estimate the missing number of datasets for the years 2017 and 2018 for each portal, we assumed a constant annual growth rate for the period with the available data between 2019 to 2022 and projected that growth rate to the previous two years where there is no data. We computed this constant annual growth rate, named r, for each portal using the following equation:

$$datasets\_at\_2019 \cdot (1+r)^3 = datasets\_at\_2022$$

Consequently, the formula for obtaining r is as follows:

$$r = \sqrt[3]{\frac{datasets\_at\_2022}{datasets\_at\_2019}} - 1$$

With respect to the variables that measure the online social network activities on *X* connected to portals from 2017 through 2021 (NT, TFP, UT, and NI), they were gathered at the end of the year 2022 with the use of the *X* API for Academic Research [160]. This API makes it possible to get tweets whose content either references the URL of portals or their *X* accounts.

Application of the self-organizing map (SOM) technique: A Self-Organizing Map (SOM) is an artificial neural network method that performs dimensionality reduction over an input dataset. The standard SOM algorithm involves an unsupervised neural network with competitive learning

[1] https://data.europa.eu/catalogue-statistics/evolution/countryCatalogue?locale=en
[2] https://data.europa.eu/api/hub/statistics/data/ds-per-catalogue?list=true

and no hidden layers [165]. The objective is to align an input vector (representation of the variables describing an element of the dataset) with a neuron in an output matrix of neurons. SOM maintains the topology of input characteristics while simultaneously reducing the number of dimensions in a dataset and making this dataset easier to understand [166]. Figure 5.3, resents an example of a SOM trained for improving the visualisation in a lower number of dimensions of a dataset consisting of the records describing an OGD initiative at a particular year with the variables enumerated in data collection section.

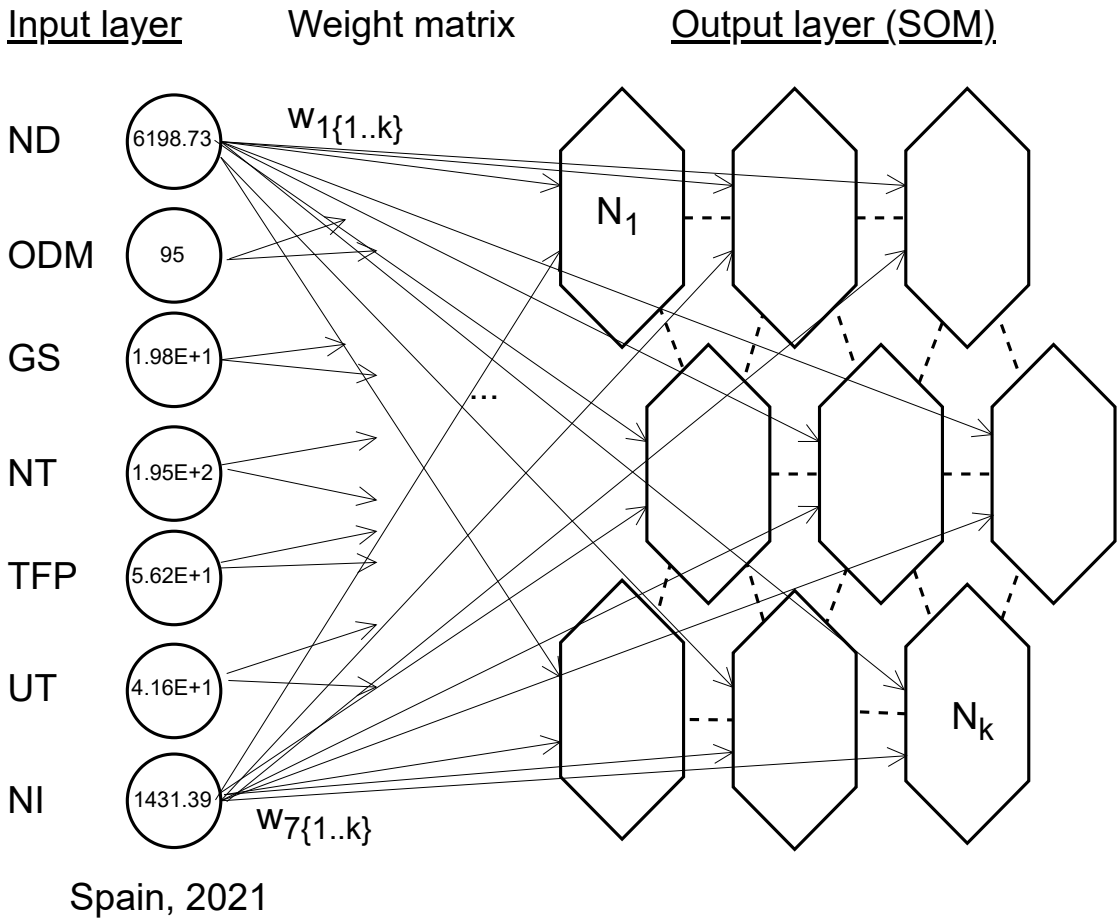

Fig. 5.3 An example of a small SOM trained with a dataset consisting of records describing OGD initiatives

During the training of this neural network, each of the vector components in the input layer is connected to the complete output map (output layer) by a weight matrix. It can be observed in the example that we have chosen a hexagonal topology, i.e. each central neuron has 6 neighbours. In contrast to other networks, the surrounding neurons of the output topology have an impact on the weights when the network is being trained. In addition, the output map in the example has a size of 9 ($k = 9 = 3 \times 3$). The size of the network must be set up during the experimental phase in order to

accommodate an appropriate number of neurons in x and y directions according to the original size of the dataset.

Clustering: The use of a SOM model allows a clearer analysis in terms of the linkages between the produced aggregations. However, as the number of nodes in the SOM map is too big to identify similar levels in the status of development of OGD initiatives, this phase of the methodology proposes to apply a clustering to the nodes of the SOM map. Data points (i.e. initiatives at different years) that are similar to one another and are grouped together in clusters according to the underlying patterns and connections that they share.

We propose to use an agglomerative clustering algorithm [167], i.e. a hierarchical clustering technique that follows a bottom-up approach to partitioning the data generating a hierarchical structure progressively. In particular, we use use the Ward clustering algorithm [168].

5.3 Experiments and Results

This section shows an extensive experimental evaluation of the proposed approach. For the evolution of Open Government Data Initiatives, the proposed two techniques are the core part of our methodology. The first technique is observing the evolution of OGD initiatives for one year and the second technique is observing the evolution of OGD initiatives for 5 years.

5.3.1 Results for Snapshot-Based Analysis of OGD Portals

This section displays the outcomes of applying the proposed methodology on the national Open Data Portals of 27 EU member countries and their X (Twitter) activity in 2021. We selected 2021 as this is a year with complete information on X (Twitter) activity. In addition, the values obtained from the Open Data Maturity report or from available APIs also reflect the situation after year 2021 had finished (when the experiment was performed).

Table 5.3 shows the results about the values of variables for the 27 portals under observation with the mean and coefficient of variation (CV) corresponding to each of them. The variables describing X (Twitter) activity (NT, TFP, UT, NI) and the number of use cases (NU) are the ones with the greatest relative variability. Similarly, in terms of relevance to portals, France, Spain, and Austria have the highest nominal values for the parameters of X (Twitter) conversation, number of use cases, and Google Scholar mentions. The Hungarian platform is the only one that does not follow a catalogue structure and does not have values for most of the indicators under observation.

Furthermore, the Spearman rank correlation coefficient is used to measure the strength and direction of association between pairs of variables, which is shown in Table 5.4. While looking at the X (Twitter) metrics (except for the number of tweets by the portal account itself), the number of use cases and mentions in Google Scholar are strongly and positively correlated with each other. Moreover, the Metadata Quality Assurance rating correlates positively and moderately with the

Table 5.3 Values of variables for Open Government Data portals of the EU countries and their X (Twitter) activity in 2021

Country*	Portal URL	ND	ODM	MQA	URL	NU	GS	NT	TFP	UT	NI
FR	data.gouv.fr	41,881	98	172	67	3,099	556	1,843	45	921	33,750
ES	datos.gob.es	60,102	95	196	46	400	137	1,384	448	294	9,974
AT	data.gv.at	38,586	92	199	93	689	158	258	110	85	3,696
IT	dati.gov.it	53,490	92	152	54	0	31	214	44	35	1,041
IE	data.gov.ie	13,815	95	185	42	23	73	173	3	46	1,377
LV	data.gov.lv	612	77	165	49	0	19	144	0	30	1,914
PL	dane.gov.pl	26,180	95	166	99	45	104	116	0	57	1,481
LU	data.public.lu	1,613	66	131	97	150	24	104	37	25	319
NL	data.overheid.nl	21,259	92	192	89	118	53	95	40	40	363
DE	govdata.de	51,275	89	240	56	24	118	85	9	55	1,502
CZ	data.gov.cz	142,554	74	276	99	0	19	62	43	11	702
GR	data.gov.gr	10,446	82	106	29	0	36	44	0	37	303
BG	data.egov.bg	10,680	78	47	0	0	6	37	15	12	119
FI	avoindata.fi	2,058	86	203	4	77	60	28	3	17	571
RO	data.gov.ro	2,753	76	98	7	10	15	28	0	18	24
DK	opendata.dk	823	91	164	42	0	22	27	14	12	137
PT	dados.gov.pt	4,928	66	183	81	51	43	25	0	14	242
CY	data.gov.cy	1,210	91	226	12	47	9	8	3	6	52
SE	www.dataportal.se	7,825	84	170	30	0	15	8	0	6	153
HR	data.gov.hr	1,141	84	96	52	6	22	6	0	3	23
BE	data.gov.be	13,056	55	218	31	81	12	3	0	3	15
SI	podatki.gov.si	5,098	92	120	61	14	17	2	0	2	12
EE	avaandmed.eesti.ee	879	94	0	0	150	5	0	0	0	0
LT	data.gov.lt	1,721	89	99	56	28	3	0	0	0	0
MT	open.data.gov.mt	205	51	0	0	0	2	0	0	0	0
SK	data.gov.sk	2,862	50	124	0	11	12	0	0	0	0
HU	kozadat.hu	0	58	0	0	0	0	0	0	0	0
Mean		19150.1	81.2	145.5	44.3	186.0	58.2	173.9	30.1	64.0	2139.6
CV		1.6	0.2	0.5	0.8	3.2	1.9	2.4	2.9	2.8	3.1

*We are using the two letter-code of ISO-639 to refer to the country of the Open Data initiatives that have been analysed.

number of datasets (0.54) and the percentage of accessible URLs (0.54). The remaining correlations are weak.

Given the high correlation between the X (Twitter) conversation variables, we removed UT and NI before factor analysis to reduce the effect of multicollinearity. The outcome for a three-factor solution accounting for 72% of the variance is shown in Table 5.5. The number of use cases, mentions in Google Scholar, and tweets are the variables that best explain factor 1. The number of datasets, MQA rating, and the percentage of accessible URLs are the most representative variables for factor 2. Finally, the number of tweets by portal account is considered the best variable for factor 3, with a small contribution of the number of tweets. The ODM score did not load in any of the three factors. From the factor loadings, factor scores are computed for each portal.

Observing high factor loadings associated with particular variables implies that these variables

Table 5.4 Spearman correlation

	ND	ODM	MQA	URL	NU	GS	NT	TFP	UT	NI
ND	1									
ODM	0.19	1								
MQA	0.54**	0.30	1							
URL	0.49*	0.32	0.54**	1						
NU	0.19	0.29	0.12	0.21	1					
GS	0.28	0.39*	0.26	0.30	0.96**	1				
NT	0.33	0.36	0.20	0.20	0.84**	0.87**	1			
TFP	0.39*	0.26	0.23	0.17	0.19	0.26	0.63**	1		
UT	0.25	0.33	0.16	0.19	0.97**	0.96**	0.94**	0.33	1	
NI	0.26	0.32	0.16	0.19	0.97**	0.96**	0.93**	0.32	1.00**	1

** $p < 0.01$ and *$p < 0.05$ indicate significant correlation.

contribute more to this component. Therefore, portals with high values on these variables tend to have higher factor scores on this particular dimension and vice versa for low values.

Table 5.5 Rotated matrix for factor analysis

Variable	Factor		
	1	2	3
ND		0.64	
ODM			
MQA		0.78	
URL		0.72	
NU	0.97		
GS	0.96		
NT	0.84		0.52
TFP			0.94
Eigenvalues	2.70	1.80	1.30
Variance	0.34	0.22	0.16
Cum. Variance	0.34	0.56	0.72

Note: Loadings with absolute values below 0.40 are omitted from the table

The next stage of the research process involved clustering methods to group the EU data portals. Figure 5.4 shows the clustering dendrogram, which is the result of the hierarchical clustering algorithm. In addition, Figure 5.5 shows the best clustering solution for k-means (k = 5) using a cluster profiling plot in parallel coordinates (the clustering profiling plot obtained with K-medians is almost identical). Parallel coordinates are a frequent way of visualising how the Open Government Data Initiatives differ from each other across factors.

Moreover, the composition of Open Data initiatives that form each cluster for K-means, K-medians and hierarchical clustering is shown in Table 5.6. In general, the techniques converge for the identification of 5 groups that are clearly distinguished from each other based on their behaviour

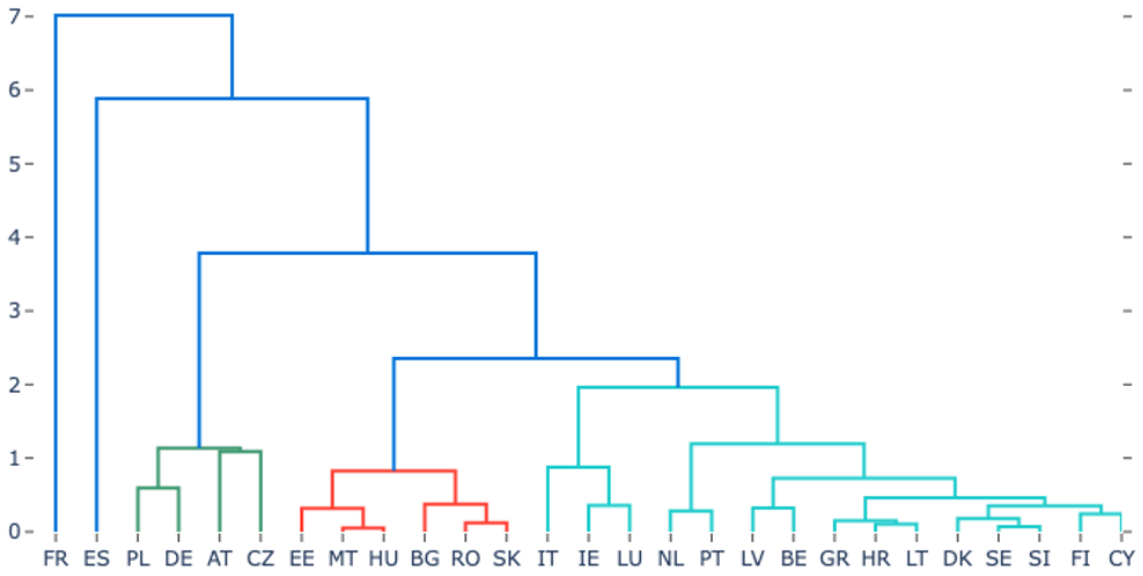

Fig. 5.4 Cluster dendrogram

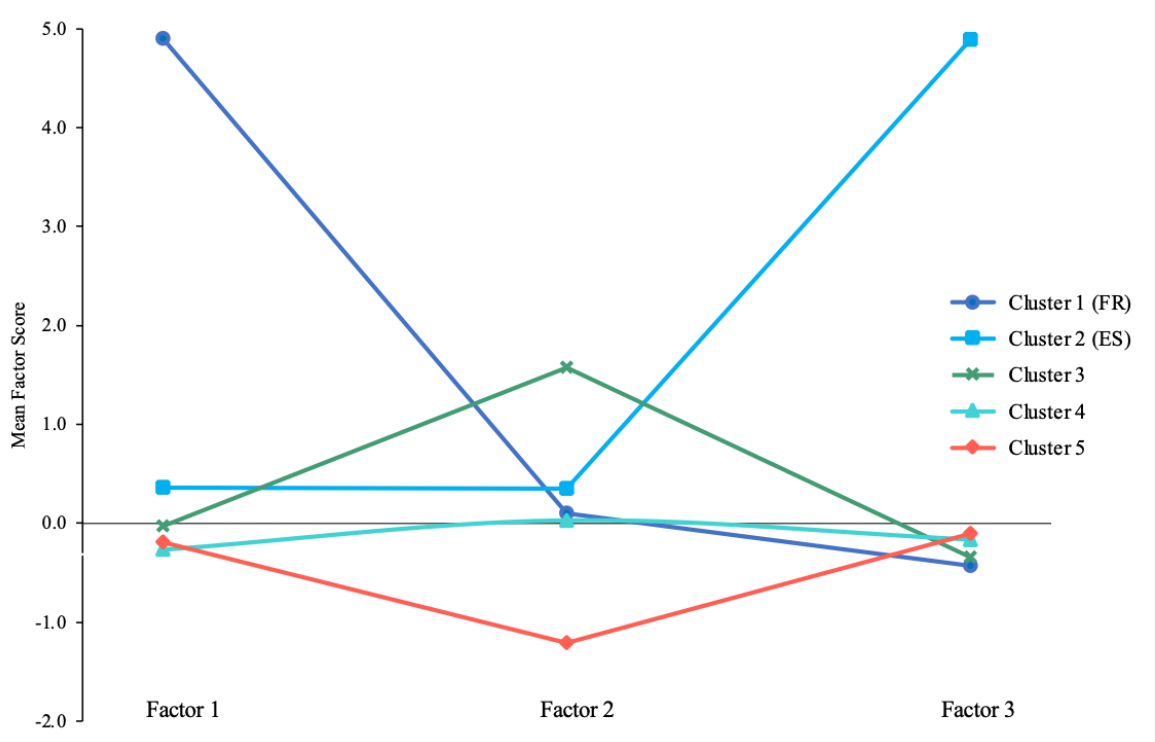

Fig. 5.5 Cluster profile plot for mean factor score of the clusters obtained with K-means

through the factors. The resulting assignments of hierarchical clustering (with a cutoff distance equal to 2) and K-medians were identical. K-means reallocated one member (NL).

The first cluster is determined by the preeminent performance of France in factor 1. In the second

cluster, Spain stands out for its score in factor 3, and has a slightly higher value for factor 1. Cluster 4 is the most numerous and shows an intermediate behaviour in the three factors. Cluster 3 and cluster 5 are characterised respectively by high and low values of the variables in factor 2.

Table 5.6 Cluster membership

Cluster	K-means	K-medians	Hierarchical Clustering
1	FR	FR	FR
2	ES	ES	ES
3	AT, PL, NL, DE, CZ	AT, PL, DE, CZ	AT, PL, DE, CZ
4	IT, IE, LV, LU, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT	IT, IE, LV, LU, NL, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT	IT, IE, LV, LU, NL, GR, FI, DK, PT, CY, SE, HR, BE, SI, LT
5	BG, RO, EE, MT, SK, HU	BG, RO, EE, MT, SK, HU	BG, RO, EE, MT, SK, HU

5.3.2 Results for Temporal Analysis of OGD Portals

This section presents the outcomes of applying the proposed methodology on the national Open Data portals of 27 EU member countries and their online social network activities on *X* during the temporal range from year 2017 to year 2021. This represents an input dataset consisting of 135 records: each record describes the status of the 27 national initiatives at each of the analysed years. It must be noted that Hungary, also belonging to the European Union, has not been considered in this study because there is not an official open data portal.

Figure 5.6 presents an overview of the input records and the values contained in the 7 considered variables over a bi-dimensional space using principal component analysis (PCA). This figure helps to guess the clouds of points that could be the origin of the clusters that are later identified. PCA is a method for reducing the number of dimensions that is often used to convert complicated data into a space with fewer dimensions while maintaining the fundamental variance of data [169]. In can be observed that there are 3 separate clouds of points in this graphical representation: a small cloud of points on the left upper corner; a small cloud of points on the right side; and a bigger cloud of points on the left side. As indicated in the description of our methodology, the core tool for the analysis of input data is the generation of a SOM map combined with the clustering of map nodes. The first parameter to be decided for the generation of a SOM map is its dimension, i.e. the number of rows and columns in this map. According to the recommendations of Vesanto and Alhomieni [170], we decided to approximate the number of neurons (cells) in the map to $5 \times \sqrt{data\ input\ size}$. Therefore, we selected 8×7 neurons, i.e. 8 rows and 7 columns in a bi-dimensional map.

Figure 5.7 shows the SOM (8×7 dimensions) map obtained after training the SOM neural network with 100,000 steps (epochs). This figure also helps in the identification of the 4 clusters grouping the nodes. The SOM map nodes represent the classification of records (initiatives with a particular status at each year) in the different output neurons. The colour gradation in these nodes provides an

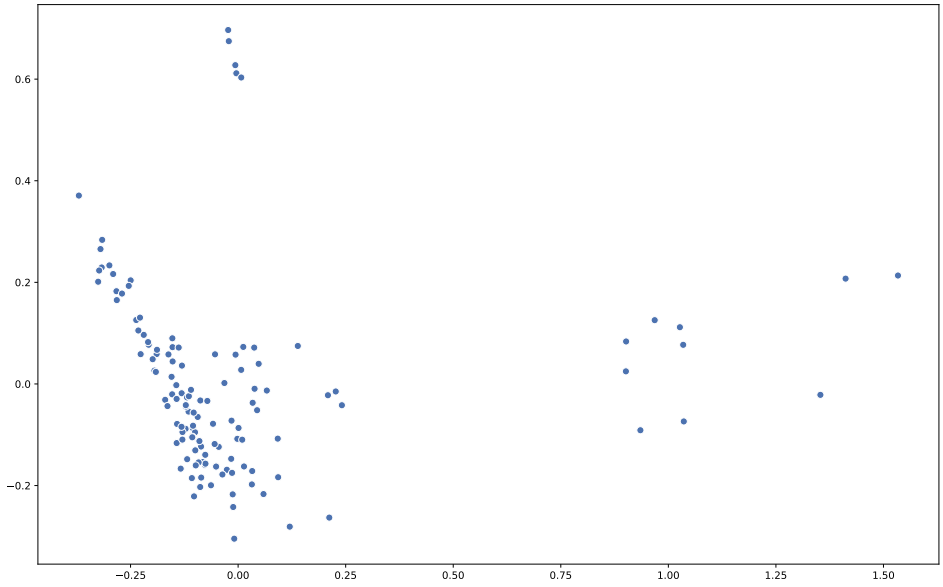

Fig. 5.6 Dispersion of records in input dataset after applying PCA over a bi-dimensional space

indication of the distance of variable values with respect to the neurons in the neighbourhood. The map in Figure 5.7 also displays the classification of the neurons (node maps) in four clusters after applying an agglomerative clustering algorithm.

We performed our experiments as exploratory analysis on the different values of k (number of clusters) to see the optimal result. We tested from $k = 2$ until $k = 7$ and we found that with $k = 4$ we get the optimal results for this study. Hence, we chose $k = 4$ as the number of clusters. For a better selection of the number of clusters that identify similar levels of development in Open Data initiatives we also generated a dendrogram (see Figure 5.8). A dendrogram is a complementary output to verify that the selected number of clusters groups appropriately the records of the input.

In order to have a better understanding of the meaning of these clusters, Figure 5.9 provides a cluster profiling plot with the average normalised score of the seven considered variables for each cluster which is represented by a line. Some details of these clusters are as follows:

- Cluster 0 has the highest values for most of the variables. This cluster contains the largest proportions of number of tweets (NT), user tweets (UT), and tweets from the portal (TFP), showing that this category sees a substantial level of X activity overall. In addition to this, it stands out in terms of the number of interactions (NI), which suggests that the information included inside this cluster produces a significant amount of engagement. The comparatively high values for Google Scholar (GS) mentions, Open Data Maturity (ODM) Score, and number

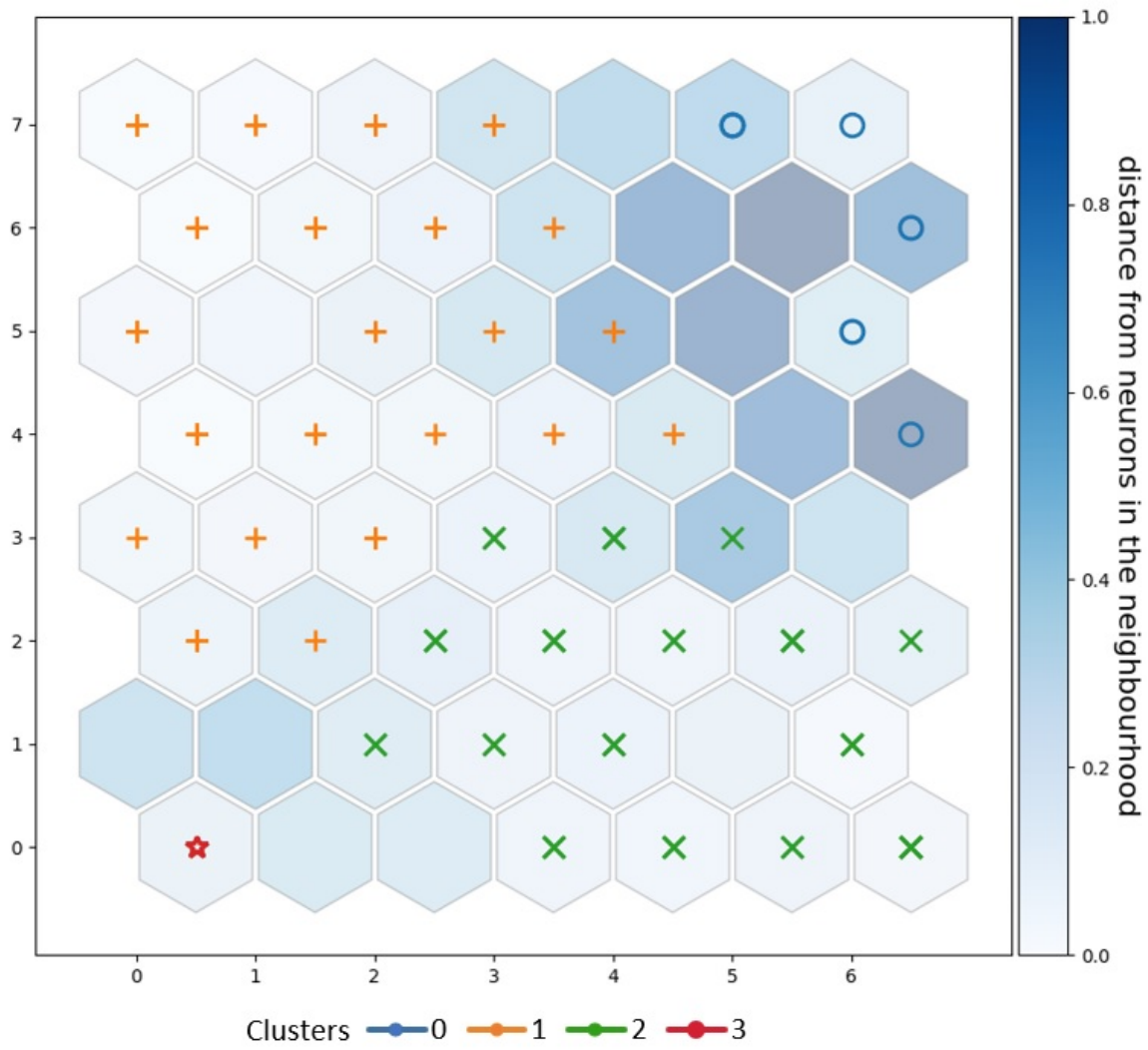

Fig. 5.7 Output SOM with 4 clusters and 8×7 dimension

of datasets (ND) all point to the fact that this cluster is academically acknowledged, advanced in terms of open data policies, and abundant in datasets that are readily accessible.

- Cluster 1, on the other hand, appears to reflect lower values than cluster 0 across the board of all parameters. This cluster shows a decrease in the number of tweets (NT), user engagements (UT), and portal activity (TFP). In addition, the number of Google Scholar (GS) mentions, the Open Data Maturity (ODM) Score, and the number of datasets (ND) associated with this cluster are all much fewer than those associated with the initiatives classified in cluster 0. Based on this information, cluster 1 seems to have less impact, a less developed set of open data standards, and maybe fewer datasets available.

- Clusters 2 and 3 are situated below clusters 0 and 1 for most of the considered variables:

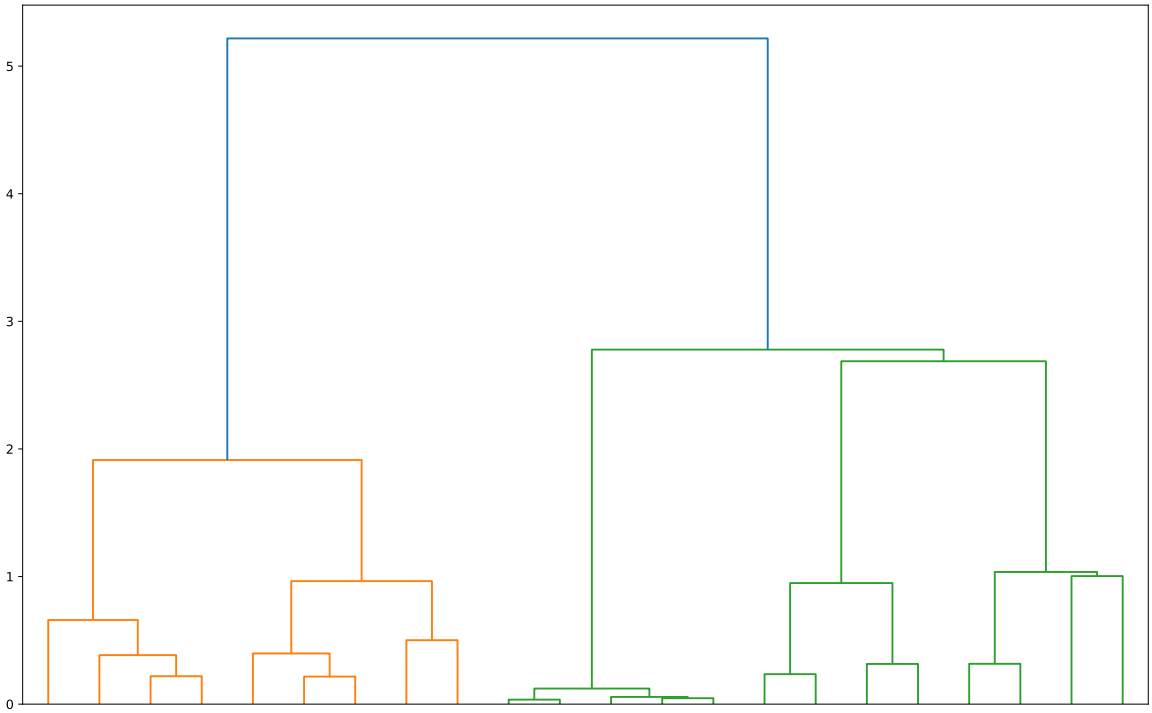

Fig. 5.8 Cluster dendrogram

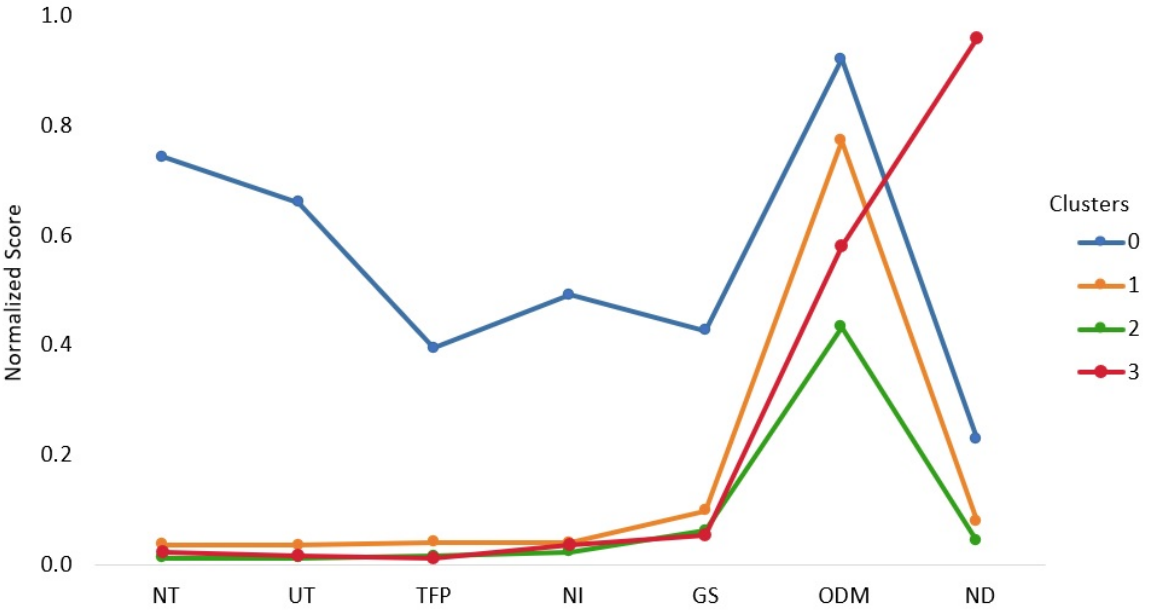

Fig. 5.9 Profiling of the Clusters

– Cluster 2 reveals a moderate presence of tweets (NT), user interaction (UT), and portal activity (TFP), in addition to relatively modest values for Google Scholar (GS) mentions,

Open Data Maturity (ODM) Score, and number of datasets (ND). This cluster also reveals a moderate presence of open datasets.

– Cluster 3 provides somewhat higher statistics across all parameters compared to cluster 2, showing a better degree of activity, recognition, maturity, and dataset availability. Its relevance within the existing context is shown by the fact that the even number of datasets is greater than that of the other three clusters.

In addition, for each variable we conducted an ANOVA test to detect the presence of differences between clusters. In the cases where this test was significant ($p < 0.05$), we accompanied it with post hoc tests to compare the clusters and detect the precise origin of the differences. We observed that for the seven variables the ANOVA test was significant, which means that at least one of the clusters is statistically different from the rest. For the NT, UT, TFP, NI and GS variables, the post hoc test revealed that the cluster 0 is statistically different from the rest and there is no difference between the rest of the clusters. For the ODM and ND variables, all the comparisons were significant, which means that all the clusters have significantly different values.

Making a deeper analysis of the countries behind the initiatives and years grouped in the different nodes and clusters displayed in the map shown in Figure 5.7, it can be derived the following:

- The red star in cluster 3 represents the open data portal for the Czech Republic. Based on the findings of our experiment, this is an anomaly or an outlier because the number of datasets (ND) published each year in the Czech Republic open data portal are significantly higher than in the other EU Open Data portals.

- The Open Data portals of Spain and France are clustered together in the blue circles of cluster 0, primarily due to their consistently high values across various selected variables. France and Spain stand out with the highest nominal values for key aspects, such as X (Twitter) discussions (NT, TFP, UT, and NI) and Google Scholar (GS) mentions, reflecting their greater significance within the open data portal landscape.

- Meanwhile, the development of other Open Data portals is represented with orange plus signs in cluster 1 and green crosses in cluster 2.

The distribution of records representing the initiatives at different years can be also observed in Figure 5.10, which is equivalent to Figure 5.6 displaying the records in a bi-dimensional space using PCA. The novelty of Figure 5.10 is that we highlight now the assignment of records (points in the plot) to the four clusters obtained after applying the clustering algorithm on map nodes. The records corresponding to the Czech Republic in cluster 3 are the points on the left-upper corner. The points corresponding to the records of Spain and France in cluster 0 are the ones on the right side. The points corresponding to clusters 1 and 2 are on the left side. The points of cluster 1 corresponding to initiatives with a relatively more mature status are closer to the horizontal axis.

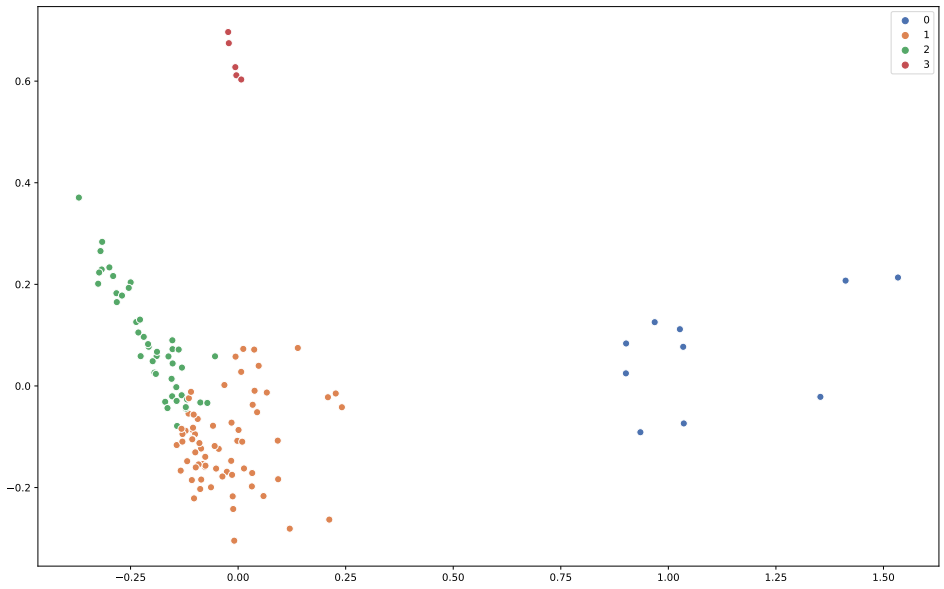

Fig. 5.10 Dispersion of records in input dataset after applying PCA over a bi-dimensional space, with assigned cluster

Another advantage of using this approach for the analysis of the evolution of OGD initiatives is the possibility of analysing the trajectories or movements of initiatives along the clusters in the studied time period. Table 5.7 shows the evolution of each country's open data portal along the 5-year time frame period. Some observations can be highlighted about countries remaining or moving between the clusters:

- With respect to countries remaining on the same cluster, it is worth noting that France and Spain have stayed in cluster 0 from 2017 to 2021. This is due to the fact that both of these countries are doing extremely well in the field of open data and in open data society. In a similar fashion, Austria, Cyprus, Ireland, Italy, the Netherlands, and Slovenia have all remained in cluster 1 over the whole of the time period covered by the experiment in this research study. In addition, given the values of the chosen variables that were taken into consideration for the purposes of this study, the open data portal for the Czech Republic is an outlier that has not shifted from cluster 3 over the temporal range of years from 2017 to 2021. This is because the open data portal for the Czech Republic is the only one that publishes a higher number of datasets (ND) than any other country, making it an outlier, as can be clearly seen from Figure 5.9. Although this cluster contains only one country, we decided to include it because it is representative of a strategy that prioritises the quantity of datasets over other variables.
- On the other hand, there have been some movements between cluster 1 and cluster 2, or vice

Table 5.7 Cluster shifting of the countries

Country	2017	2018	2019	2020	2021
AUSTRIA	1	1	1	1	1
BELGIUM	2	1	2	2	2
BULGARIA	1	1	2	2	2
CROATIA	1	2	1	1	1
CYPRUS	1	1	1	1	1
CZECH	3	3	3	3	3
DENMARK	2	2	1	1	1
ESTONIA	2	2	2	1	1
FINLAND	1	2	1	1	1
FRANCE	0	0	0	0	0
GERMANY	2	1	1	1	1
GREECE	1	1	2	1	1
IRELAND	1	1	1	1	1
ITALY	1	1	1	1	1
LATVIA	2	1	1	1	2
LITHUANIA	2	2	2	1	1
LUXEMBOURG	1	1	2	2	2
MALTA	2	2	2	2	2
NETHERLANDS	1	1	1	1	1
POLAND	2	1	1	1	1
PORTUGAL	2	2	2	2	2
ROMANIA	1	2	2	2	2
SLOVAKIA	1	1	2	2	2
SLOVENIA	1	1	1	1	1
SPAIN	0	0	0	0	0
SWEDEN	2	2	2	1	1

versa. Denmark, Estonia, Germany or Poland started in cluster 2 and moved to cluster 1 in the following years. This can be interpreted as an improvement in the maturity status of their initiatives (value of ODM score), as the values of variables in cluster 1 are higher than those in cluster 2. On the contrary, Bulgaria, Luxembourg, Romania and Slovakia started in cluster 1 and move to cluster 2, which can be interpreted as a decrease in their maturity status.

- Last, several national Open Data portals have moved from an initial cluster to another cluster, and then returned back to the initial cluster. For instance, countries like Belgium started out in cluster 2 in the year 2017, moved up to cluster 1 in the year 2018, and finally returned back to cluster 2 for the next three years. The same thing happened to countries like Croatia, but with a different cluster: Croatia started out in cluster 1 in 2017, moved up to cluster 2 in 2018, and then shifted back to cluster 1 for the next three years.

The full code of the experiments is available on a GitHub repository.[3] We employed external Python libraries as well as functions that we developed ourselves.

[3]Github Repository https://github.com/abdulunizar/Evolution-of-OGD-Initiatives/

5.4 Discussion

The feasibility of our analysis methodology was tested by evaluating the development of 27 European OGD portals between 2017 and 2021. Using as input the values of the selected variables at a yearly basis, we were able to compare the output of our methodology with the conclusions reported in the Open Data Maturity Report for the years ranging from 2017 to 2021.

In the course of our experiment, we used a SOM-based model and a clustering algorithm in order to identify different maturity levels in the evolution of OGD initiatives according to the variables that were selected. Our first observation is that cluster 0 in Figure 5.9 could be considered as "user community driven" because the initiatives in this cluster have the highest values for many variables and the highest concentrations of Number of Tweets (NT), User Tweets (UT), and Tweets from the portal (TFP), all of which indicate a high volume of *X* activity. Not only does the content within this cluster generate a large number of interactions (NI), but it also stands out in terms of engagement volume. In addition, it is highly interdependent. The relatively high values for Google Scholar (GS) mentions, Open Data Maturity (ODM) Score, and Number of Datasets (ND) indicate that this cluster is well-known among academics, mature in terms of open data policy, and abundant in freely accessible datasets. Cluster 0 Open Data portals are in the greatest position overall and may be defined as user-centric, consistent with the findings of the 2017–2021 Open Data Maturity Report. This may be largely attributable to the activities of the Open Data portals in cluster 0. These initiatives monitor user feedback through multiple channels, such as message forums dedicated to each dataset (e.g., in case of the French open data portal). The Open Data Maturity Reports of 2017–2021 highlight the efforts of cluster 0 Open Data portals to cultivate editorial content, improve search and findability, and make active use of social media platforms such as Facebook, LinkedIn, YouTube, and Flickr. In addition, from Figure 5.9 it can be observed that cluster 3 includes open data initiatives with a remarkable number and quantity of released datasets. However, these activities have no direct effect on *X* activity. This is likely due to the fact that the bodies responsible for coordinating these efforts are not as effective as they could be at promoting their work on social media platforms. Lastly, we are able to determine that the open data initiatives in clusters 1 and 2 report a moderate level of quality and social network activity. This most likely indicates that the initiatives in question are improving and in their infancy, making them less appealing to potential consumers. In addition, it is worth noting the work involved in collecting the values of the various variables that were special for this research. The information may be used by governments and decision-makers to assess Open Data initiatives based on the level of user participation that they have received. In particular, the variables on X activity indicators (NT, UT, and NI) and Google Scholar mentions (GS) were specifically chosen to capture the user engagement. Moreover, it must be noted that to increase the applicability of our experiment and generate a better distribution of initiatives in the SOM map, we decided to normalise the raw values of the variables dividing them by the logarithm of the population of each country in each of the analysed years. This allowed us to decrease the effect of comparing initiatives in countries with big differences in terms of population and governmental complexity. The only exception was

the ODM variable, as this indicator is a score assigned manually by experts considering the overall context of each initiative and country. During the development and testing of our methodology, we also tested the inclusion of a variety of other variables and their permutations, such as the sentiment score of tweets from each country over the span of each year, both individually and collectively, and the number of likes from each year in each country and number of positive tweets of each country for our temporal range of years from 2017 to 2021. However, as a consequence of their negative impact on the generation of the SOM map and the clustering phase, certain variables and combinations were eliminated.

5.5 Summary

This chapter introduces an innovative, data-driven methodology to evaluate the evolution and user engagement of European OGD initiatives. By combining snapshot-based assessments with longitudinal temporal analysis, and integrating both portal characteristics and public discussions captured through social media activity on platform X (formerly Twitter), this work provides a comprehensive view of how OGD portals perform and connect with their user communities over time. The research addresses the lack of standardised, comparable user engagement across OGD portals by proposing a framework that integrates multidimensional data with temporal analysis. A key innovation lies in the combined application of Self-Organizing Maps (SOM) with agglomerative clustering and Factor Analysis technique to identify developmental trajectories and engagement patterns across 27 national OGD portals. The temporal range for the Self-Organizing Maps (SOM) is from 2017 to 2021 whereas for the Factor Analysis technique is 2021.

The findings reveal distinct clusters of OGD initiatives, highlighting varying degrees of user engagement and portal maturity. Notably, one cluster characterised by high tweet activity, strong academic visibility, and mature policy frameworks, emerged as "user community driven," aligning with best practices identified in Open Data Maturity Reports. Contrarily, other clusters displayed limited interaction despite significant dataset publication, suggesting a gap between data supply and public impact. The study also emphasises the added value of normalising activity metrics by population size to account for structural differences among countries, thereby enabling a more equitable comparison.

Despite the contributions of this study, the methodology has certain limitations. It relies on publicly accessible data that are under access restrictions, limiting public availability and ease of access. Additionally, while SOM and clustering techniques facilitate visual and structural analysis, they still require manual interpretation to derive actionable insights. Future work may explore content-based tweet analysis to capture more nuanced dimensions of user feedback, extend the framework beyond Europe and evaluate broader engagement indicators across regions and time. Furthermore, we can perform sentiment analysis on the available X data, categorizing each tweet into one of three classes: positive, negative, or neutral. These advancements could support policymakers in designing more user-responsive and impactful OGD ecosystems.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

This chapter presents the overall discussion, key contributions, and limitations of the research, followed by directions for future work. It synthesises the findings from the previous chapters, highlighting how the proposed frameworks contribute to building sustainable and user-centered open data ecosystems. It also reflects on how the proposed solutions address key gaps in current practices and support more inclusive and dynamic interactions between data providers and users. Finally, the chapter outlines future research directions.

6.1 Summary of contributions

This dissertation aimed to investigate methodologies for enhancing inclusiveness in OGD portals, with a tangible focus on data accessibility and user interaction. To be more precise, accessibility is treated in this thesis as the tangible focus of analysis, while inclusiveness represents its broader implication. By improving dataset discoverability, creating feedback mechanisms, and extending user engagement beyond portals, the research demonstrates ways to make OGD ecosystems more inclusive of diverse stakeholders. This distinction clarifies that inclusiveness is not directly engineered as a technical property but emerges from the combined effect of accessibility and user participation.

In this final chapter, the findings are synthesized to show how the objective of the research study has been achieved and how each research question has been answered, together with the scientific and practical implications of the study. Moreover, this dissertation addresses two main challenges in bringing inclusiveness to open data portals. The first is to ensure the easy discovery of the resources by various stakeholders. These stakeholders, including government agencies, researchers, citizens, often have different needs, levels of technical expertise, and contextual requirements. The second is to facilitate responsive interaction when resources do not meet user needs. Furthermore, Figure 7.1 provides a bird's eye view of this dissertation, showing how conceptual foundations (Chapter 2), supplier-driven approaches (Chapter 3), user-driven feedback mechanisms (Chapter 4), and evaluation of OGD initiatives (Chapter 5) are connected.

Chapter 2 laid the foundation of this thesis by exploring the origins and definitions of open government, open data, open government data, open data ecosystems, and open data portals. The

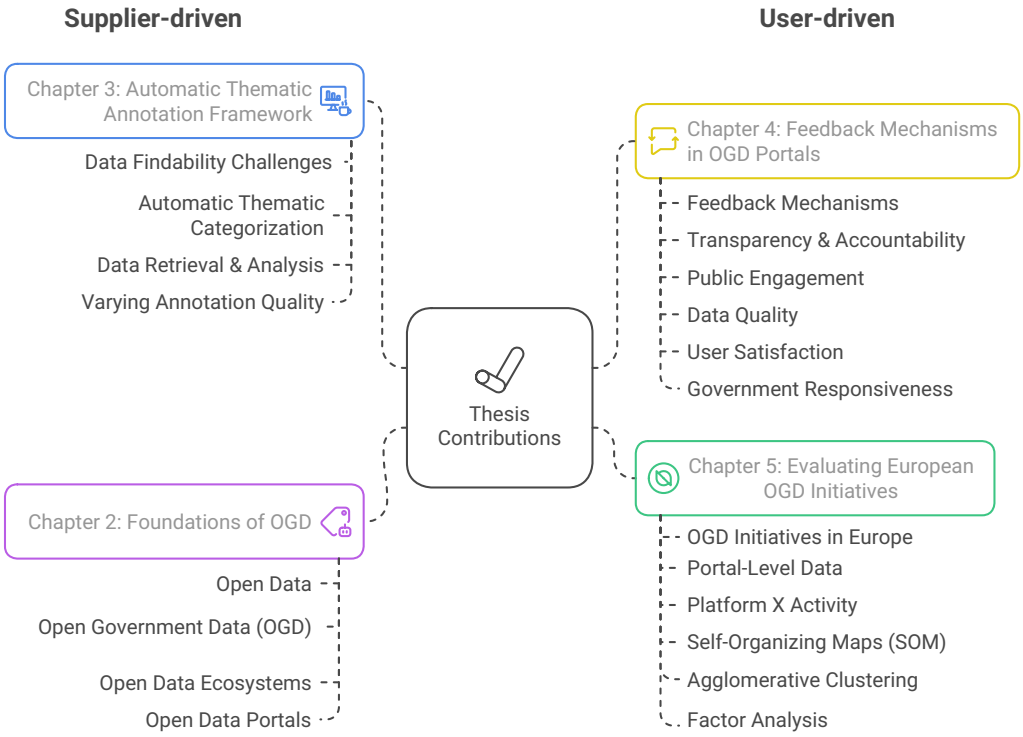

Fig. 6.1 A bird's eye view of the contributions

findings of this chapter are published in the Doctoral Consortium Workshop, 22nd International Conference on Perspectives in Business Informatics Research [149].

Chapter 3 is one of the central contributions of this thesis which aims at the development of a robust framework for the automatic thematic annotation of OGD. The framework addresses a fundamental challenge in data findability across local, national, and international open data portals. Given the pyramidal architecture of many OGD portals, where datasets are harvested and aggregated across layers. Hence, thematic coherence and metadata accuracy are vital to ensure meaningful access for users. This work acknowledges the diversity in annotation quality across portals and provides strategies for both well-annotated and poorly annotated metadata corpus.

The framework has two methods. One is supervised and another is unsupervised. In the supervised method, the framework explores combinations of metadata properties (title, description, keywords) and various techniques for feature representation and classification. In unsupervised method, embedding-based similarity between datasets and theme definitions provides an effective alternative for portals with minimal manual labeling. This dual strategy ensures the applicability of the method across a

range of portal maturity levels, enabling enhanced thematic structure and improved user navigation throughout the open data landscape.

Experimental validation on a representative sample from the European Data Portal confirmed the feasibility of both approaches. These findings support the broader thesis perspective that theme classification is a critical enabler for open data usability and findability. The findings presented in this chapter formed the basis for the preparation and submission of a journal article [171].

Chapter 4 explores the role of feedback mechanisms in OGD portals as a means to support greater transparency, accountability, and public engagement. OGD portals act as central gateways to government data at various administrative levels. However, to ensure their effectiveness, it is crucial to go beyond simply offering access to datasets. This work focuses on analysing how feedback mechanisms are designed and implemented, and how they can be improved to create a more dynamic, responsive relationship between data publishers and users.

To understand the current state of feedback in OGD portals, the study begins by proposing a conceptual framework that defines both input and output feedback channels. This is followed by detailed case studies and a broader analysis of OGD portals using automated techniques to detect existing feedback flows. Input feedback channels such as email, feedback forms, comments, and dataset ratings provide users with direct means to express their specific needs, suggestions, and concerns. These channels are invaluable in capturing granular user perspectives that inform data quality and portal improvements. Output feedback channels, including blogs and newsletters, are primarily one-way communication tools designed to disseminate updates and announcements. While they play a critical role in maintaining user awareness, they do not facilitate interactive engagement. Input/output channels like social media and discussion forums, on the other hand, serve as bi-directional platforms that enable real-time interactions and foster collaborative contributions from users. This bidirectional view highlights opportunities for portals to evolve into more participatory and accountable platforms. The findings reveal wide differences in how portals allow users to provide feedback and how, in turn, data publishers respond. The findings of this chapter are published in the journal of Online Information Review [172].

Chapter 5 introduces a framework for evaluating the evolution and user engagement of OGD initiatives throughout Europe. This study examines the existing gap in standardised and comparable engagement metrics by integrating descriptive portal-level data with publicly accessible activity on platform X (formerly Twitter). The methodology incorporates Self-Organizing Maps (SOM) with agglomerative clustering, and Factor Analysis technique, facilitating the identification of engagement patterns and development trajectories across 27 EU national OGD portals Europe. This research involves the years 2017 to 2021 for the study of SOM and concentrates on 2021 for the Factor Analysis technique, offering both temporal and structural insights into the growth and interaction of OGD portals with their user communities.

The analysis identifies three distinct clusters of national OGD portals. A notable group is distinguished by high user activity on X, strong academic linkage, and well-established policy frameworks, driven by engaged user communities. This cluster corresponds with the best practices outlined in the

Open Data Maturity Report published by the EU. In contrast, other clusters exhibit minimal social media engagement despite generating significant amounts of data, highlighting a disparity between data availability and public utilisation. A key contribution of this work is the use of normalised engagement indicators, adjusted for population size, which helps to enable fairer cross-country comparisons and highlights structural challenges within smaller or less mature data ecosystems. The findings of this chapter are published in the International Conference on Theory and Practice of Digital Libraries [173] and in IEEE Access journal [174].

While the methodology provides valuable insights, it is not without limitations. The reliance on publicly accessible data introduces constraints, particularly as access to social media data becomes increasingly restricted. The temporal dimension of the results of this chapter is a critical factor in interpretation. The data collected from 2017 to 2021 is the foundation for numerous analyses, including the evaluation of feedback mechanisms and the large-scale study of European OGD portals. For this contribution we started working and collecting data using the X (formerly Twitter) in 2022 and started experiments but later X, eliminated its open access API, replacing free access with a paid tier. Hence we already had the data for the temporal period (from 2017 to 2021) of the selected years and we continued our experiments with that collected data. Since then, open data practices, portal technologies, and policy frameworks have evolved significantly, which could affect the applicability of the findings of this study to the present moment. For instance, the introduction of new European directives on data governance, enhancements in portal software, and the increasing utilization of artificial intelligence for metadata curation may have resulted in modifications that were not accounted for in this research study. Moreover, the experiments have been performed on data that is open to the public, and the reliability of the results is dependent on the quality of the data sources. But in some case, missing values, inconsistent metadata fields, or limited documentation posed challenges for analysis. In addition, our method of clustering is predicated on a number of different characteristics. Nevertheless, there may be more important variables that have an impact on the level of user participation and data compliance.

Our research study has both theoretical and practical implications for the field of OGD. On the theoretical side, it advances understanding in three areas: (i) how structured thematic annotations can improve dataset findability and enrich metadata quality; (ii) how feedback mechanisms in OGD portals function as drivers of transparency, accountability, and public engagement; and (iii) how OGD portals evolve over time in terms of inclusiveness and user interaction. These insights extend prior work on open data ecosystems by offering new methodological tools (e.g., automated annotation) and conceptual frameworks (e.g., roundtrip feedback).

On the practical side, the findings suggest concrete applications for policymakers, portal managers, and developers. For instance, the thematic annotation framework can be adapted as a plugin for widely used open data platforms such as CKAN, DKAN, or Socrata, provided that the themes for classification are defined in alignment with portal needs. The results on feedback mechanisms also have direct implications for the governance of OGD initiatives: portals with similar features can benchmark themselves against others and adjust their user engagement policies accordingly. Moreover,

the guidelines derived from the input and output feedback channels studied in this research can help administrative teams identify underutilized communication channels and initiate new interaction flows with end users. In this way, the study provides both conceptual clarity and actionable strategies for fostering more inclusive and user-responsive OGD portals.

This PhD research was carried out as part of the "ODECO: Towards a Sustainable Open Data Ecosystem" project. ODECO was a four-year Innovative Training Network funded under the Horizon 2020 Marie Skłodowska-Curie programme (H2020-MSCA-ITN-2020, Grant Agreement No. 955569). The overarching goal of the project was to tackle both current and emerging challenges in building open data ecosystems that are inclusive, user-centric, and based on circular principles. Alongside the original research contributions presented in this dissertation, the project also enabled active engagement in collaborative academic and outreach activities, particularly focusing on open data infrastructure and user experience enhancement. Additionally with the collaboration of ODECO colleagues we have published a conference paper in the Proceedings of the 25th Annual International Conference on Digital Government Research [175] and an article in the journal of Information Polity [176].

6.2 Open Issues

While this dissertation has presented several frameworks to improve thematic classification, feedback mechanisms, and engagement analysis in Open Government Data (OGD) ecosystems, there remain multiple avenues for future research. The dynamic and socio-technical nature of open data platforms requires continuous adaptation to emerging technologies, evolving user expectations, and shifting policy landscapes. Building upon the foundational work presented in this thesis, future directions can further enhance the sustainability, responsiveness, and inclusiveness of OGD ecosystems by addressing both technical and human-centric dimensions.

Chapter 3 aims at the development of a robust framework for the automatic thematic annotation of OGD. As a future work in this chapter, it was suggested that it would be better to explore if the information related to the application schema of the different distributions of datasets can help us to improve the automatic thematic classification of datasets. Available distributions in machine readable formats such as CSV or RDF can provide in some cases meaningful names of thematic attributes of the dataset content. Even in the case of RDF (graph data), these attributes are usually selected from well-known vocabularies, and this may be used to infer links with the themes that can be assigned automatically.

Chapter 4 explores the role of feedback mechanisms in OGD portals. Future research could focus on developing reliable metrics to evaluate the success and impact of feedback mechanisms in OGD portals. Indicators such as response rates from data publishers, the emergence of discussions in dedicated forums, and levels of user satisfaction after submitting feedback can offer valuable insights into system effectiveness. Furthermore, user profiling is a potential path, especially in terms of understanding how various stakeholder groups (such as students, journalists, non-governmental

organisations, commercial enterprises, and intermediaries) interact with feedback channels. This understanding could possibly lead to the development of strategies that can be more appropriately tailored and efficient. Lastly, when it comes to expanding the adoption and effectiveness of feedback systems, complementary activities such as co-creation initiatives, targeted outreach, and trust-building campaigns may also play a significant role.

Chapter 5 introduces a framework for evaluating the evolution and user engagement of OGD initiatives. A promising future direction involves conducting a more detailed content analysis of user interactions on platforms like X (formerly Twitter) using sentiment and semantic analysis techniques. This could reveal deeper insights into user engagement, including recurring concerns, reuse examples, and suggestions for improvement. Although initial results suggest limited influence of sentiment features on cluster construction, categorizing tweet content may still help uncover nuanced patterns of interaction. Expanding this framework to include OGD portals outside Europe and analyzing longer time periods would enable more comprehensive comparisons across regions. Such work could inform policymakers and practitioners about regional dynamics and best practices, contributing to the development of more responsive and user-driven open data ecosystems.

Finally, it must be acknowledged that the case studies used for the validation of the contributions in this thesis have focused on analysing the characteristics and metadata content of the European data portal (data.europa.eu) and the national portals of the European Union member states that it aggregates. However, the proposals and experiments with these case studies could be also applicable to OGD portals in other countries with the necessary work of customization to the features of each portal. Furthermore, the possibility of adapting the contributions to non-government data portals could also be considered as a future line of research.

CHAPTER 7

CONCLUSIONES Y TRABAJO FUTURO

Este capítulo presenta la discusión general, las principales contribuciones y las limitaciones de la investigación, seguido de orientaciones para trabajos futuros. Sintetiza los hallazgos de los capítulos anteriores, destacando cómo los marcos propuestos contribuyen a la construcción de ecosistemas de datos abiertos sostenibles y centrados en el usuario. También reflexiona sobre cómo las soluciones propuestas abordan lagunas clave en las prácticas actuales y respaldan interacciones más inclusivas y dinámicas entre proveedores de datos y usuarios. Finalmente, el capítulo esboza líneas futuras de investigación.

7.1 Resumen de contribuciones

Esta tesis doctoral tuvo como objetivo investigar metodologías para mejorar la inclusividad en los portales de DAG, con un enfoque en la accesibilidad de los datos y la interacción con los usuarios. Para ser más precisos, la accesibilidad se trata en esta tesis como el enfoque tangible de análisis, mientras que la inclusividad representa su implicación más amplia. Al mejorar la capacidad de descubrimiento de los conjuntos de datos, crear mecanismos de retroalimentación y ampliar la participación de los usuarios más allá de los portales, la investigación demuestra maneras de hacer que los ecosistemas de OGD sean más inclusivos para los diferentes actores implicados. Esta distinción aclara que la inclusividad no se concibe directamente como una propiedad técnica, sino que surge del efecto combinado de la accesibilidad y la participación de los usuarios.

En este capítulo final, los resultados se sintetizan para mostrar cómo se ha alcanzado el objetivo de la investigación y cómo se ha respondido a cada pregunta de investigación, junto con las implicaciones científicas y prácticas del estudio. Además, esta tesis aborda dos retos principales en la incorporación de la inclusividad en los portales de datos abiertos. El primero es garantizar el fácil descubrimiento de los recursos por parte de los distintos grupos de interés. Estos grupos, que incluyen organismos gubernamentales, investigadores y ciudadanos, a menudo tienen diferentes necesidades, niveles de experiencia técnica y requisitos contextuales. El segundo es facilitar una interacción receptiva cuando los recursos no satisfacen las necesidades de los usuarios. Además, la Figura 7.1 ofrece una visión global de esta tesis, mostrando cómo los fundamentos conceptuales (Capítulo 2), los enfoques

impulsados por los proveedores (Capítulo 3), los mecanismos de retroalimentación impulsados por los usuarios (Capítulo 4) y la evaluación de las iniciativas de DAG (Capítulo 5) están conectados.

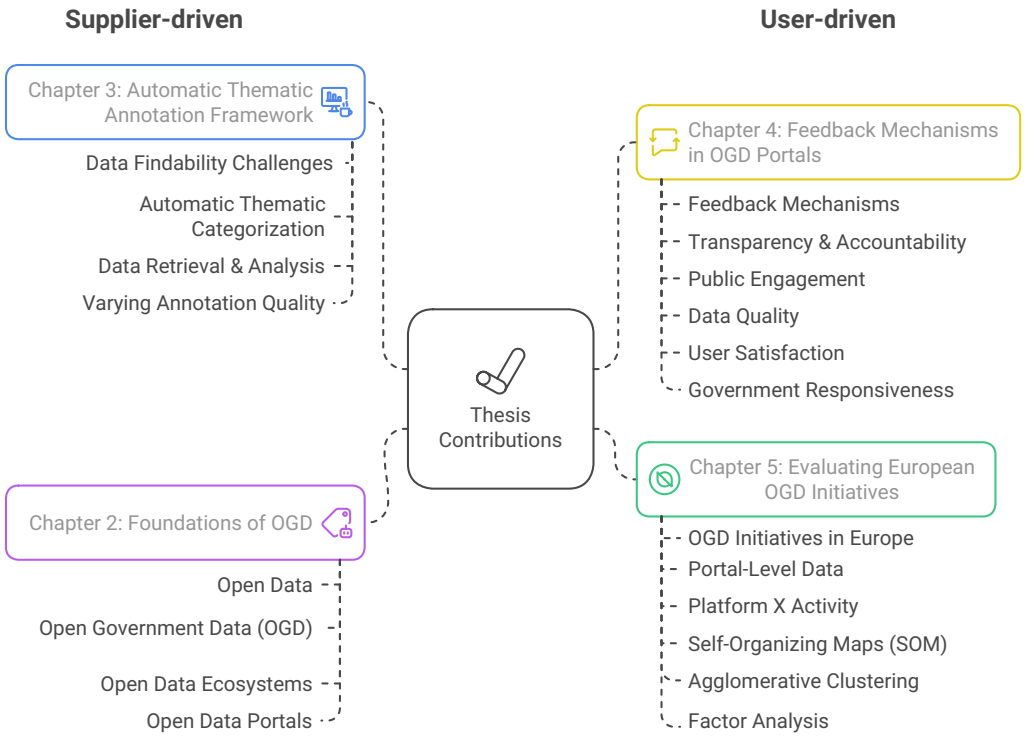

Fig. 7.1 Una visión global de las contribuciones

El Capítulo 2 sentó las bases de esta tesis explorando los orígenes y definiciones de gobierno abierto, datos abiertos, datos abiertos gubernamentales, ecosistemas de datos abiertos y portales de datos abiertos. Los resultados de este capítulo se publicaron en el Doctoral Consortium Workshop, 22nd International Conference on Perspectives in Business Informatics Research [149].

El Capítulo 3 constituye una de las contribuciones centrales de esta tesis, cuyo objetivo es el desarrollo de un marco robusto para la anotación temática automática de DAG. El marco aborda un desafío fundamental en la capacidad de descubrimiento de datos a través de portales locales, nacionales e internacionales de datos abiertos. Dada la arquitectura piramidal de muchos portales de DAG, donde los conjuntos de datos se cosechan y agregan en distintas capas, la coherencia temática y la exactitud de los metadatos son esenciales para garantizar un acceso significativo para los usuarios. Este trabajo reconoce la diversidad en la calidad de la anotación entre portales y proporciona estrategias tanto para corpus de metadatos bien anotados como para aquellos deficientemente anotados.

El marco incluye dos métodos: uno supervisado y otro no supervisado. En el método supervisado,

el marco explora combinaciones de propiedades de los metadatos (título, descripción, palabras clave) y diversas técnicas de representación de características y clasificación. En el método no supervisado, la similitud basada en embeddings entre conjuntos de datos y definiciones temáticas ofrece una alternativa eficaz para portales con etiquetado manual mínimo. Esta estrategia dual garantiza la aplicabilidad del método en diferentes niveles de madurez de portales, permitiendo una estructura temática mejorada y una navegación más eficiente de los usuarios en el ecosistema de datos abiertos.

La validación experimental en una muestra representativa del Portal Europeo de Datos confirmó la viabilidad de ambos enfoques. Estos resultados respaldan la perspectiva más amplia de la tesis de que la clasificación temática es un habilitador crítico para la usabilidad y la capacidad de descubrimiento de los datos abiertos. Los hallazgos presentados en este capítulo sirvieron de base para la preparación y envío de un artículo de revista [171].

El Capítulo 4 explora el papel de los mecanismos de retroalimentación en los portales de DAG como medio para apoyar una mayor transparencia, rendición de cuentas y participación pública. Los portales de DAG actúan como puertas de acceso centrales a los datos gubernamentales en varios niveles administrativos. Sin embargo, para garantizar su eficacia, es crucial ir más allá de simplemente ofrecer acceso a conjuntos de datos. Este trabajo se centra en analizar cómo se diseñan e implementan los mecanismos de retroalimentación y cómo pueden mejorarse para crear una relación más dinámica y receptiva entre los publicadores de datos y los usuarios.

Para comprender el estado actual de la retroalimentación en los portales de DAG, el estudio comienza proponiendo un marco conceptual que define tanto los canales de retroalimentación de entrada como de salida. A continuación, se presentan estudios de caso detallados y un análisis más amplio de portales de DAG utilizando técnicas automatizadas para detectar flujos de retroalimentación existentes. Los canales de retroalimentación de entrada, como el correo electrónico, los formularios, los comentarios y las valoraciones de conjuntos de datos, proporcionan a los usuarios medios directos para expresar sus necesidades, sugerencias y preocupaciones. Estos canales son invaluables para captar perspectivas granulares de los usuarios que contribuyen a mejorar la calidad de los datos y de los portales. Los canales de salida, como los blogs y boletines, son principalmente herramientas de comunicación unidireccionales diseñadas para difundir actualizaciones y anuncios. Si bien desempeñan un papel importante en el mantenimiento de la atención de los usuarios, no facilitan un compromiso interactivo. Por otro lado, los canales de entrada/salida, como las redes sociales y los foros de discusión, sirven como plataformas bidireccionales que permiten interacciones en tiempo real y fomentan contribuciones colaborativas de los usuarios. Esta visión bidireccional resalta oportunidades para que los portales evolucionen hacia plataformas más participativas y responsables. Los hallazgos revelan grandes diferencias en cómo los portales permiten a los usuarios proporcionar retroalimentación y cómo, a su vez, los publicadores de datos responden. Los resultados de este capítulo se publicaron en la revista Online Information Review [172].

El Capítulo 5 introduce un marco para evaluar la evolución y la participación de los usuarios en las iniciativas de DAG en toda Europa. Este estudio examina la brecha existente en métricas estandarizadas y comparables de participación integrando datos descriptivos a nivel de portal con

la actividad accesible públicamente en la plataforma X (anteriormente Twitter). La metodología incorpora Mapas Autoorganizados (SOM) con clustering aglomerativo y técnica de Análisis Factorial, lo que facilita la identificación de patrones de participación y trayectorias de desarrollo en 27 portales nacionales de DAG de la UE. Esta investigación abarca los años 2017 a 2021 para el estudio de SOM y se centra en 2021 para la técnica de Análisis Factorial, ofreciendo tanto perspectivas temporales como estructurales sobre el crecimiento y la interacción de los portales de DAG con sus comunidades de usuarios.

El análisis identifica tres grupos distintos de portales nacionales de DAG. Un grupo destacado se caracteriza por una alta actividad de usuarios en X, fuertes vínculos académicos y marcos normativos consolidados, impulsados por comunidades de usuarios activas. Este grupo corresponde con las mejores prácticas descritas en el Open Data Maturity Report publicado por la UE. En contraste, otros grupos muestran un compromiso mínimo en redes sociales a pesar de generar grandes volúmenes de datos, lo que pone de manifiesto una disparidad entre disponibilidad de datos y su uso público. Una contribución clave de este trabajo es el uso de indicadores de participación normalizados, ajustados por tamaño de población, lo que permite comparaciones más justas entre países y resalta los desafíos estructurales de ecosistemas de datos más pequeños o menos maduros. Los resultados de este capítulo se publicaron en la International Conference on Theory and Practice of Digital Libraries [173] y en la revista IEEE Access [174].

Si bien la metodología proporciona ideas valiosas, no está exenta de limitaciones. La dependencia de datos de acceso público introduce restricciones, particularmente a medida que el acceso a datos de redes sociales se vuelve más limitado. La dimensión temporal de los resultados de este capítulo es un factor crítico en su interpretación. Los datos recopilados entre 2017 y 2021 constituyen la base de numerosos análisis, incluida la evaluación de mecanismos de retroalimentación y el estudio a gran escala de los portales de DAG europeos. Desde entonces, las prácticas de datos abiertos, las tecnologías de portales y los marcos normativos han evolucionado significativamente, lo que podría afectar a la aplicabilidad de los hallazgos de este estudio en la actualidad. Por ejemplo, la introducción de nuevas directivas europeas sobre gobernanza de datos, las mejoras en el software de portales y la creciente utilización de inteligencia artificial para la curación de metadatos pueden haber dado lugar a modificaciones que no fueron consideradas en esta investigación. Además, los experimentos se realizaron con datos de acceso público, y la fiabilidad de los resultados depende de la calidad de las fuentes de datos. Pero en algunos casos, los valores ausentes, los campos de metadatos inconsistentes o la documentación limitada plantearon desafíos para el análisis. Además, nuestro método de clustering se basa en una serie de características diferentes. No obstante, puede haber variables más importantes que influyan en el nivel de participación de los usuarios y en la conformidad de los datos.

Nuestro estudio de investigación tiene implicaciones tanto teóricas como prácticas para el campo de los DAG. En el plano teórico, avanza en la comprensión de tres áreas: (i) cómo las anotaciones temáticas estructuradas pueden mejorar la capacidad de descubrimiento de conjuntos de datos y enriquecer la calidad de los metadatos; (ii) cómo los mecanismos de retroalimentación en los portales de DAG funcionan como motores de transparencia, responsabilidad y participación pública; y (iii)

cómo los portales de DAG evolucionan a lo largo del tiempo en términos de inclusividad e interacción con los usuarios. Estos conocimientos amplían el trabajo previo sobre ecosistemas de datos abiertos al ofrecer nuevas herramientas metodológicas (p. ej., anotación automatizada) y marcos conceptuales (p. ej., retroalimentación circular).

En el plano práctico, los resultados sugieren aplicaciones concretas para responsables políticos, gestores de portales y desarrolladores. Por ejemplo, el marco de anotación temática puede adaptarse como un complemento para plataformas de datos abiertos ampliamente utilizadas como CKAN, DKAN o Socrata, siempre que los temas para la clasificación se definan en alineación con las necesidades del portal. Los resultados sobre mecanismos de retroalimentación también tienen implicaciones directas para la gobernanza de iniciativas de DAG: los portales con características similares pueden compararse con otros y ajustar sus políticas de implicación de usuarios en consecuencia. Además, las directrices derivadas de los canales de retroalimentación de entrada y salida estudiados en esta investigación pueden ayudar a los equipos administrativos a identificar canales de comunicación infrautilizados e iniciar nuevos flujos de interacción con los usuarios finales. De este modo, el estudio aporta tanto claridad conceptual como estrategias prácticas para fomentar portales de DAG más inclusivos y sensibles a los usuarios.

Esta investigación doctoral se llevó a cabo en el marco del proyecto “ODECO: Towards a Sustainable Open Data Ecosystem”. ODECO fue una Red de Formación Innovadora de cuatro años financiada bajo el programa Horizon 2020 Marie Skłodowska-Curie (H2020-MSCA-ITN-2020, Grant Agreement No. 955569). El objetivo general del proyecto fue abordar los desafíos actuales y emergentes en la construcción de ecosistemas de datos abiertos inclusivos, centrados en el usuario y basados en principios circulares. Además de las contribuciones originales presentadas en esta disertación, el proyecto también permitió una participación activa en actividades académicas y de divulgación colaborativas, centrándose especialmente en la infraestructura de datos abiertos y en la mejora de la experiencia de usuario. Asimismo, con la colaboración de colegas de ODECO hemos publicado un artículo en las actas de la 25th Annual International Conference on Digital Government Research [175] y un artículo en la revista Information Polity [176].

7.2 Cuestiones abiertas

Si bien esta tesis ha presentado varios marcos para mejorar la clasificación temática, los mecanismos de retroalimentación y el análisis de la participación en los ecosistemas de Datos Abiertos Gubernamentales (DAG), aún quedan múltiples vías para futuras investigaciones. La naturaleza dinámica y socio-técnica de las plataformas de datos abiertos requiere una adaptación continua a las tecnologías emergentes, las expectativas cambiantes de los usuarios y los marcos políticos en evolución. Sobre la base del trabajo fundacional presentado en esta tesis, las líneas futuras pueden mejorar aún más la sostenibilidad, la capacidad de respuesta y la inclusividad de los ecosistemas de DAG abordando tanto las dimensiones técnicas como las centradas en las personas.

El Capítulo 3 tiene como objetivo el desarrollo de un marco robusto para la anotación temática

automática de DAG. Como trabajo futuro en este capítulo, se sugirió que sería conveniente explorar si la información relacionada con el esquema de aplicación de las diferentes distribuciones de conjuntos de datos puede ayudarnos a mejorar la clasificación temática automática de los conjuntos de datos. Las distribuciones disponibles en formatos legibles por máquinas, como CSV o RDF, pueden proporcionar en algunos casos nombres significativos de atributos temáticos del contenido del conjunto de datos. Incluso en el caso de RDF (datos en grafo), estos atributos suelen seleccionarse de vocabularios bien conocidos, lo que puede utilizarse para inferir vínculos con los temas que pueden asignarse automáticamente.

El Capítulo 4 explora el papel de los mecanismos de retroalimentación en los portales de DAG. La investigación futura podría centrarse en desarrollar métricas fiables para evaluar el éxito y el impacto de los mecanismos de retroalimentación en los portales de DAG. Indicadores como las tasas de respuesta de los publicadores de datos, la aparición de debates en foros dedicados y los niveles de satisfacción de los usuarios tras enviar retroalimentación pueden ofrecer información valiosa sobre la eficacia del sistema. Además, la elaboración de perfiles de usuarios es una vía potencial, especialmente en términos de comprensión de cómo los distintos grupos de actores (como estudiantes, periodistas, organizaciones no gubernamentales, empresas comerciales e intermediarios) interactúan con los canales de retroalimentación. Esta comprensión podría conducir al desarrollo de estrategias más adecuadas y eficientes. Por último, en lo que respecta a ampliar la adopción y eficacia de los sistemas de retroalimentación, actividades complementarias como iniciativas de co-creación, acciones de divulgación específicas y campañas de construcción de confianza también pueden desempeñar un papel importante.

El Capítulo 5 introduce un marco para evaluar la evolución y la participación de los usuarios en las iniciativas de DAG. Una dirección futura prometedora consiste en realizar un análisis de contenido más detallado de las interacciones de los usuarios en plataformas como X (anteriormente Twitter) utilizando técnicas de análisis de sentimiento y semántico. Esto podría revelar conocimientos más profundos sobre la participación de los usuarios, incluidas preocupaciones recurrentes, ejemplos de reutilización y sugerencias de mejora. Aunque los resultados iniciales sugieren una influencia limitada de las características de sentimiento en la construcción de clústeres, la categorización del contenido de los tuits aún podría ayudar a descubrir patrones más matizados de interacción. Ampliar este marco para incluir portales de DAG fuera de Europa y analizar períodos de tiempo más largos permitiría comparaciones más completas entre regiones. Dicho trabajo podría informar a los responsables políticos y a los profesionales sobre dinámicas regionales y mejores prácticas, contribuyendo al desarrollo de ecosistemas de datos abiertos más receptivos y centrados en el usuario.

Por último, es necesario reconocer que los estudios de caso utilizados para la validación de las contribuciones en esta tesis se han centrado en analizar las características y el contenido de metadatos del portal europeo de datos (data.europa.eu) y de los portales nacionales de los Estados miembros de la Unión Europea que este agrega. Sin embargo, las propuestas y experimentos con estos estudios de caso también podrían aplicarse a portales de DAG en otros países, con el trabajo necesario de personalización a las características de cada portal. Además, la posibilidad de adaptar las

contribuciones a portales de datos no gubernamentales también podría considerarse como una futura línea de investigación.

REFERENCES

[1] Susana de Juana-Espinosa and Sergio Luján-Mora. Open government data portals in the european union: A dataset from 2015 to 2017. Data in brief, 29:105156, 2020.

[2] Luigi Reggi and Sharon S Dawes. Creating open government data ecosystems: Network relations among governments, user communities, NGOs and the media. Government Information Quarterly, 39(2):101675, 2022.

[3] Publications Office of the European Union. The economic impact of open data: Opportunities for value creation in europe, 2022. https://data.europa.eu/sites/default/files/the-economic-impact-of-open-data.pdf.

[4] European Commission. Digital agenda: Commission's open data strategy, questions & answers, 2011. https://ec.europa.eu/commission/presscorner/api/files/document/print/en/memo_11_891/MEMO_11_891_EN.pdf.

[5] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4):258–268, 2012.

[6] Angela M Evans and Adriana Campos. Open government initiatives: Challenges of citizen participation. Journal of policy analysis and management, pages 172–185, 2013.

[7] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. Journal of Data and Information Quality (JDIQ), 8(1):1–29, 2016.

[8] W. Carrara, W.-S. Chan, S. Fischer, and E. Van-Steenbergen. Creating value through open data: Study on the impact of re-use of public data resources, 2015. https://data.europa.eu/sites/default/files/edp_creating_value_through_open_data_0.pdf.

[9] Abiola Paterne Chokki, Anthony Simonofski, Benoît Frénay, and Benoit Vanderose. Engaging citizens with open government data: The value of dashboards compared to individual visualizations. Digital Government: Research and Practice, 3(3):1–20, 2022.

[10] Nigel Bevan, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. New iso standards for usability, usability reports and usability measures. In Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I 18, pages 268–278. Springer, 2016.

[11] Juho-Pekka Mäkipää. Explaining accessibility: Possible variables in users' abilities, tasks, and contexts in it artefact use. AIS Transactions on Human-Computer Interaction, 15(4):414–441, 2023.

[12] W3C. Web content accessibility guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/, 2024. Accessed: 2025-02-15.

[13] Renáta Máchová, Miloslav Hub, and Martin Lnenicka. Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. Aslib Journal of Information Management, 70(3):252–268, 2018.

[14] Katrin Braunschweig, Julian Eberius, Maik Thiele, and Wolfgang Lehner. The state of open data: Limits of current open data platforms categories and subject descriptors. In Conference Proceedings World Wide Web Conference, WWW, 2012.

[15] Olayiwola Bello, Victor Akinwande, Oluwatoyosi Jolayemi, and Ahmed Ibrahim. Open data portals in africa: An analysis of open government data initiatives. African Journal of Library, Archives and Information Science, 26(2):97–106, 2016.

[16] Publications Office of the European Union. Rethinking the impact of open data – A first step towards a European impact assessment for open data. Publications Office, 2023. https://op.europa.eu/en/publication-detail/-/publication/62cee051-a11a-11ed-b508-01aa75ed71a1/.

[17] Renee E Sieber and Peter A Johnson. Civic open data at a crossroads: Dominant models and current challenges. Government information quarterly, 32(3):308–315, 2015.

[18] Barbara Ubaldi. Open government data: Towards empirical analysis of open government data initiatives. Technical Report 22, OECD Publishing, Paris, 2013. DOI: https://doi.org/10.1787/5k46bj4f03s7-en.

[19] Daniel Lathrop and Laurel Ruma. Open government: Collaboration, transparency, and participation in practice. " O'Reilly Media, Inc.", 2010.

[20] Teresa M Harrison and Djoko Sigit Sayogo. Transparency, participation, and accountability practices in open government: A comparative study. Government information quarterly, 31(4):513–525, 2014.

[21] Luigi Reggi, Sharon S Dawes, and J Ramon Gil-Garcia. The effects of open government data on the inclusiveness of governance networks: Identifying management strategies and success factors. Information Polity, 27(4):473–490, 2022.

[22] Paul Plantinga and Rachel Adams. Rethinking open government as innovation for inclusive development: Open access, data and ict in south africa. African Journal of Science, Technology, Innovation and Development, 13(3):315–323, 2021.

[23] Theodora Kouvara, Rozita Tsoni, and Vassilios S Verykios. Openness is not enough without inclusiveness: The learning analytics parable. In Digital Reset: European Universities Transforming for a Changing World: Overview of papers as presented during the Innovating Higher Education Conference 2022: 19-21 October 2022 in Athens, pages 271–285, 2022.

[24] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. A systematic review of open government data initiatives. Government information quarterly, 32(4):399–418, 2015.

[25] Adegboyega Ojo, Lukasz Porwol, Mohammad Waqar, Arkadiusz Stasiewicz, Edobor Osagie, Michael Hogan, Owen Harney, and Fatemeh Ahmadi Zeleti. Realizing the innovation potentials from open data: Stakeholders' perspectives on the desired affordances of open data environment. In Working Conference on Virtual Enterprises, pages 48–59. Springer, 2016.

[26] Ahmad Luthfi and Marijn Janssen. Toward a reference architecture for user-oriented open government data portals. In International Symposium on Business Modeling and Software Design, pages 259–267. Springer, 2022.

[27] Javier Nogueras-Iso, Miguel Ángel Latre, Ruben Bejar, Pedro R Muro-Medrano, and F Javier Zarazaga-Soria. A model driven approach for the development of metadata editors, applicability to the annotation of geographic information resources. Data & Knowledge Engineering, 81:118–139, 2012.

[28] Abiola Paterne Chokki, Charalampos Alexopoulos, Stuti Saxena, Benoît Frénay, Benoît Vanderose, and Mohsan Ali. Metadata quality matters in open government data (ogd) evaluation! an empirical investigation of ogd portals of the gcc constituents. Transforming Government: People, Process and Policy, 17(3):303–316, 2022.

[29] Charalampos Alexopoulos, Euripidis Loukis, and Yannis Charalabidis. A platform for closing the open data feedback loop based on web 2. 0 functionality. JeDEM-eJournal of eDemocracy and Open government, 6(1):62–68, 2014.

[30] Iryna Susha, Åke Grönlund, and Marijn Janssen. Organizational measures to stimulate user engagement with open data. Transforming Government: People, Process and Policy, 9(2):181–206, May 2015.

[31] Arie Purwanto, Anneke Zuiderwijk-van Eijk, and Marijn Janssen. Citizen engagement with open government data: Lessons learned from indonesia's presidential election. Transforming Government: people, process and policy, 14(1):1–30, 2020.

[32] Marijn Janssen and Natalie Helbig. Innovating and changing the policy-cycle: Policy-makers be prepared! Government Information Quarterly, 35(4, Supplement):S99–S105, 2018.

[33] OECD. Open government data report: Enhancing policy maturity for sustainable impact, 2018.

[34] Andrew Booth, Diana Papaioannou, and Anthea Sutton. Systematic Approaches to a Successful Literature Review. SAGE Publications, Los Angeles; Thousand Oaks, CA, 2012.

[35] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. British Medical Journal, 372, 2021.

[36] Bernd W Wirtz, Jan C Weyerer, and Michael Rösch. Open government and citizen participation: an empirical analysis of citizen expectancy towards open government data. International Review of Administrative Sciences, 85(3):566–586, 2019.

[37] Charles W Bailey. What is open access. Open access: key strategic, technical and economic aspects. Oxford: Chandos Publishing, pages 13–26, 2006.

[38] Laura Fortunato and Mark Galassi. The case for free and open source software in research and scholarship. Philosophical Transactions of the Royal Society A, 379(2197):20200079, 2021.

[39] Luciano Floridi. The logic of information: A theory of philosophy as conceptual design. Oxford University Press, 2019.

[40] Jeanne G Harris. How to turn data into a strategic asset. Outlook Journal, Accenture, 2010.

[41] Mohammad Alamgir Hossain, Yogesh K Dwivedi, and Nripendra P Rana. State-of-the-art in open data research: Insights from existing literature and a research agenda. Journal of organizational computing and electronic commerce, 26(1-2):14–40, 2016.

[42] Zhe Zhu, Michael A Wulder, David P Roy, Curtis E Woodcock, Matthew C Hansen, Volker C Radeloff, Sean P Healey, Crystal Schaaf, Patrick Hostert, Peter Strobl, et al. Benefits of the free and open landsat data policy. Remote Sensing of Environment, 224:382–385, 2019.

[43] Grace M Begany and J Ramon Gil-Garcia. Understanding the actual use of open data: Levels of engagement and how they are related. Telematics and Informatics, 63:101673, 2021.

[44] Thorhildur Jetzek. The value generating mechanisms of open government data. Geoforum Perspektiv, 12(23), 2013.

[45] Bonnie A Nardi and Vicki O'Day. Information ecologies: Using technology with heart. Mit Press, 2000.

[46] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. Innovation with open data: Essential elements of open data ecosystems. Information polity, 19(1-2):17–33, 2014.

[47] Bastiaan Van Loenen, Anneke Zuiderwijk, Glenn Vancauwenberghe, Francisco J Lopez-Pellicer, Ingrid Mulder, Charalampos Alexopoulos, Rikke Magnussen, Mubashrah Saddiqa, Melanie Dulong de Rosnay, Joep Crompvoets, et al. Towards value-creating and sustainable open data ecosystems: A comparative case study and a research agenda. JeDEM-eJournal of eDemocracy and Open Government, 13(2):1–27, 2021.

[48] Sharon S Dawes, Lyudmila Vidiasova, and Olga Parkhimovich. Planning and designing open government data programs: An ecosystem approach. Government Information Quarterly, 33(1):15–27, 2016.

[49] Teresa M Harrison, Theresa A Pardo, and Meghan Cook. Creating open government ecosystems: A research and development agenda. Future Internet, 4(4):900–928, 2012.

[50] Thorhildur Jetzek. Innovation in the open data ecosystem: Exploring the role of real options thinking and multi-sided platforms for sustainable value generation through open data. Analytics, Innovation, and Excellence-Driven Enterprise Sustainability, pages 137–168, 2017.

[51] Anneke M. G. Zuiderwijk. Open Data Infrastructures: The design of an infrastructure to enhance the coordination of open data use. Phd thesis, Delft University of Technology, 2015.

[52] Kawaljeet Kapoor, Vishanth Weerakkody, and Uthayasankar Sivarajah. Open data platforms and their usability: Proposing a framework for evaluating citizen intentions. In Open and Big Data Management and Innovation: 14th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2015, Delft, The Netherlands, October 13-15, 2015, Proceedings 14, pages 261–271. Springer, 2015.

[53] Ingrid Mulder, Tomasz Jaskiewicz, and Nicola Morelli. On digital citizenship and data as a new commons: Can we design a new movement? Cuadernos del Centro de Estudios en Diseño y Comunicación. Ensayos, (73):96–108, 2019.

[54] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. Government information quarterly, 31(1):17–29, 2014.

[55] OECD. Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact. OECD Digital Government Studies. OECD Publishing, Paris, 2018. DOI: https://doi.org/10.1787/9789264305847-en.

[56] Tim Davies, Stephen B Walker, Mor Rubinstein, and Fernando Perini. The state of open data: Histories and horizons. African Minds, 2019.

[57] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9, 2016.

[58] Rob Kitchin. The data revolution: Big data, open data, data infrastructures and their consequences. Sage, 2014.

[59] Luis Felipe Luna-Reyes, John C Bertot, and Sehl Mellouli. Open government, open data and digital government. Government Information Quarterly, 31(1):4–5, 2014.

[60] Sylvain Kubler, Jerermy Robert, Sebastian Neumaier, Jürgen Umbrich, and Yves Le Traon. Comparison of metadata quality in open data portals using the analytic hierarchy process. Government Information Quarterly, 35(1):13–29, 2018.

[61] Jenn Riley. Understanding Metadata: What Is Metadata, and What Is It For? NISO Primer. National Information Standards Organization, Baltimore, MD, 2017.

[62] Anthony Simonofski, Anneke Zuiderwijk, Antoine Clarinval, and Wafa Hammedi. Tailoring open government data portals for lay citizens: A gamification theory approach. International Journal of Information Management, 65:102511, 2022.

[63] Lorea Akerreta Escribano and Julián Moyano Collado. Contar historias con los datos: Aragón Open Data Focus, una experiencia innovadora de reutilización de los datos del sector público. Scire: representación y organización del conocimiento, 27(1):31–43, 2021.

[64] Joss Winn. Open data and the academy: An evaluation of CKAN for research data management. 2013. https://rd-alliance.org/system/files/documents/CKANEvaluation.pdf.

[65] Renata Máchová and Martin Lněnička. Evaluating the quality of open data portals on the national level. Journal of theoretical and applied electronic commerce research, 12(1):21–41, 2017.

[66] Ahmad Assaf, Raphaël Troncy, and Aline Senart. HDL-Towards a harmonized dataset model for open data portals. In Usewod-profiles@ ESWC, pages 62–74, 2015.

[67] Gennaro Cordasco, Delfina Malandrino, Donato Pirozzi, Vittorio Scarano, and Carmine Spagnuolo. A layered architecture for open data: Design, implementation and experiences. In Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, pages 371–381, 2018.

[68] Arshad Ali Khan. Exploiting Linked Open Data (LoD) and Crowdsourcing-based Semantic Annotation & Tagging in Web Repositories to Improve and Sustain Relevance in Search Results. Phd thesis, University of Southampton, 2018. https://eprints.soton.ac.uk/428046/.

[69] Tim G Davies and Zainab Ashraf Bawa. The promises and perils of open government data (ogd). The Journal of Community Informatics, 8(2), 2012.

[70] Syed Iftikhar Hussain Shah, Vassilios Peristeras, and Ioannis Magnisalis. A conceptual framework for the government big data ecosystem ('datagov.eco'). Data Knowledge Engineering, 154:102348, 2024.

[71] Charalampos Alexopoulos, Lefkothea Spiliotopoulou, and Yannis Charalabidis. Open data movement in greece: a case study on open government data sources. In Proceedings of the 17th Panhellenic Conference on Informatics, pages 279–286, 2013.

[72] Martin Lnenicka, Anastasija Nikiforova, Mariusz Luterek, Petar Milic, Daniel Rudmark, Sebastian Neumaier, Karlo Kević, Anneke Zuiderwijk, and Manuel Pedro Rodríguez Bolívar. Understanding the development of public data ecosystems: From a conceptual model to a six-generation model of the evolution of public data ecosystems. Telematics and informatics, page 102190, 2024.

[73] Martin Lněnička, Renata Machova, Jolana Volejníková, Veronika Linhartová, Radka Knezackova, and Miloslav Hub. Enhancing transparency through open government data: The case of data portals and their features and capabilities. Online Information Review, 45(6):1021–1038, 2021.

[74] Mona Mohamed, Sharma Pillutla, and Stella Tomasi. Extraction of knowledge from open government data: The knowledge iterative value network framework. VINE Journal of Information and Knowledge Management Systems, 50(3):495–511, 2020.

[75] Peter Morville and Louis Rosenfeld. Information Architecture for the World Wide Web. O'Reilly Media, Inc., Canada, 2015.

[76] Javier Nogueras-Iso, Javier Lacasta, Manuel Antonio Ureña-Cámara, and Francisco Javier Ariza-López. Quality of metadata in open data portals. IEEE Access, 9:60364–60382, 2021.

[77] Bianca Wentzel, Fabian Kirstein, Torben Jastrow, Raphael Sturm, Michael Peters, and Sonja Schimmler. An extensive methodology and framework for quality assessment of DCAT-AP datasets. In International Conference on Electronic Government, pages 262–278. Springer, 2023.

[78] Fabian Kirstein, Benjamin Dittwald, Simon Dutkowski, Yury Glikman, Sonja Schimmler, and Manfred Hauswirth. Linked data in the european data portal: A comprehensive platform for applying DCAT-AP. In International Conference on Electronic Government, pages 192–204. Springer, 2019.

[79] Makx Dekkers, Stefanos Kotoglou, Chris Nelson, Marco Pellegrino, Norbert Hohn, and Vassilios Peristeras. StatDCAT-AP, a common layer for the exchange of statistical metadata in open data portals. In 6th International Workshop on Semantic Statistics co-located with the 17th International Semantic Web Conference, 2016.

[80] W3C. Data Catalog Vocabulary (DCAT). W3C Recommendation 16 January 2014, 2014. https://www.w3.org/TR/vocab-dcat/.

[81] European Commission. DCAT Application profile for data portals in Europe, DCAT-AP Version 2.1.0, 2021. https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe/release/210.

[82] Luigi Arlotta, Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. Automatic annotation of data extracted from large web sites. In International Workshop on the Web and Databases, pages 7–12, 2003.

[83] Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. Automatic annotation suggestions and custom annotation layers in webanno. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 91–96, 2014.

[84] Matthew Petrillo and Jessica Baycroft. Introduction to manual annotation. Fairview research, pages 1–7, 2010.

[85] Sebastian Haunss, Jonas Kuhn, Sebastian Padó, Andre Blessing, Nico Blokker, Erenay Dayanik, and Gabriella Lapesa. Integrating manual and automatic annotation for the creation of discourse network data sets. Politics and Governance, 8(2):326–339, 2020.

[86] Alaa Qasim Mohammed Salih. Towards from manual to automatic semantic annotation: based on ontology elements and relationships. International Journal of Web & Semantic Technology, 4(2):21, 2013.

[87] Publications Office of the European Union. Metadata quality assessment methodology. Online, 2025. https://data.europa.eu/api/mqa/reporter/report/en/pdf. Accessed: June 6, 2025.

[88] Anthony Simonofski, Anastasija Nikiforova, Martin Lnenicka, and Nicolas Bono Rossello. Artificial intelligence as a catalyzer for open government data ecosystems: A typological theory approach. Proceedings of the 58th Hawaii International Conference on System Sciences, 2025. https://hdl.handle.net/10125/109107.

[89] Umair Ahmed. Reimagining open data ecosystems: a practical approach using AI, CI, and knowledge graphs. In BIR Workshops, pages 235–249, 2023.

[90] Robert Enríquez-Reyes, Susana Cadena-Vela, Andrés Fuster-Guilló, Jose-Norberto Mazón, Luis Daniel Ibáñez, and Elena Simperl. Systematic mapping of open data studies: Classification and trends from a technological perspective. IEEE Access, 9:12968–12988, 2021.

[91] Pierre-Yves Vandenbussche and Bernard Vatant. Metadata recommendations for linked open data vocabularies. 1:2011–12, 2011. https://lov.linkeddata.es/Recommendations_Vocabulary_Design.pdf.

[92] Sophie Pavia, Nickolas Piraino, Kazi Islam, Anna Pyayt, and Michael N Gubanov. Hybrid metadata classification in large-scale structured datasets. Journal of Data Intelligence, 3(4):460–473, 2022.

[93] Giulio Carducci, Marco Leontino, Daniele P Radicioni, Guido Bonino, Enrico Pasini, and Paolo Tripodi. Semantically aware text categorisation for metadata annotation. In Italian Research Conference on Digital Libraries, pages 315–330. Springer, 2019.

[94] Suzan Verberne, Eva D'hondt, Antal Van den Bosch, and Maarten Marx. Automatic thematic classification of election manifestos. Information Processing & Management, 50(4):554–567, 2014.

[95] Umair Ahmed, Charalampos Alexopoulos, Marco Piangerelli, and Andrea Polini. BRYT: Automated keyword extraction for open datasets. Intelligent Systems with Applications, 23:200421, 2024.

[96] Kevin Kliimask and Anastasija Nikiforova. TAGIFY: LLM-powered tagging interface for improved data findability on OGD portals. In 2024 Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA), pages 18–27. IEEE, 2024.

[97] Ramil Huseynov, Anastasija Nikiforova, Dimitrios Symeonidis, and David Duenas-Cid. May the Data Be with You: Towards an AI-Powered Semantic Recommender for Unlocking Dark Data. In International Conference on Electronic Government. Springer, 2025.

[98] Julia Sasse, Johannes Darms, and Juliane Fluck. Semantic metadata annotation services in the biomedical domain—a literature review. Applied Sciences, 12(2):796, 2022.

[99] Alexandre Hudon, Mélissa Beaudoin, Kingsada Phraxayavong, Laura Dellazizzo, Stéphane Potvin, and Alexandre Dumais. Use of automated thematic annotations for small data sets in a psychotherapeutic context: systematic review of machine learning algorithms. JMIR Mental Health, 8(10):e22651, 2021.

[100] Alexandre Hudon, Mélissa Beaudoin, Kingsada Phraxayavong, Laura Dellazizzo, Stéphane Potvin, and Alexandre Dumais. Implementation of a machine learning algorithm for automated thematic annotations in avatar: A linear support vector classifier approach. Health Informatics Journal, 28(4):14604582221142442, 2022.

[101] Suppawong Tuarob, Line C Pouchard, Natasha F Noy, Jeffery S Horsburgh, and Giri Palanisamy. ONEMercury: Towards automatic annotation of environmental science metadata. In Proceedings of the Second International Workshop on Linked Science 2012 - Tackling Big Data, Boston, MA, USA,, volume 951 of CEUR Workshop Proceedings, 2012.

[102] Jeffrey S Ellen, Casey A Graff, and Mark D Ohman. Improving plankton image classification using context metadata. Limnology and Oceanography: Methods, 17(8):439–461, 2019.

[103] Yonghong Peng, Zhiqing Wu, and Jianmin Jiang. A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics, 43(1):15–23, 2010.

[104] Phivos Mylonas, Yorghos Voutos, and Anastasia Sofou. A collaborative pilot platform for data annotation and enrichment in viticulture. Information, 10(4):149, 2019.

[105] Mingfang Wu, Hans Brandhorst, Maria-Cristina Marinescu, Joaquim More Lopez, Margorie Hlava, and Joseph Busch. Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence, 5(1):122–138, 2023.

[106] Luis-Daniel Ibáñez, Ian Millard, Hugh Glaser, and Elena Simperl. An assessment of adoption and quality of linked data in european open government data. In International Semantic Web Conference, pages 436–453. Springer, 2019.

[107] Eirini Kaldeli, Orfeas Menis-Mastromichalakis, Spyros Bekiaris, Maria Ralli, Vassilis Tzouvaras, and Giorgos Stamou. CrowdHeritage: crowdsourcing for improving the quality of cultural heritage metadata. Information, 12(02):64, 2021.

[108] A. Miles and D. Brickley. SKOS Core Guide. W3C Working Draft 2 November 2005, 2005. https://www.w3.org/TR/swbp-skos-core-guide.

[109] Publications Office of the European Union. Data theme authority table, 2022. https://op.europa.eu/s/zBx4.

[110] Dayany Díaz-Corona, Javier Lacasta, Miguel Ángel Latre, F Javier Zarazaga-Soria, and Javier Nogueras-Iso. Profiling of knowledge organisation systems for the annotation of linked data cultural resources. Information Systems, 84:17–28, 2019.

[111] Patricia Martín-Chozas, Elena Montiel-Ponsoda, and Víctor Rodríguez-Doncel. Language resources as linked data for the legal domain. In Knowledge of the Law in the Big Data Age, pages 170–180. IOS Press, 2019.

[112] Wang Tao, Jiang Yongjia, and Rong Xiangsheng. A novel two-level one-vs-rest classifier. In 2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE), pages 645–648. IEEE, 2019.

[113] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of North American Association for Computational Linguistics, volume 1, page 2, 2019.

[114] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9119–9130, Online, November 2020. Association for Computational Linguistics.

[115] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. In Findings of the Association for Computational Linguistics: ACL 2022, pages 1864–1874, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[116] International Organization for Standardization (ISO). Sampling Procedures for Inspection by Attributes–Part 2: Sampling Plans Indexed by Limiting Quality (LQ) for Isolated Lot Inspection, Standard ISO 2859-2:1985, 1985.

[117] Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. A/B testing: A systematic literature review. Journal of Systems and Software, 211:112011, 2024.

[118] E. Simperl and 2020. Walker, J. The future of open data portals. Publications Office of the European Union., 2020.

[119] Martin Lnenicka and Anastasija Nikiforova. Transparency-by-design: What is the role of open data portals? Telematics and Informatics, 61:101605, 2021.

[120] Thorhildur Jetzek, Michel Avital, and Niels Bjørn-Andersen. Generating value from open government data. In Proceedings of the 34th International Conference on Information Systems. ICIS 2013. Association for Information Systems. AIS Electronic Library (AISeL), 2013.

[121] Fredrick Ishengoma and Deo Shao. A framework for aligning e-government initiatives with the sustainable development goals. Journal of Innovative Digital Transformation, 2(1):73–89, 2025.

[122] Abiola Paterne Chokki and Benoit Vanderose. From conventional open government data portals to storytelling portals: The storyogd prototype. In Proceedings of the 24th Annual International Conference on Digital Government Research, dg.o '23, page 642–643. Association for Computing Machinery, 2023.

[123] Mila Gascó-Hernández, Erika G Martin, Luigi Reggi, Sunyoung Pyo, and Luis F Luna-Reyes. Promoting the use of open government data: Cases of training and engagement. Government Information Quarterly, 35(2):233–242, 2018.

[124] Unisse C Chua, Kyle L Santiago, Ian Benedict M Ona, Romeo Manuel N Peña, Geremiah Zachary S Marasigan, Paolo Gabriel A Delos Reyes, and Briane Paul V Samson. From access to effective use: Open data portals for everyday citizens. In Proceedings of the 2020 Symposium on Emerging Research from Asia and on Asian Contexts and Cultures, pages 61–64, 2020.

[125] European Commission. Enable feedback channels for improving the quality of existing government data, 2016. https://data.europa.eu/sites/default/files/report/enable-feedback-channels-for-improving-the-quality-of-existing-government-data.pdf.

[126] Bev Wilson and Cong Cong. Beyond the supply side: Use and impact of municipal open data in the us. Telematics and Informatics, 58:101526, 2021.

[127] Erna HJM Ruijer and Evelijn Martinius. Researching the democratic impact of open government data: A systematic literature review. Information Polity, 22(4):233–250, 2017.

[128] Peter A Johnson. Reflecting on the success of open data: How municipal government evaluates their open data programs. International Journal of E-Planning Research (IJEPR), 5(3):1–12, 2016.

[129] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. Journal of management information systems, 24(3):45–77, 2007.

[130] Robert K Yin. Case study research: Design and methods, volume 5. sage, 2009.

[131] Joseph Kekeya. Qualitative case study research design: The commonalities and differences between collective, intrinsic and instrumental case studies. Contemporary PNG Studies, 36:28–37, 2021.

[132] Kathleen M Eisenhardt. Building theories from case study research. Academy of management review, 14(4):532–550, 1989.

[133] Sharon S. Dawes and Natalie Helbig. Information strategies for open government: Challenges and prospects for deriving public value from government transparency. In Maria A. Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans J. Scholl, editors, Electronic Government, pages 50–60, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[134] Rodrigo Hickmann Klein, Deisy Barbiero Klein, and Edimara M Luciano. Identification of mechanisms for the increase of transparency in open data portals: an analysis in the brazilian context. Cadernos EBAPE.BR, 16:692, 11 2018.

[135] Martin Lnenicka and Renata Machova. How intentions to use of OGD and respective portals changed over time: evidence from a three-year study among Czech students. Online Information Review, 49(3):517–533, May 2025.

[136] Panom Gunawong. Open government and social media: A focus on transparency. Social Science Computer Review, 33(5):587–598, 2015.

[137] Publications Office of the European Union. Open Data Maturity Report 2021. Publications Office, Luxembourg, 2022. https://data.europa.eu/doi/10.2830/394148.

[138] Nushrat Khan, Mike Thelwall, and Kayvan Kousha. Measuring the impact of biodiversity datasets: data reuse, citations and altmetrics. Scientometrics, 126(4):3621–3639, April 2021.

[139] Publications Office of the European Union. Insights into the user experience of the European Data Portal. https://data.europa.eu/en/publications/datastories/insights-user-experience-european-data-portal, 2020. Accessed: 2025-06-11.

[140] Han Zhang, Ying Bi, Fei Kang, and Zhong Wang. Incentive mechanisms for government officials' implementing open government data in China. Online Information Review, 46(2):224–243, January 2022. Publisher: Emerald Publishing Limited.

[141] Anneke Zuiderwijk, Marijn Janssen, and Iryna Susha and. Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators. Journal of Organizational Computing and Electronic Commerce, 26(1-2):116–146, 2016.

[142] Renáta Máchová, Miloslav Hub, and Martin Lnenicka. Usability evaluation of open data portals: Evaluating data discoverability, accessibility, and reusability from a stakeholders' perspective. Aslib Journal of Information Management, 70(3):252–268, May 2018.

[143] Anastasija Nikiforova. Comparative analysis of national open data portals or whether your portal is ready to bring benefits from open data. In Piet Kommers and Guo Chao Peng, editors, Proceedings of the International Conference on e-Society 2020, pages 81–88. IADIS Press, 2020.

[144] John Venable and R. Baskerville. Eating our own cooking: Toward a more rigorous design science of research methods. Electronic Journal of Business Research Methods, 10:141–153, 01 2012.

[145] Nataša Veljković, Sanja Bogdanović-Dinić, and Leonid Stoimenov. Benchmarking open government: An open data perspective. Government information quarterly, 31(2):278–290, 2014.

[146] Gema Santos-Hermosa, Alfonso Quarati, Eugenia Loría-Soriano, and Juliana E Raffaghelli. Why does open data get underused? a focus on the role of (open) data literacy. In Data Cultures in Higher Education: Emergent Practices and the Challenge Ahead, pages 145–177. Springer, 2023.

[147] Stefanie Haustein, Rodrigo Costas, and Vincent Larivière. Characterizing social media metrics of scholarly papers: The effect of document properties and collaboration patterns. PloS one, 10(3):e0120495, 2015.

[148] Rui Pedro Lourenço. Open government portals assessment: a transparency for accountability perspective. In Electronic Government: 12th IFIP WG 8.5 International Conference, EGOV 2013, Koblenz, Germany, September 16-19, 2013. Proceedings 12, pages 62–74. Springer, 2013.

[149] Abdul Aziz. Technical aspects for inclusiveness across user domains in data portals. In CEUR Workshop Proceedings, volume 3514, pages 271–280, 2023. DOI: https://doi.org/10.5281/zenodo.10039867.

[150] Mohsan Ali, Charalampos Alexopoulos, and Yannis Charalabidis. A comprehensive review of open data platforms, prevalent technologies, and functionalities. In Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance, ICEGOV '22, page 203–214, New York, NY, USA, 2022. Association for Computing Machinery.

[151] Andreiwid Sheffer Correa, Pär-Ola Zander, and Flavio Soares Correa da Silva. Investigating open data portals automatically: a methodology and some illustrations. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, dg.o '18, New York, NY, USA, 2018. Association for Computing Machinery.

[152] Anastasija Nikiforova and Keegan McBride. Open government data portal usability: A user-centred usability analysis of 41 open government data portals. Telematics and Informatics, 58:101539, 2021.

[153] Xiaohua Zhu and Mark Antony Freeman. An evaluation of US municipal open data portals: A user interaction framework. Journal of the Association for Information Science and Technology, 70(1):27–37, 2019.

[154] Hamidreza Shahbaznezhad, Rebecca Dolan, and Mona Rashidirad. The role of social media content format and platform in users' engagement behavior. Journal of Interactive Marketing, 53(1):47–65, 2021.

[155] José Luis Alonso Berrocal, Carlos G Figuerola, and Ángel F Zazo Rodriguez. Propuesta de índice de influencia de contenidos (Influ@ RT) en Twitter. Scire: representación y organización del conocimiento, pages 21–26, 2015.

[156] Min Zhang, Dongxin Zhang, Yin Zhang, Kristin Yeager, and Taylor N Fields. An exploratory study of twitter metrics for measuring user influence. Journal of Informetrics, 17(4):101454, 2023.

[157] Fereshteh Didegah, Niels Mejlgaard, and Mads P Sørensen. Investigating the quality of interactions and public engagement around scientific papers on twitter. Journal of informetrics, 12(3):960–971, 2018.

[158] Jianhua Hou, Yuanyuan Wang, Yang Zhang, and Dongyi Wang. How do scholars and non-scholars participate in dataset dissemination on twitter. Journal of Informetrics, 16(1):101223, 2022.

[159] Publications Office of the European Union. European Data Portal SPARQL Endpoint. https://data.europa.eu/sparql, Last accessed: 2022-05-30.

[160] X API v2. Twitter API v2. https://developer.twitter.com/en/docs/api-reference-index#twitter-api-v2, Last accessed: 2022-05-30.

[161] Joseph Hair, Barry Balbin, William Black, and Rolph Anderson. Multivariate Data Analysis. Cengage Learning EMEA, 2019.

[162] Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. Explainable k-means and k-medians clustering. In International conference on machine learning, pages 7055–7065. PMLR, 2020.

[163] Margarita Argüelles, María del Carmen Benavides, and I Fernández. A new approach to the identification of regional clusters: hierarchical clustering on principal components. Applied Economics, 46(21):2511–2519, 2014.

[164] Susana de Juana-Espinosa and Sergio Luján-Mora. Open government data portals in the european union: Considerations, development, and expectations. Technological Forecasting and Social Change, 149:119769, 2019.

[165] André Skupin and Pragya Agarwal. Self-Organising Maps: Applications in Geographic Information Science, chapter Introduction: What is a self-organizing map?, pages 1–20. Wiley Online Library, 2008.

[166] Ana Ruiz-Varona, Javier Lacasta, and Javier Nogueras-Iso. Self-organizing maps to evaluate multidimensional trajectories of shrinkage in spain. ISPRS International Journal of Geo-Information, 11(2):77, 2022.

[167] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1):86–97, 2012.

[168] Vladimir Batagelj. Classification and related methods of data analysis, chapter Generalized Ward and related clustering problems, pages 67–74. North-Holland Amsterdam, the Netherlands, 1988.

[169] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1):20–30, 2021.

[170] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. IEEE Transactions on neural networks, 11(3):586–600, 2000.

[171] Abdul Aziz, Mohsan Ali, Dagoberto Jose Herrera-Murillo, Maria Ioanna Maratsi, Francisco J. Lopez-Pellicer, and Javier Nogueras-Iso. A framework for the thematic annotation of open government data. ACM Journal of Data and Information Quality, 2025. Paper under review.

[172] Abdul Aziz, Mohsan Ali, Javier Nogueras-Iso, Charalampos Alexopoulos, and Francisco J Lopez-Pellicer. Roundtrip feedback in open data portals: analysis of input and output channels. Online Information Review, 49(8):134–151, 2025. https://doi.org/10.1108/OIR-12-2024-0807.

[173] Dagoberto Jose Herrera-Murillo, Abdul Aziz, Javier Nogueras-Iso, and Francisco J Lopez-Pellicer. Analysing user involvement in open government data initiatives. In Linking Theory and Practice of Digital Libraries. TPDL 2022, volume 13541 of Lecture Notes in Computer Science, pages 175–186. Springer, 2022. DOI: https://doi.org/10.1007/978-3-031-16802-4_14.

[174] Abdul Aziz, Dagoberto José Herrera-Murillo, Javier Nogueras-Iso, Javier Lacasta, and Francisco J. Lopez-Pellicer. Identifying the evolution of open government data initiatives and their user engagement. IEEE Access, 12:84556–84566, 2024. DOI: 10.1109/ACCESS.2024.3414282.

[175] Mohsan Ali, Georgios Papageorgiou, Abdul Aziz, Euripidis Loukis, Yannis Charalabidis, and Francisco Javier Lopez Pellicer. Towards the development of interoperable open data ecosystems: Harnessing the technical, semantic, legal, and organizational (TSLO) interoperability framework. In Proceedings of the 25th Annual International Conference on Digital Government Research, dg.o '24, pages 909–919, New York, NY, USA, 2024. Association for Computing Machinery. DOI: https://doi.org/10.1145/3657054.3657160.

[176] Mohsan Ali, Georgios Papageorgiou, Abdul Aziz, Euripidis Loukis, Yannis Charalabidis, Charalampos Alexopoulos, and Francisco Javier Lopez-Pellicer. A framework for the multi-dimensional assessment of interoperability for open data ecosystems development. Information Polity, 29(4):439–466, 2024. DOI: https://doi.org/10.1177/15701255241297172.