TESIS DE LA UNIVERSIDAD DE ZARAGOZA
2026
41

Dagoberto José Herrera Murillo

Evaluation of User Interfaces in Open Data Portals

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

EVALUATION OF USER INTERFACES IN OPEN DATA PORTALS

Autor

Dagoberto José Herrera Murillo

Director/es

Nogueras Iso, Francisco Javier
López Pellicer, Francisco Javier

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

2025

Universidad Zaragoza
1474

Tesis Doctoral

Evaluation of User Interfaces in Open Data Portals

Autor

Dagoberto José Herrera-Murillo

Directores

Dr. Javier Nogueras-Iso
Dr. Francisco J. López-Pellicer

Programa de Doctorado en Ingeniería de Sistemas e Informática
Escuela de Doctorado
2025

Evaluation of User Interfaces in Open Data Portals

Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

Dagoberto José Herrera-Murillo

Supervisors: Dr. Javier Nogueras-Iso
Dr. Francisco J. Lopez-Pellicer

Computer Science and Systems Engineering Department
Universidad de Zaragoza

A dissertation submitted to the Doctorate Program in Systems Engineering and Computer Science at Universidad de Zaragoza in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Computer Science and Systems Engineering
Department

September 2025

Evaluation of User Interfaces in Open Data Portals

Dagoberto José Herrera-Murillo

Abstract

This thesis investigates how user interfaces within open data portals can be systematically evaluated to support sustainable, user-centered open data ecosystems. While open data portals play a pivotal role in facilitating data discovery and reuse, their design often fails to reflect user needs, limiting their effectiveness. This research addresses this gap through a multi-dimensional evaluation of user interfaces, emphasizing three key sustainability dimensions: user-drivenness, circularity, and inclusiveness. Three core research questions guide the study: (1) How can existing software testing process methodologies be adapted to evaluate user interfaces effectively? (2) In what ways can usability testing techniques be used to assess whether the conceptual design of user interfaces is user-driven and inclusive with respect to different levels of expertise? (3) How can the collective intelligence of users be evaluated through their circular interactions with open data ecosystems?

This research makes three contributions to the field of human-computer interaction by proposing methods to evaluate the user interfaces of open data ecosystems. First, it introduces a framework for acceptance testing that adapts software engineering principles to assess whether platforms meet user expectations before deployment. Second, it extends usability testing through process mining to identify user mental models by analyzing interaction logs. Third, it proposes an approach to evaluate collective intelligence by examining collaborative behaviors via user interface data. To demonstrate the feasibility of the three contributions, they have been applied to two case studies: a geospatial search engine developed by the National Geographic Institute of Spain, used to test functionality and usability as well as mental model alignment; and the HOT Tasking Manager, a platform for coordinating humanitarian mapping, which illustrates collective intelligence in action. Together, these contributions propose a novel, interdisciplinary approach for UI evaluation that combines methods from software engineering, human-computer interaction, and data science.

The findings demonstrate that sustainable, effective open data ecosystems require rigorous, user-aware interface evaluation methods. By embedding these evaluation practices into the design and deployment cycle, institutions can better align open data platforms with the needs and capacities of diverse users, ultimately strengthening the impact and inclusivity of open data initiatives.

Resumen

Esta tesis investiga cómo pueden evaluarse sistemáticamente las interfaces de usuario de los portales de datos abiertos para apoyar ecosistemas de datos abiertos sostenibles y centrados en el usuario. Aunque los portales de datos abiertos desempeñan un papel fundamental a la hora de facilitar el descubrimiento y la reutilización de datos, su diseño a menudo no refleja las necesidades de los usuarios, lo que limita su eficacia. Esta investigación aborda esta carencia mediante una evaluación multidimensional de las interfaces de usuario, haciendo hincapié en tres dimensiones clave de la sostenibilidad: orientación al usuario, circularidad e inclusividad. El estudio se guía por tres preguntas básicas de investigación: (1) ¿Cómo pueden adaptarse las metodologías existentes de proceso de pruebas de software para evaluar eficazmente las interfaces de usuario? (2) ¿De qué manera pueden utilizarse las técnicas de pruebas de usabilidad para evaluar si el diseño conceptual de las interfaces de usuario está orientado al usuario y es inclusivo con respecto a los distintos niveles de experiencia? (3) ¿Cómo puede evaluarse la inteligencia colectiva de los usuarios a través de sus interacciones circulares con los ecosistemas de datos abiertos?

Esta investigación hace tres aportaciones al campo de la interacción persona-ordenador al proponer métodos para evaluar las interfaces de usuario de los ecosistemas de datos abiertos. En primer lugar, introduce un marco de pruebas de aceptación que adapta los principios de la ingeniería de software para evaluar si las plataformas cumplen las expectativas de los usuarios antes de su despliegue. En segundo lugar, amplía las pruebas de usabilidad mediante la minería de procesos para identificar los modelos mentales del usuario analizando los registros de interacción. En tercer lugar, propone un enfoque para evaluar la inteligencia colectiva examinando los comportamientos colaborativos a través de los datos de la interfaz de usuario. Para demostrar la viabilidad de las tres contribuciones, se han aplicado a dos estudios de caso: un motor de búsqueda geoespacial desarrollado por el Instituto Geográfico Nacional de España, utilizado para probar la funcionalidad y usabilidad, así como la alineación de modelos mentales; y el HOT Tasking Manager, una plataforma para coordinar la cartografía humanitaria, que ilustra la inteligencia colectiva en acción. En conjunto, estas contribuciones proponen un enfoque novedoso e interdisciplinar para la evaluación de la interfaz de usuario que combina métodos de la ingeniería de software, la interacción persona-ordenador y la ciencia de datos.

Las conclusiones demuestran que los ecosistemas de datos abiertos sostenibles y eficaces requieren métodos de evaluación de interfaces rigurosos y conscientes de las necesidades de los usuarios. Al integrar estas prácticas de evaluación en el ciclo de diseño y despliegue, las instituciones pueden adaptar mejor las plataformas de datos abiertos a las necesidades y capacidades de los distintos usuarios y, en última instancia, reforzar el impacto y la inclusividad de las iniciativas de datos abiertos.

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.

Dagoberto José Herrera-Murillo
September 2025

ACKNOWLEDGEMENTS

First and foremost, I give thanks to God, whose love imbues all human endeavors with meaning, and to Our Lady of Pilar, whose presence has filled my days in Zaragoza with joy and hope.

I am grateful to my supervisors, Dr. Javier Nogueras-Iso and Dr. Francisco J. López-Pellicer, for their support and guidance throughout the course of this project.

My thanks go to my family, who have accompanied me at every stage of this journey, and to my community in Costa Rica, which continues to give purpose and direction to my efforts.

I am also thankful to the CNIG team in Madrid—especially Paloma Abad-Power—for facilitating such an enriching and productive stay. Special thanks to my friend Héctor Ochoa-Ortiz, whose support, expertise, and infectious enthusiasm were instrumental in making this project a success.

To all my colleagues from the ODECO project and the IAAA, thank you for your companionship and support throughout these years.

Finally, this thesis is dedicated to all those who work tirelessly in the service of open government and open data. The movement you lead holds the promise of a more prosperous, democratic, transparent, and just world for everyone.

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 Sklodowska-Curie grant agreement No. 955569.

CONTENTS

List of Figures **xv**

List of Tables **xvii**

List of Abbreviations **xix**

1 Introduction **1**
1.1 Motivation 1
1.2 Challenge and research questions 5
1.3 Contributions and methodological approach 7
1.4 The Structure of the thesis 9

2 A framework for the acceptance testing of user interfaces **11**
2.1 Introduction 11
2.2 Related Work 13
2.2.1 Functional testing 13
2.2.2 Effectiveness 14
2.2.3 Usability testing 14
2.2.4 Test software processes 15
2.3 Testing Framework 16
2.3.1 Functionality quality attribute 18
2.3.2 Effectiveness quality attribute 20
2.3.3 User-friendliness quality attribute 20
2.4 Case study: the evaluation of the new IGN search engine 22
2.4.1 Functional testing results 24
2.4.2 Relevance evaluation results 26
2.4.3 Usability testing results 27
2.5 Discussion 32
2.6 Summary 33

3 Identification of mental models from usability tests **35**
3.1 Introduction . . . 35
3.2 Related work . . . 37
3.2.1 Usability testing . . . 37
3.2.2 Process mining . . . 38
3.2.3 Mental models in human-computer interaction . . . 39
3.2.4 The information-search process . . . 41
3.3 Methodological framework . . . 42
3.3.1 Examination of the target system to identify its conceptual model . . . 42
3.3.2 Design and execution of a search task with representative users . . . 43
3.3.3 Analysis of interactions to infer mental models of users . . . 45
3.3.4 Comparison of the conceptual model with the inferred mental model of users 45
3.4 Experiments and Results . . . 46
3.4.1 Examination of the target system . . . 46
3.4.2 Design and execution of a search task with representative users . . . 47
3.4.3 Analysis of interactions . . . 48
3.4.4 Model comparison . . . 54
3.5 Discussion . . . 56
3.6 Summary . . . 57

4 Identification of collective intelligence in open data ecosystems **59**
4.1 Introduction . . . 59
4.2 Related work . . . 63
4.2.1 Crowdsourced Work in OSM . . . 63
4.2.2 Collective Action in OSM . . . 65
4.2.3 Intelligent Action in OSM . . . 66
4.3 Background . . . 67
4.3.1 Understanding the Context of Humanitarian Mapping and HOT-TM . . . 67
4.3.2 Understanding the HOT Tasking Manager . . . 68
4.4 Methodology . . . 71
4.4.1 Data Collection . . . 72
4.4.2 Data Analysis . . . 73
4.5 Results . . . 75
4.5.1 Profiling of Humanitarian Mapping Projects . . . 75
4.5.2 Data Analysis . . . 76
4.6 Discussion . . . 89
4.6.1 Towards a More Sustainable Group Composition of Humanitarian Mappers: Developing Novice Mappers and Enhancing Local Participation . . . 90

4.6.2 Towards a More Meaningful Collective Action in Humanitarian Mapping: Facilitating Effective Collaboration Among Mappers 92
4.6.3 Towards More Intelligent Collective Action in Humanitarian Mapping: Fostering Greater Equity Without Compromising Productivity 93
4.7 Summary 95

5 Conclusions and Future Work **99**
5.1 Summary of contributions 99
5.2 Work in progress 101
5.3 Future work 102

6 Conclusiones y trabajo futuro **105**
6.1 Resumen de contribuciones 105
6.2 Trabajo en curso 107
6.3 Trabajo futuro 109

References **111**

Appendix Test scenarios for the acceptance testing of user interfaces written in Gherkin **127**

Appendix A Closer Look At the Mapping Process **129**

List of Figures

1.1 Towards value-creating and sustainable open data ecosystems . . . 3
1.2 Processes facilitated by user interfaces in open data ecosystems . . . 5
1.3 The user interfaces of the geospatial search engine and the HOT Tasking Manager . . 7
1.4 Linking the research questions to the sustainability framework of open data ecosystems 7

2.1 Activity diagram describing the life cycle model of the acceptance test level. Relevant test products generated along the test activities are highlighted in blue face. 17
2.2 Geospatial search engine interface. 23
2.3 Use case diagram illustrating all the functionalities related to the search process. . . . 24
2.4 Activity diagram illustrating the search workflow. 24
2.5 Schematic graph with complete branches as single edges. 25
2.6 Box plots of System Usability Scale (SUS) scores for each testing group. 31

3.1 Methodological framework at a glance. 42
3.2 Conceptual model of the geospatial search engine. 43
3.3 Flowchart representing the test workflow. 44
3.4 Geospatial search engine interface. 47
3.5 Navigation process map. 49
3.6 Search process map including pages and actions for all participants. 50
3.7 Dotted chart distribution of the events over absolute time. 51
3.8 Dotted chart distribution of the events over relative time. 51
3.9 Variants explorer sequence of events. 52
3.10 Interaction with resources found by participants. 54

4.1 Project page in HOT-TM 69
4.2 Task state diagram showing the mapping workflow in HOT-TM 70
4.3 Methodology at a glance. 71
4.4 Analyzed projects at a glance 77
4.5 Completeness of the contributor profile by mapping level - % of total contributors- . 78
4.6 Participation structure in projects - weighted proportion of contributors per project- . 80
4.7 Frequency and duration map of task states and transitions (85% of most frequent traces) 81

4.8 Execution of mapped states based on contributor location - % of mapped states where the location of the mapper is known- 83
4.9 Number of contributors per task, depending on the occurrence of Split or Invalidated states (average number, standard deviation, and quartiles) 84
4.10 Tasks mapped by more than one contributor by project difficulty and priority 85
4.11 Variant explorer of mapping level profiles for tasks with multiple contributors 86
4.12 Handover of mapping tasks 87

List of Tables

2.1 Test strategy for the acceptance test level. 18
2.2 Precision@10. 27
2.3 Nielsen heuristics. 29
2.4 Demographics of study participants (%). 30
2.5 Median System Usability Scale (SUS) scores for each item and testing group. 31

4.1 Overview of total unique contributors by mapping level 77
4.2 Origin and destination of HOT contributions - weighted proportion of contributors from project locations. 79
4.3 State execution based on the contributor mapping level - % of states- 82
4.4 Observed and expected frequency of handovers -% of handovers- 88
4.5 Logistic Regression for validation state of a task (Validated vs. Invalidated) 89

1 Test scenarios for the search feature written in Gherkin 127
2 Cognitive walkthroughs written in Gherkin. 128

3 Task states according to frequency and case coverage 129
4 Median duration of task states and transitions (85% of most frequent traces) 130

List of Abbreviations

API	Application Programming Interface
GIS	Geographic Information System
GUI	Graphical User Interface
HOT-TM	Humanitarian OpenStreetMap Team Tasking Manager
IDE	Spatial Data Infrastructure
IGN	Spanish National Geographic Institute
OD	Open Data
OSM	OpenStreetMap
TMAP	Test Management Approach
UI	User Interface
UX	User Experience
VGI	Volunteered Geographic Information

CHAPTER 1

INTRODUCTION

1.1 Motivation

The fundamental idea behind open data is to guarantee that data can be freely accessed, used, and shared without restrictions, regardless of format [1]. According to projections by the European Commission, the open data market within the European Union is expected to reach a net value of nearly €200 billion by 2025, influencing over one million jobs in the sector [2]. Beyond fostering economic and social value, open data has the potential to enhance public services, promote transparency, boost citizen engagement, and support sustainable development goals [3, 4].

Since the open data movement began, a wide range of open datasets across Europe have become available, helping to spur innovation and produce meaningful insights. Open data now plays a central role in the European digital agenda, encouraging Member States to adapt their national policies to support this vision [5, 1]. Governments are actively promoting open data through various initiatives and by creating dedicated platforms that offer data in formats ready for reuse. Consequently, more open data repositories, directories, and platforms have been developed to meet this growing demand. Over the last ten years, these efforts have greatly expanded access to data.

An open data ecosystem is a cyclical, sustainable, and demand-driven system that relies on mutual interdependence between stakeholders to create and deliver value (See Figure 1.1) [4]. Open data ecosystems are increasingly recognized as effective frameworks for unlocking the full potential of open data, moving beyond traditional 'one-way street' models toward more dynamic and interactive environments. A successful open data ecosystem must be:

- User-driven: while many open data initiatives have traditionally been supplier-driven (releasing datasets according to administrative convenience or internal agendas) this approach often produces a disconnect between what is published and what diverse user groups actually require. A user-driven ecosystem, by contrast, places the needs of users at its core [6–8]. Achieving this shift calls for systematic feedback mechanisms that enable citizens, journalists, entrepreneurs, NGOs, and researchers to identify data gaps, flag quality issues, and specify preferred formats or levels of granularity. It also demands governance models and technological frameworks

that are sufficiently flexible and adaptive, allowing publication strategies to evolve in step with societal needs. By recognizing the heterogeneity of user profiles, such ecosystems move beyond the mere release of data to ensure that information is not only accessible, but also relevant, usable, and capable of generating both public and private value.

- Circular: traditional open data systems are often linear: providers release datasets, users consume them, and the flow stops there. Such one-way models frequently result in underutilization and uneven value distribution, with benefits concentrated among a small number of technically advanced actors. A circular ecosystem, by contrast, fosters continuous exchange and regeneration of value [9–11]. It integrates processes that enable users not only to consume data but also to contribute back by enriching, validating, or extending it. This cyclical dynamic extends the life cycle of data, maintains its relevance, and spreads benefits more equitably across stakeholders. To succeed, however, circularity must be carefully designed to avoid closed loops where only a limited group of actors participate; the goal is to build open and regenerative cycles of use that multiply value throughout the ecosystem.

- Inclusive: this dimension emphasizes that open data ecosystems should involve a broad range of stakeholders rather than privileging governments as providers and businesses as the main re-users. Inclusivity requires that citizens, NGOs, civil society groups, journalists, businesses, and marginalized communities are not merely passive recipients of data but active participants in its production, reuse, and governance. This involves integrating both government and non-government data sources, fostering cross-sector collaboration, and designing applications that are accessible to users with different levels of technical ability [12–14]. Moreover, inclusiveness is not only about who participates but also about who benefits. Addressing barriers such as lack of digital access, limited resources, unclear licensing, or concerns over privacy and reputation is required to ensure that open data contributes to social justice and broad societal well-being, rather than concentrating advantages among a narrow group of technologically skilled actors.

- Skill-based: the effectiveness of open data ecosystems ultimately depends on the skills and capacities of those who participate in them. Different user groups (from non-specialist citizens and students to companies, NGOs, intermediaries, and even artificial agents) have diverse needs and varying levels of technical competence. A skill-based ecosystem recognizes this heterogeneity and addresses it through sustained investments in capacity-building and digital literacy. Ultimately, a skill-based ecosystem is not limited to producing technically competent data workers. It requires cultivating interdisciplinary competencies that combine technical, domain, and social skills. This includes the ability to frame societal challenges, interpret data ethically, and collaborate across organizational and disciplinary boundaries.

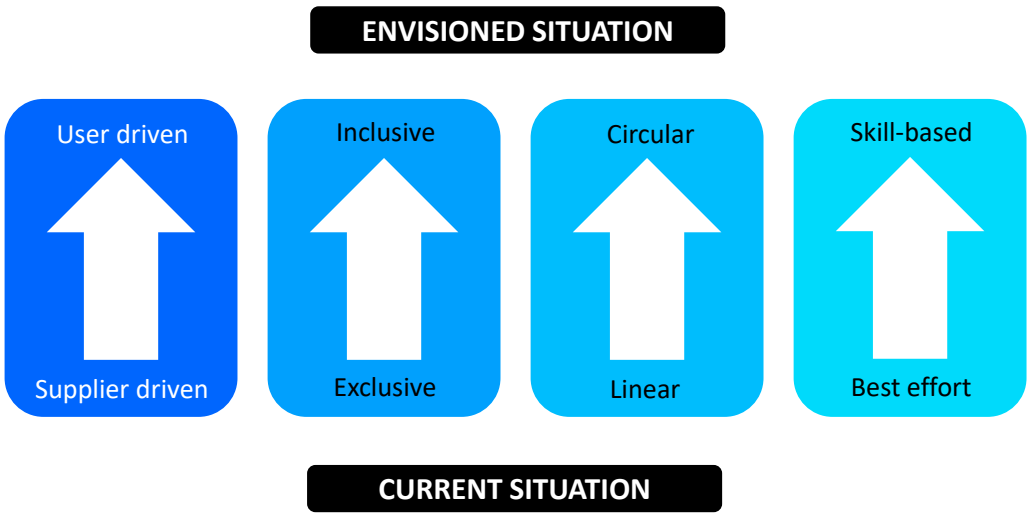


Fig. 1.1 Towards value-creating and sustainable open data ecosystems. [4]

Open data portals serve as key entry points to open data ecosystems. These online platforms provide structured and detailed descriptions of datasets, including information about their authorship, origin, and licensing [15]. Available in various forms, open data portals support essential processes such as data discovery, access, evaluation, improvement, and ingestion. The underlying premise of open data initiatives is that sharing datasets through these portals will stimulate demand for high-quality data, thereby encouraging improvements in data standards and the portals themselves. Moreover, by facilitating access to public sector data, open data portals play a pivotal role in advancing the broader open government data movement [16].

The interaction between users and open data portals is primarily mediated through user interfaces. User interfaces are widely understood in human–computer interaction as the mediating layer, or the visible part, of a system through which users interpret its capabilities and limitations [17]. They determine not only how effectively tasks can be carried out, but also the quality of the overall experience, shaping perceptions of trust, efficiency, and inclusiveness. The literature identifies several interface types, each shaping interaction in distinct ways: command-line interfaces (CLIs) provide precise control but require technical expertise; graphical user interfaces (GUIs) rely on windows, icons, menus, and pointers to enhance accessibility [18]; and more recent natural user interfaces (NUIs), including gesture- and voice-based interaction or conversational agents, expand possibilities beyond the desktop paradigm [19]. Despite this variety, all interfaces serve three fundamental functions: they translate user intentions into system actions, render system states and feedback comprehensible, and support task completion while preventing or recovering from errors [17]. Beyond their technical role, interfaces should also be viewed as socio-technical artifacts that condition participation, collaboration, and meaning-making [20]. In this sense, they shape not only what tasks are possible, but also who can realistically engage with them, given differences in expertise, literacy, or access.

This thesis contends that these interfaces should function as enablers—actively supporting the sustainability of the open data ecosystem. To do so, they must reflect the four sustainability perspectives

discussed earlier. As a result, the methods used to evaluate these interfaces should also incorporate this sustainability-oriented framework.

In the geoinformation domain, which is particularly relevant to this dissertation, well-designed interfaces can support a wide range of processes that are fundamental to open data ecosystems (see Figure 1.2). These processes typically revolve around four interconnected functions: discovery, access, evaluation, and ingestion. Discovery enables users to explore spatial datasets through keyword searches or interactive maps, while access provides standardized download formats and APIs for retrieving layers such as administrative boundaries. Evaluation involves displaying metadata, quality indicators, and version histories, allowing users to assess the reliability and relevance of a dataset. Ingestion extends these capabilities by enabling users to upload, annotate, or integrate new information—such as crowdsourced mapping data—into existing repositories. Figure 1.2 illustrates how these processes are implemented in practice. Platforms such as the European Data Portal emphasize discovery and access; the IGN geospatial search engine provides robust mechanisms for evaluation; and the HOT Tasking Manager demonstrates ingestion through the contribution of volunteered geographic information. Together, these examples highlight the central role of user interfaces in shaping not only what tasks can be accomplished but also the scope of user engagement, thereby influencing the sustainability of open data ecosystems.

Although geodata portals are a type of open data portal, they have specific characteristics that distinguish them from more general platforms [21–24]. Geographic information is inherently complex, involving spatial coordinates, projections, scales, and multi-layered datasets that demand specialized metadata standards to ensure interoperability and usability. Unlike generic datasets, geospatial data often require advanced visualization tools such as interactive maps, faceted spatial search, and download services in multiple formats. These technical requirements make the user interface of geodata portals particularly relevant, as it must balance accessibility for non-experts with powerful functions for domain specialists. Furthermore, geodata portals frequently integrate with broader spatial data infrastructures, which adds institutional and governance complexities not always present in general-purpose open data platforms. For these reasons, geodata portals provide a demanding but highly illustrative domain for evaluating user interfaces, and insights drawn from this context have the potential to inform the design and assessment of open data portals more broadly.

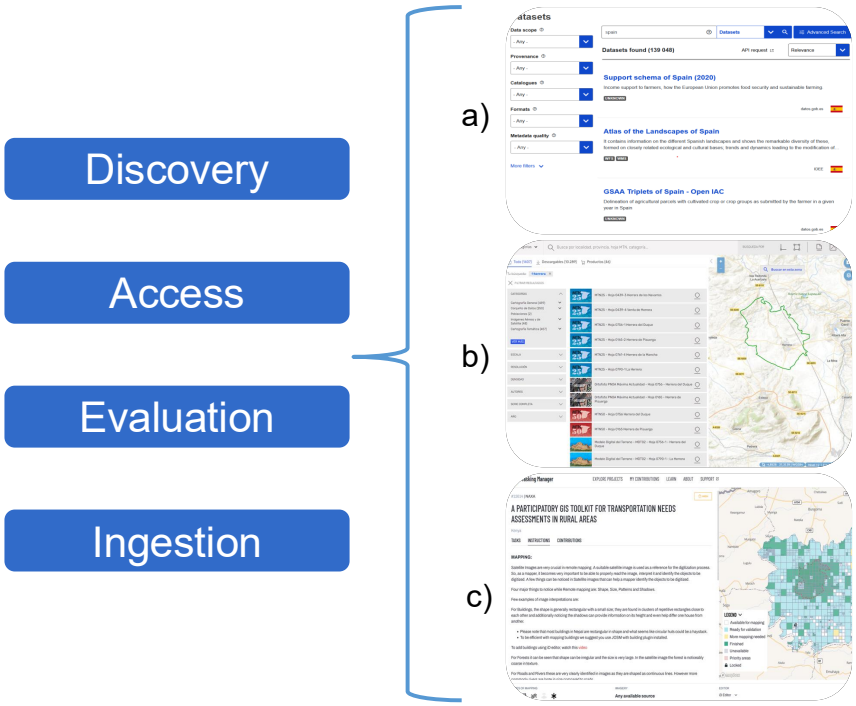


Fig. 1.2 Typical processes facilitated by user interfaces in open data ecosystems—discovery, access, evaluation, and ingestion. Examples include (a) the European Data Portal, which emphasizes discovery and access; (b) the IGN geospatial search engine, which provides mechanisms for evaluation; and (c) the HOT Tasking Manager, which demonstrates ingestion through volunteered geographic information. The figure was created using images from the European Data Portal (https://data.europa.eu), IGN Spain (https://www.ign.es), and the Humanitarian OpenStreetMap Team (https://tasks.hotosm.org).

1.2 Challenge and research questions

This PhD research tackles the challenge of evaluating user interfaces within open data ecosystems in order to foster systems that are user-driven, inclusive, and circular. Although open data initiatives have made progress, portal design is still typically led by institutions and often falls short of addressing the diverse needs of end users [4]. As a result, three persistent problems continue to undermine the sustainability of open data ecosystems. First, misalignment between supply and demand: many portals remain supplier-driven, releasing data without systematic mechanisms to capture user needs, which leads to underused or irrelevant datasets. Second, barriers to inclusivity: complex or unintuitive interfaces exclude less technical participants, and reduce meaningful engagement. Third, limited support for circular interactions and collaboration: interfaces often fail to enable users to enrich, validate, or reuse data, constraining the regenerative dynamics required for sustainable ecosystems, and weaken the collective intelligence that open data ecosystems depend on. These challenges point to deeper research gaps that remain insufficiently addressed in the literature.

- Fragmented evaluation approaches: Prior research in human–computer interaction and software engineering has typically examined isolated aspects of user interfaces, such as functional correctness, usability, or information retrieval effectiveness, but few studies integrate these perspectives into a coherent evaluation framework tailored to open data portals.
- Limited attention to user mental models: While usability testing has been widely applied, little work has systematically analyzed the alignment between the conceptual models embedded in system design and the diverse mental models of users, especially in the context of open data search and exploration.
- Lack of systematic evaluation of collective intelligence: Although research on crowdsourcing and volunteered geographic information has acknowledged collaborative behaviors, methods to evaluate how interfaces foster or hinder collective intelligence, particularly in microtasking workflows, remain underdeveloped.

User interface design plays a critical role in addressing these shortcomings, yet it remains inherently complex. In practice, many interfaces are hard to navigate, unintuitive, or fail to support sustained engagement. To illustrate these common challenges, Figure 1.3 presents two examples of functionalities provided by open data platforms that will be discussed in detail in later chapters: a geospatial search engine (the geospatial search engine of a geographic information infrastructure deployed by the National Geographic Institute of Spain) and an interface for the ingestion of volunteered geographic information (the interface of the HOT Tasking Manager). Although designed for different purposes—the former to query and explore geographic information resources, and the latter to coordinate the creation of crowdsourced data for disaster response—both platforms face a similar challenge: their user interfaces must support inherently complex processes for both expert users and those with limited technical expertise. For the latter group, an unpleasant first experience—such as struggling with a cumbersome search filter or facing unclear guidelines for data ingestion—can be decisive. It may determine whether the user abandons the platform entirely or becomes a long-term contributor to the open data ecosystem.

This research is therefore driven by a central question: How can we systematically assess the design of user interfaces in open data portals to anticipate deployment failures and guide improvements that align with sustainability-oriented principles?

The following research questions guide this PhD investigation, each addressing a key aspect of evaluating user interfaces within open data ecosystems from both technical and human-centered perspectives:

RQ1: How can existing software testing process methodologies be adapted to evaluate user interfaces effectively?

RQ2: In what ways can usability testing techniques be used to assess whether the conceptual design of user interfaces is user-driven and inclusive with respect to different levels of expertise?

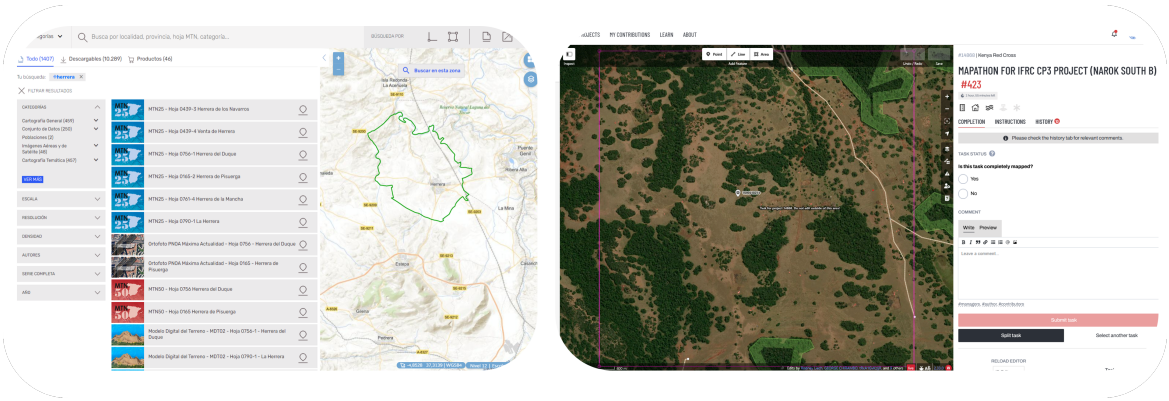


Fig. 1.3 The user interfaces of the geospatial search engine (left) and the HOT Tasking Manager (right) share a common challenge: facilitating inherently complex processes for users with varying levels of expertise.

RQ3: How can the collective intelligence of users be evaluated through their circular interactions with Open Data ecosystems?

Figure 1.4 illustrates how the three research questions are anchored in the sustainability dimensions of open data ecosystems, with each one addressing a specific research gap that motivates the study.

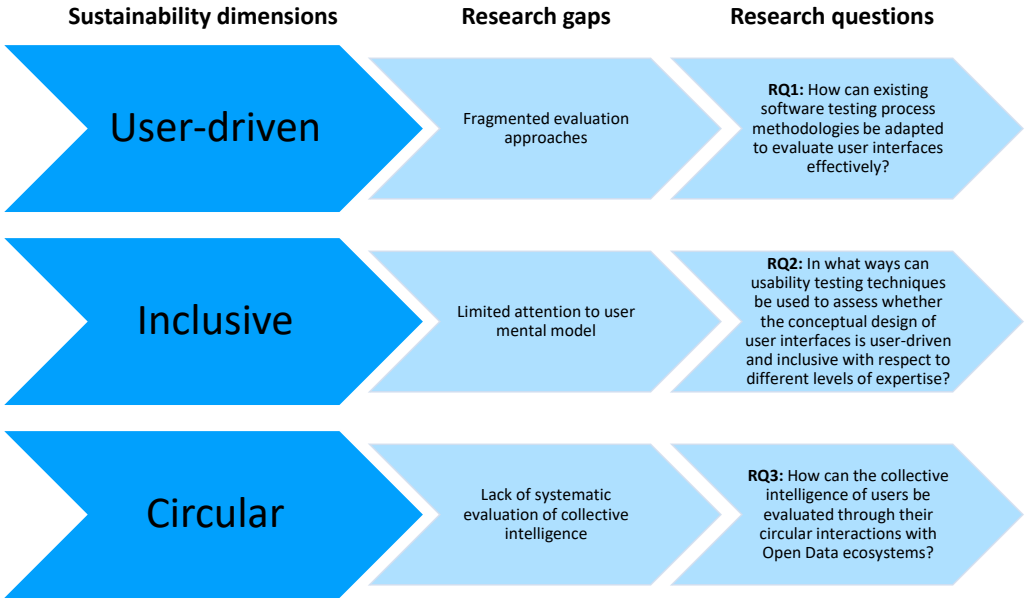


Fig. 1.4 Linking the research questions to the sustainability framework of open data ecosystems

1.3 Contributions and methodological approach

This PhD research makes the following contributions to the field of human-computer interaction for evaluating the user interface of open data ecosystems:

- A framework for acceptance testing of user interfaces, which adapts principles from software engineering to assess whether open data platforms meet the functional and usability expectations of end users before deployment. This contribution responds to the problem of misalignment between supply and demand typically observed in open data ecosystems.
- An extension of usability testing techniques through process mining, enabling the identification and analysis of user mental models by tracing interaction patterns in user interface logs. This makes it possible to detect mismatches between system design and the ways of thinking of diverse groups of users. In addition to addressing the supply–demand gap, this contribution tackles the barriers to inclusivity that are commonly observed in open data ecosystems.
- An approach to evaluating collective intelligence, using process mining techniques to analyze user interactions and collaborative behaviors within open data ecosystems, thereby uncovering how collective problem-solving emerges through interface use. This contribution responds to the problem of limited support for circular interactions typically observed in open data ecosystems.

The research methodology applied in each of these contributions consists of three main phases: a systematic review of the research literature, the design of a method/process focused on the specific contribution, and the demonstration of the viability of the contribution with specific case studies. The work with specific studies allows the refinement of the design of the proposed methods. The case studies selected to evaluate user interfaces in the context of open data ecosystems are focused on the thematic domain of geographic information.

The first case study is a geospatial search engine developed by the National Geographic Institute of Spain. This search engine uses a knowledge graph that contains more than 2 million geographical resources semantically represented in more than 150 million triples. Its objective is to make profit from a better annotation of resources to provide users with more precise results accompanied with contextual information and recommendations. On the one hand, the acceptance testing framework is customized to evaluate the functionality, effectiveness, and usability of this search engine. On the other hand, the same case study is used to analyze the discordance between the conceptual model of the user interface and the mental models of the users.

The second case study focuses on OpenStreetMap (OSM) and the Humanitarian OpenStreetMap Team (HOT) Tasking Manager—a platform built on top of OSM that facilitates coordinated humanitarian mapping efforts. OSM is an initiative supported by a global community of approximately 10.5 million registered members who have collectively contributed over 9.1 billion geographic elements to the database [25]. The humanitarian dimension of this work involves thousands of volunteers mapping critical infrastructure and features in highly vulnerable regions around the world, often in support of disaster response and preparedness initiatives. This case study facilitates information about the interactions performed by different users, which can be analyzed with process mining techniques to profile the collective intelligence emerging on this kind of platforms.

In addition to addressing each contribution individually, the thesis adopts a cyclical perspective on the software design, development, and testing of user interfaces in open data ecosystems. Evaluation is conceived not as a one-time activity but as an iterative process that continually feeds insights back into subsequent design and development phases. This perspective aligns with the sustainability principle of circularity introduced earlier: user interactions, testing outcomes, and collective behaviors are reintegrated into the lifecycle of platform improvement. By embedding acceptance testing, usability testing, and collective-intelligence analyses into this iterative cycle, the thesis shows how open data platforms can evolve more responsively, better meet user needs, and support the long-term sustainability of open data ecosystems.

The scope of this dissertation is threefold. First, thematically, the thesis is situated within the evaluation of user interfaces in open data ecosystems, with a specific focus on sustainability principles of user-drivenness, inclusivity, and circularity. Second, methodologically, the work integrates approaches from software testing, usability testing combined with process mining, and collective intelligence analysis, applying them to the context of open data portals. Third, in terms of domain, the empirical scope is centered on geographic information systems, using two case studies, the Spanish National Geographic Institute's geospatial search engine and the Humanitarian OpenStreetMap Team Tasking Manager, to demonstrate the applicability and transferability of the proposed methods. The thesis does not attempt to evaluate all types of open data portals or to provide a universal framework for every context; rather, it focuses on developing, testing, and refining methodological approaches that can be adapted to diverse domains facing similar challenges.

In this thesis, the terms portal, platform, and engine are used in a largely interchangeable manner, as they all refer to components of open data ecosystems that mediate user access to information. Strictly speaking, an open data portal denotes the web-based entry point that provides access to datasets; a platform highlights the broader technical environment that integrates the portal with governance structures, user communities, and collaborative processes; and an engine refers to the technical core that supports discovery and retrieval functions. However, given the overlapping nature of these concepts in practice and the fact that many systems combine these functions within the same artifact, the thesis uses the terms flexibly when referring to user-facing systems for data discovery and interaction.

1.4 The Structure of the thesis

The remainder of this thesis is divided into chapters covering different dimensions of user interface evaluation in the context of open data ecosystems.

- Chapter 2 introduces a framework for the acceptance testing of user interfaces. This framework is applied to assess a geospatial semantic search engine developed by the National Center for Geographic Information in Spain.

- Chapter 3 centers on extending usability tests with process mining. Specifically, it examines the alignment between the user interface and the mental models of users.
- Chapter 4 focuses on the evaluation of collective intelligence through the process mining of user interface logs. Specifically, it analyzes the HOT Tasking Manager, a prominent initiative in the field of Volunteered Geographic Information (VGI).
- Chapter 5 summarizes the key contributions of the thesis and outlines potential directions for future research and development.

CHAPTER 2

A FRAMEWORK FOR THE ACCEPTANCE TESTING OF USER INTERFACES

2.1 Introduction

Open data portals have become fundamental instruments for unlocking the value of public data. A core objective of these portals is to facilitate user access to data, often by means of integrated search engines that allow users to explore, discover, and retrieve relevant datasets. Yet, despite their potential, many portals remain supplier-driven, releasing data without systematically accounting for user needs, which contributes to a persistent supply–demand gap in open data ecosystems [4]. Addressing this gap requires rigorous evaluation methods that ensure user interfaces are not only functional but also aligned with diverse user expectations. The objective of this chapter is therefore to address the research question (RQ1) posed in the introduction about the possibility of adapting software testing methodologies for the evaluation of user interfaces in open data portals. To this end, we propose a framework for the acceptance testing level of open data portals that combines the principles of the Test Management Approach (TMAP) [26] with domain-specific adaptations and the selection of appropriate testing tools.

According to the IEEE Standard Glossary of Software Engineering Terminology [27], software testing is the process of evaluating a system to verify that it meets specified requirements or to identify any discrepancies between actual and expected results. Within this broad definition, it is important to frame the specific context of acceptance testing as used in this research. Testing activities are usually distinguished in different levels according to the party that takes the leading responsibility in each level: on the one hand, the supplying party (the development team) is in charge of the development test level (unit and integration testing) and the system test level in order to assure that the delivered system complies with the expected system requirements, technical specifications and technical design; on the other hand, the accepting party (the contractor) is in charge of the acceptance test level to assure that the received system is the one really expected by contractors, i.e., it meets the user needs and is ready for operational use.

Independently of the design architecture behind search engines, the user interface plays an essential role in the success of a search engine product and this is typically the core object of

analysis in acceptance testing. The proposed framework for acceptance testing aims at evaluating three main quality attributes of search engines: functionality (ability of the system to accurately and comprehensively process information), effectiveness (the capacity of the system to deliver a desired output), and user-friendliness (the ease with which end-users use the system). In the case of functionality testing, we propose the application of branch testing and scenario testing as testing design techniques. In the case of evaluating the effectiveness, we propose the measurement of evaluation relevance metrics. Last, in the case of the evaluation of user-friendliness, we propose the application of usability testing techniques considering both the inclusion of users (usability tests) or not (heuristics checking, cognitive walkthroughs).

For demonstrating the viability of this proposed framework we have used as a case study the context of spatial open data ecosystems, i.e. geographic information infrastructures providing open access to geographic information resources. The advances in geographic information systems (GIS), remote sensing platforms or location-aware devices, among other examples, have motivated the spread on the Web of an enormous volume of geographic information resources in various formats and representations. In order to deal with this volume of data, spatial data infrastructure (SDI) initiatives were launched since the end of the nineties at different administrative levels (regional, national or global) and with the collaboration of both public and private institutions. SDIs can be defined as a cohesive framework encompassing technologies, institutional structures, and policies designed to improve the availability and accessibility of spatial data [21]. They are structured as a hierarchical network of nodes, with key technological components including spatial data, metadata, middleware services (enabling functions such as data location, visualization, and download), and end-user applications at each node [24]. Among these components, metadata, catalogue services and geospatial search engines are instrumental for discovering geographic information resources [23, 22]. Geospatial resources play a key role in activities such as ecosystem monitoring, climate analysis, and resource management [28–30]. Many sources of relevant data remain underutilized due to interfaces that require highly specialized technical expertise [31].

In this context, geospatial search engines are a typical example of user interfaces that illustrate the common challenges faced by the search functionalities of open data portals across multiple domains. Geospatial search engines, as well as any other open data search interfaces, demand rigorous testing methodologies that address the complexity of these artifacts in an integrated manner. While existing research has explored individual testing aspects for software products in the geo-information domain, such as functionality or usability [32–35], there is limited work on frameworks specifically tailored for geospatial search engines handling large-scale datasets. The insights gained are broadly transferable to comparable interfaces across other open data ecosystems

In particular, the proposed framework has been applied to the geospatial search engine developed by the Spanish National Geographic Institute (IGN). This project aims to make the cartographic resources of the institution more accessible to the public by means of the development of a Knowledge Graph that integrates various sources in the IGN geospatial data ecosystem [36]: the database of

datasets in multiple GIS formats available through the IGN Download Center;[1] the catalogue of the IGN Map Library,[2] which contains historical cartography assets; and the catalogue used as back-end at the Online Shop of the National Center for Geographic Information for selling IGN products in hardcopy format.[3] Many of these information resources are directly related to natural environmental systems — such as satellite and aerial images, land use data, or digital elevation models — which are of interest to both experts and the general public.

The rest of this chapter is structured as follows. Section 2.2 presents an overview of related work. In Section 2.3, we delineate the proposed framework adapting TMAP for the acceptance testing of a geospatial search engine. Section 2.4 describes how to instantiate this framework in a real case study: the newly developed geospatial search engine of the Spanish National Geographic Institute (IGN). Finally, Sections 2.5 and 2.6 present a discussion of the results and outline the main conclusions along with directions for future research, respectively.

2.2 Related Work

This section reviews the state of the art of work related to the role of testing in the development of geographic information software products and the three types of testing relevant to our case study of geospatial search engines: functional testing, effectiveness evaluation and usability testing. In addition, the concept and role of test process methodologies are also presented.

Well-developed testing improves the quality, competitiveness, and demand for geographic information software products [32, 33]. Galimova [32] draws a distinction between manual testing, valued for its flexibility and ability to closely replicate user actions, and automated testing, appreciated for its ease of reuse. Through her research, which focused on three classes of geographic information systems (mobile, server, desktop), the author concludes that, for all the GIS classes under consideration, semi-automated testing emerges as the preferred approach.

2.2.1 Functional testing

With respect to functional testing for acceptance tests, there is a line of work in software engineering known as early testing [37] that aims at facilitating the automation of user acceptance tests based on the definition of requirements. There are several works in this area oriented towards the generation of test cases based on the semi-formal representation of use cases [38], and some tools of increasing use such as Cucumber[4] have popularized the automation of user tests starting from a minimally controlled plain text describing functional requirements of the system under test [39].

There are also several approaches for the design of functional test cases derived from the user interface definition. In fact, as one of the few elements on which an agreement is reached between

[1] https://centrodedescargas.cnig.es/CentroDescargas/home
[2] https://www.ign.es/web/catalogo-cartoteca/
[3] https://www.cnig.es/locale?lang=en
[4] https://cucumber.io/

client and developer during the analysis phase of an application are the user interface prototypes and the way in which they will be navigated, the automation tools that try to involve the accepting party in the testing process are based on the use of diagrams that model the behavior and interaction with the application user. For instance, tools such as Testar [40] or NDT-Suite [41] are based on this approach.

2.2.2 Effectiveness

With respect to the assessment of the effectiveness, it must be noted that the most common approach to measure the satisfaction of users in information retrieval systems is to compile metrics related to the evaluation of the relevance in the list of results returned by these systems for a list of information needs under control [42]. Therefore, since a geospatial search engine can be also considered as an information retrieval system, the evaluation of the relevance should be an appropriate indicator to assess the effectiveness of the system to retrieve relevant results. The effectiveness of a search engine providing a ranked list of results should also take into account the ability of a system to return first those results that are relevant. For that purpose, the 11-point interpolated average precision or Mean Average Precision (MAP) measures are typically computed. However, precision-recall curves or MAP consider the precision at all recall levels and this is quite unfeasible in the case of search engines indexing millions of records: relevance judgments should be available for all the documents in the collection [43]. Therefore, many search engines use as evaluation measure the precision at fixed low levels of retrieved results, i.e., *Precision at K* [44–46].

2.2.3 Usability testing

Regarding the usability assessment of search user interfaces, Hearst [47] provides a comprehensive overview of research on the evaluation of this kind of interface. Search interfaces should be evaluated in terms of efficiency, effectiveness, and satisfaction. In particular, the subjective reaction of participants to the interface is a critical factor in determining the likelihood of use. The evolving nature of interface development requires evaluations with varying levels of complexity and detail. In the early stages of comparing candidate designs, designs are shown to participants, and their responses are recorded to identify areas for improvement. These are often referred to as informal usability studies. Later, formal usability studies through controlled experiments allow us to understand how the target users use the interface and determine whether the design concepts work as expected [18]. Finally, for operational interfaces, it is important to conduct studies in which participants use the search platform in their daily routines and environments over a significant period of time [48].

User-unfriendly interfaces and poor GUI design are identified as some of the main problems in the current geospatial software ecosystem [49]. Related to the collaboration of users in the evaluation of user interfaces, it is worth noting the work of Popelka et al. [50] and Kalantari [34]. Popelka et al. [50] employed an eye-tracking method to evaluate the user-friendliness of map-based visual analytics tools, and their conclusions encourage a stronger use of mixed research designs that combine the advantages of quantitative and qualitative methods. Such designs include think-aloud protocols, which provide

deeper insights into user reasoning and the causes of errors when interacting with interactive maps. Kalantari [34] evaluated spatial metadata systems by conducting think-aloud usability testing and semi-structured interviews with users.

2.2.4 Test software processes

It is also worth mentioning test process methodologies that could be applied to define a workflow of activities for the testing of geospatial search engines. van Veenendaal [51] has asserted that the failure to implement test process methodologies is a fundamental factor contributing to system releases falling short of expectations in terms of quality, cost, and timely delivery. The same author declares that while testing theory advocates complete adherence to structured testing as the optimal and most effective solution, in real-world situations, a professional tester often is capable of choosing a minimal set of testing practices from a structured testing approach. He defines this approach as 'good enough testing'. Its success depends on a clear definition of testing priorities and appropriate risk assessment. Similarly, Vukovic [52] emphasize the importance of predefining the test process methodology rather than conducting it ad hoc. Keeping this in mind, organizations can choose from existing models or customize one to suit their needs. The same research acknowledges the challenges faced by small and medium-sized companies in formalizing their testing procedures, attributed to constraints such as limited time and human resources. There is a recognized necessity to simplify the complexity of available testing models to enhance their feasibility and implementation in such organizations.

When referring to general software test process methodologies, we can cite the Test Management Approach (TMAP) [26]. It is a well-known structured process methodology for software testing, that proposes a life cycle model to structure all the activities required for the management, preparation and execution of test processes. Van Banerveld et al. [53], who employed TMAP to assess the efficacy of a natural language processing tool, acknowledge the distinctive challenge presented by query systems dealing with massive and complex data. The case study carried out by the latter author focuses on the TMAP notion of quality attribute.

With a more specific focus on specific test processes for semantic search engines, it is worth noting the existence of frameworks like the Large-Scale Semantic Evaluation (SEALS) project [54]. The core of the SEALS project is a two-phase process. The automated phase involves collecting non-interactive metrics such as execution success, number of results returned, execution time, and system load. The user-in-the-loop phase requires real users to perform specific search objectives on the search engine. In this process, a number of user-centric metrics are collected, such as the time taken to obtain a successful response and the user impression of the tool through questionnaires. Another test process specifically designed for search engines is the proposal of Zhou et al. [55]. It proposes the application of metamorphic testing as an essential technique to evaluate search engines by comparing relationships between inputs and outputs of different search engines.

The framework proposed in the following section is derived from the TMAP methodology. In addition, it is based on core principles outlined in each of the three types of testing relevant for

geospatial search engines: automating test cases based on representative cases from the functional testing domain, *Precision at K* from the assessment of effectiveness and the use of realistic scenarios from usability tests.

It is worth noting that other widely adopted software test process models [52], such as ISO 29119-2 [56] and Test Maturity Model Integration (TMMi) [57], also provide structured methodologies for software testing. While these frameworks share the common goal of establishing systematic testing processes, the selection of TMAP as the main reference framework for this proposal is due to the alignment of its characteristics with the dynamic nature of geospatial search engines. One of its key advantages is its flexibility, agility, and lightweight structure, making it well-suited for environments where search relevance, indexing mechanisms, and knowledge graph updates are continuously evolving. Additionally, its strong support for automated testing facilitates the evaluation of search UI interactions. Finally, its emphasis on quality attributes aligns with the evaluation needs of modern geospatial search engines.

2.3 Testing Framework

In our proposed acceptance testing framework, we have adapted the Test Management Approach (TMAP) to the case of this type of geospatial search engines. For each test level, TMAP delineates a life cycle model for organizing the test activities across seven phases: planning (activities are detailed later); control, which encompasses monitoring and readjusting of the planning; setting up and maintaining the infrastructure; preparation of test cases; specification of test scripts; execution of tests and reporting of results; and completion, which consists of the evaluation of the test process and preservation of the testware for future test processes.

Figure 4.3 presents an activity diagram showing the life cycle model of an acceptance test level and the logical order of these phases. It can be observed that control and infrastructure phases are transversal activities executed in parallel to the activities devoted to preparation, execution and completion. In addition, it can be observed that the planning phase has been subdivided in various activities and that we have highlighted in blue face the test products that are relevant for our acceptance test process.

During the planning phase, after understanding the mission of an acceptance test level and identifying potential sources for a test basis, we must analyze the product risks to identify and prioritize the test objectives, i.e., determine the ‘quality attributes’ of the system that are critical for a geospatial search engine. Then, we need to determine a test strategy that defines the intensity of testing for every quality attribute. Later, we need to estimate and schedule the required human resource efforts. The following step is devoted to identify the ‘test types’ and ‘techniques’ that are more appropriate for designing the test cases associated with each quality attribute. The planning finishes with the definition of other remaining planning details related to the list of deliverables, organization, infrastructure or management. Outside the planning phase, it must be noted that the preparation phase includes the compilation of a specific ‘test basis’ (information defining the required

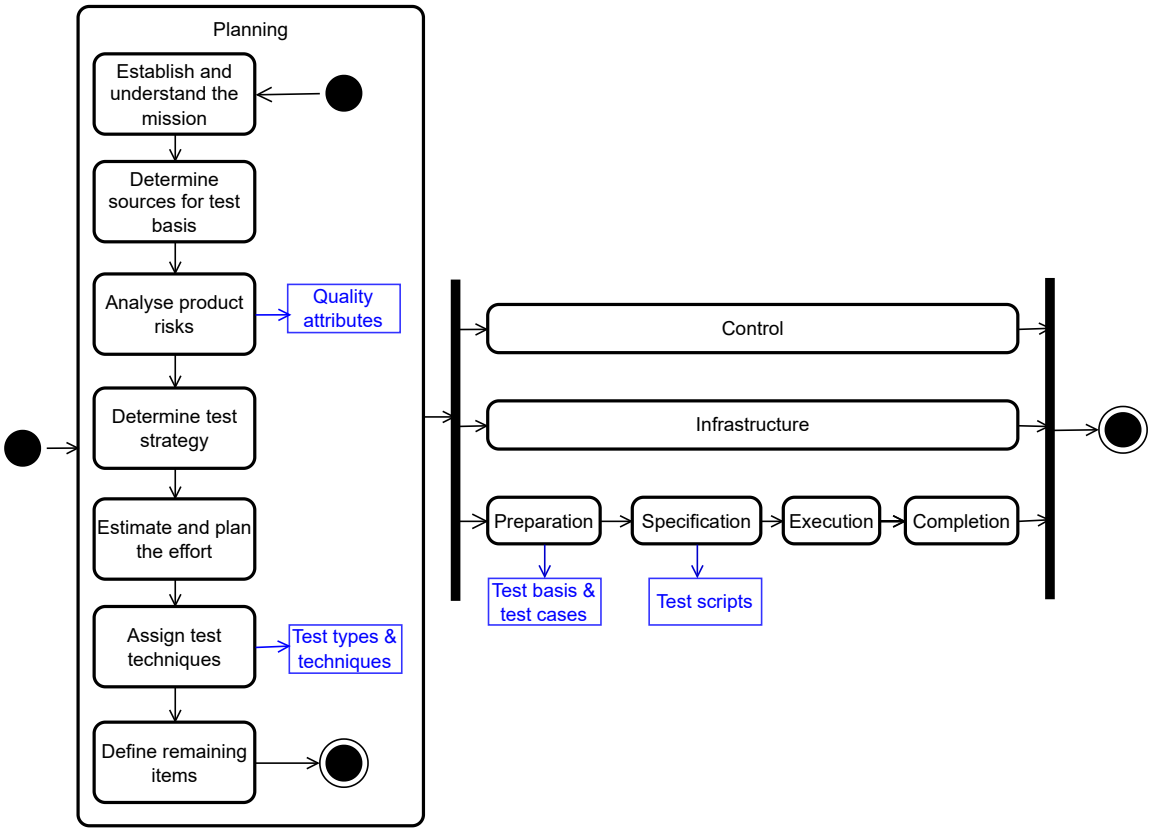


Fig. 2.1 Activity diagram describing the life cycle model of the acceptance test level. Relevant test products generated along the test activities are highlighted in blue face.

behavior of the system) for the definition of test cases. In addition, the main output of the specification phase is the test scripts: automatic programs or manual procedures for the execution of test cases.

Table 2.1 summarizes the test products that are relevant in our proposed acceptance test process for geospatial search engines. As can be observed, we have identified three main quality attributes that are critical for the development of these geospatial search engines: functionality (ability of the system to accurately and comprehensively process information), effectiveness (the capacity of the system to deliver a desired output), and user-friendliness (the ease with which end-users use the system). The choice of these three quality attributes is grounded in both international standards for software testing and established evaluation practices for search engines. We used as main source the definition of these attributes by TMAP, but they are consistent with other well-known sources for software engineering like the ISO/IEC 25010 standard [58], which also highlights functionality (the degree to which the system facilitates and cover the realization of all specified tasks and objectives), effectiveness (the degree to which a system achieves the required level of precision in delivering results) and usability (the degree to which the system facilitates and cover the realization of all specified tasks and objectives) as main attributes to describe quality in the product or usage model of software and systems. As already noted, the information retrieval literature consistently highlights

retrieval effectiveness as a defining dimension for search engine evaluation [42]. Taken together, these references justify the division into the three selected attributes: functionality ensures that the search engine processes and delivers results correctly; effectiveness guarantees that the retrieved results are relevant to information needs; and user-friendliness ensures that the search interface supports successful and satisfactory task completion. This triad therefore provides a scientifically grounded and holistic framework for acceptance testing of user interfaces in search engines.

Table 2.1 Test strategy for the acceptance test level.

Quality attribute	Test type	Testing design technique		Test basis	Execution
Functionality	Functional	Scenario testing		Use cases	Automatic
		Branch testing		Navigation map (UI workflow)	Automatic
Effectiveness	Relevance evaluation	Ranking evaluation measures (prec@10)		Relevance evaluation benchmark	By experts
User-friendliness	Usability	Without users	Heuristic evaluation	User interface	By experts
			Cognitive walkthrough	Tasks proposed by experts	Automatic
		With users	Usability test	Usability test scenario	By users

For each quality attribute, we provide recommendations on the most appropriate test types and test design techniques: functionality testing for assessing the functionality quality attribute; relevance evaluation for assessing the effectiveness of this type of system; and usability to assess user-friendliness. Each test design technique is aligned with a specific test basis, serving as the main source of information needed for both test case definition and specification of test scripts. Sections 2.3.1, 2.3.2 and 2.3.3 describe how the three different quality attributes have been assessed by selecting the most appropriate test types, test design techniques, and test tools. In addition, each design technique produces a metric for a quantitative assessment of the quality attribute.

It is important to note that the results of the attribute testing are part of a larger methodological process, in which relevant stakeholders interpret the results and incorporate them into subsequent stages of analysis. Rather than remaining as isolated measurements, these results contribute to the evaluation cycle by providing evidence for discussion and informing user interface design and development decisions.

2.3.1 Functionality quality attribute

For assessing the functionality quality attribute, we propose the application of two different test design techniques employed in functionality testing: branch testing and scenario testing. The following subsections describe the use of these techniques.

2.3.1.1 Scenario testing

Scenario testing is a specification-based technique in which testers design cases to evaluate how the software will function when end-users interact with it for specific purposes [59]. This technique involves developing a model of the interaction sequences between the test item and users to test the usage flows that involve the test item [60].

We employ a form of scenario testing called use case testing [61]. This method involves a use case model of the test item that outlines how it interacts with one or more actors. Since use cases are employed to express requirements in the early stages of development, they serve as an excellent basis for acceptance testing. From the description of use cases, we can define features and associated test scenarios using the Gherkin language [39]. This language allows to specify the expected behaviors of the software in a human-readable way: it is a minimally controlled language containing just plain text and a reduced set of reserved words. We decided to use this language because it is designed to write acceptance tests that can be implemented using Cucumber, a platform widely used for functional testing and behavior driven development. The ultimate goal of Gherkin is to facilitate the understanding of software testing or behavior by technical and non-technical team members.

With respect to the infrastructure needed for the automation of these tests, we propose the use of Behave [62], a Python implementation of Cucumber that converts Gherkin scenarios into Python test scripts. In addition, to interact with the web interface of a geospatial search engine, we propose the integration of Selenium [63]. Selenium is an open-source automation testing framework for web-based applications, which offers a Selenium WebDriver that allows interaction with most modern web browsers. That is to say, test scripts written in Python (and also other programming languages) can integrate a specialized library to interact with the Selenium WebDriver to trigger different User Interface (UI) events (e.g. open/close web pages, typing text or mouse events) in an automated way.

2.3.1.2 Branch testing

Branch testing is a structure-based technique that evaluates the system by following possible logical branches in its functional flow [64]. We conducted branch testing considering the decision points and flow restrictions outlined in Figure 2.4. In particular, we derived test cases by applying the test depth level N technique [26]. According to this technique, achieving test depth at a certain level N implies that all the combinations of N consecutive branches are covered. For instance, test depth level 1 is equivalent to achieve full branch coverage as stated in part 4 of the ISO 29119 standard [59]. With test depth level 2, all combinations of branches going in and out of each decision point are covered, or, equivalently, all subpaths of two consecutive branches starting at each decision point. Test depth level 3 covers all subpaths of three consecutive branches starting at each decision point and so on.

With respect to the infrastructure for the automation of these test cases, we propose to express these paths as Gherkin scenarios because Gherkin allows expressing a sequence of actions in an almost human-readable way. In addition, the translation of these steps into interactions with the geospatial

search engine can be implemented in the same way as proposed for scenario testing, i.e. using Behave and Selenium WebDriver.

2.3.2 Effectiveness quality attribute

As already introduced in section 2.2, the effectiveness of search engines and information retrieval systems is usually assessed in terms of the evaluation of the relevance in the list of results returned by these systems. To measure the performance of an information retrieval system in terms of relevance evaluation, we need a relevance evaluation benchmark, also known as test collection, comprising three components: a document collection; queries expressing information needs; and a set of relevance judgments (usually a binary assessment) for each query-document pair. The most common measures for information retrieval effectiveness without taking into account the ranking of results are precision (the proportion of retrieved documents that are considered relevant), recall (the proportion of relevant documents retrieved) and the F-measure (a weighted harmonic mean of the previous measures). However, the effectiveness of a search engine providing a ranked list of results should also take into account the ability of a system to return first those results that are relevant. For that purpose, the 11-point interpolated average precision or Mean Average Precision (MAP) measures are typically computed: on the one hand, the 11-point interpolated average precision is a precision-recall curve consisting of the average interpolated precision of the considered information needs at 11 fixed recall points; on the other hand, MAP computes the arithmetic mean of the average precision of each considered information need, which averages the precisions whenever a relevant document is retrieved.

The problem of precision-recall curves or MAP is that they consider the precision at all recall levels. However, this is quite unfeasible in the case of search engines indexing millions of records: despite using a small test suite of information needs for relevance evaluation, it is not possible to have relevance judgments for all the documents in the collection. On the other hand, this may not be relevant to final users of search engines because they are usually interested only in the first page of results. Therefore, for the purpose of relevance evaluation in the proposed framework of this work, we propose to measure the precision at fixed low levels of retrieved results. This is referred to *Precision at K* or *Precision@k*, i.e., the precision computed when the top k documents are retrieved. This measure is widely used for the evaluation of web search engines and offers sufficient conditions for acceptance testing, where testers do not have direct access to the full list of resources ranked by relevance.

2.3.3 User-friendliness quality attribute

In the case of the evaluation of the user-friendliness quality attribute, we propose the application of usability testing techniques considering both the inclusion of users (usability tests) or not (heuristics checking, cognitive walkthroughs). The customization of these techniques for the case of geospatial search engines is explained in the following subsections.

2.3.3.1 Tests without users: Heuristics checking

Heuristic evaluation entails expert raters applying a usability checklist to a user interface in order to spot potential usability issues that could prevent users from carrying out their intended tasks [65]. In a classic heuristic evaluation, the user interface is checked to verify whether specific design criteria are followed to enhance the user experience.

In particular, we propose the use of the ten heuristics established by Nielsen [66] for evaluating user interfaces: visibility of system status (1); match between system and the real world (2); user control and freedom (3); consistency and standards (4); error prevention (5); recognition rather than recall (6); flexibility and efficiency of use (7); aesthetic and minimalist design (8); help users recognize, diagnose, and recover from errors (9); and help and documentation (10). An expert evaluator executes typical browser search tasks, annotates violations in a standardized worksheet, and assigns severity ratings to the violated heuristics.

2.3.3.2 Tests without users: Cognitive walkthroughs

Cognitive walkthrough is a usability evaluation technique that connects the interface review to a cognitive model [67]. The evaluator simulates the experience of a typical user by performing tasks on the interface. The process compares the user expectations with the actual steps required by the interface to complete the tasks.

The accepting party must collaborate in the identification of specific examples of information needs required by new users of the geospatial search engine. After defining these information needs, it is necessary to think about the actions that should be performed in the search engine to accomplish these information needs. Then, these information needs are expressed as a sequence of steps in a Gherkin scenario. On the one hand, Gherkin facilitates a well-known language to express these actions. On the other hand, these actions can be automated in the same way as proposed for scenario testing and branch testing.

Finally, in order to identify potential deviation of the automated executions of Gherkin scenarios from the expected behavior, we propose to record the execution time each scenario and the number of results that were returned.

2.3.3.3 Tests with users: Usability tests

In a usability test, a researcher asks a participant to perform representative tasks using one or more specific user interfaces. During the task completion, the researcher observes the participant behavior and listens for feedback [68]. The primary objectives of usability tests are to identify any issues with the system design and gain insight into the behavior and preferences of our target users.

For the testing sessions we propose the use of the ‘think-aloud strategy’ [69], a well-established method for gathering data in user studies where participants are asked to vocalize their thoughts while

performing tasks. This provides insight into the user reasoning, perception, and difficulties with the search tasks through the interface.

To understand the patterns of perception and use by novice and specialized users, the recruitment of participants must take into account individual differences in search performance, as studied in the literature [47]. Factors such as knowledge of the task domain, experience as searchers, and cognitive differences are considered to define three types of participants: I. Non-experts, II. Non-familiar experts, III. Familiar experts. The difference between unfamiliar and familiar expert users is that the latter regularly use the platforms and products of the institution whose search engine is going to be tested.

Finally, participants must complete the System Usability Scale (SUS) questionnaire, a widely used survey [70], after completing the task. This questionnaire consists of 10 items rated on a 5-point Likert scale, and the final score ranges from 0 to 100.

2.4 Case study: the evaluation of the new IGN search engine

This section illustrates how the proposed framework can be instantiated in a real case study such as the geospatial search engined developed by IGN. Figure 2.2 shows the main web page of the search engine interface. This interface allows final users to have access to different functionalities related to the search process as illustrated in the use case diagram of Figure 2.3. The diagram highlights how the core functionality begins with the Search action, which is followed by the Display results use case containing multiple extension points. From this point, users can refine their interaction by applying filters, viewing metadata, locating specific resources, downloading them, or, when applicable, proceeding to purchase. This structured representation provides a clear overview of the system capabilities and the optional paths available to users once results are displayed. It also serves as the basis for defining functional test cases, ensuring that acceptance testing covers the different scenarios and extensions that may arise in real user interactions.

In order to describe better the overall search process enabled by the geospatial search engine, Figure 2.4 provides an activity diagram of the search workflow. The diagram begins with the selection of a search mechanism, which can include free text, spatial queries such as selecting a point or geometry on the map, uploading a geometry file, or specifying coordinates or cadastral references. After submitting a query, the system processes the request and displays the results, which then serve as the basis for further user actions. Users may refine results by applying faceted filters, narrowing down to downloadable resources, or filtering by product categories. Once results are filtered, users can select individual items and choose from several actions, such as viewing metadata, centering the map on the resource location, displaying download options, or initiating purchase. This diagram highlights the logical flow and decision points of the search process, making it possible to design branch testing cases that ensure all potential paths and outcomes are systematically evaluated.

The selection of the IGN search engine as the case study was guided by several criteria. First, the platform was in the final stages of development at the time of this research, providing an opportunity to apply acceptance testing prior to its full deployment. Second, the system complexity, based on a

large-scale knowledge graph that integrates millions of geospatial resources, made it a demanding and representative environment for testing the framework. Third, the institutional role of the IGN as the national provider of geospatial data in Spain ensured the platform was relevant for open data infrastructures of national scope. Finally, practical feasibility was secured through the availability of collaboration with the IGN technical team and access to system documentation and stakeholders. While the findings are grounded in this specific case, the methodological approach was designed to be transferable to other open data portals of similar scale and complexity, thereby supporting the external validity of the research.

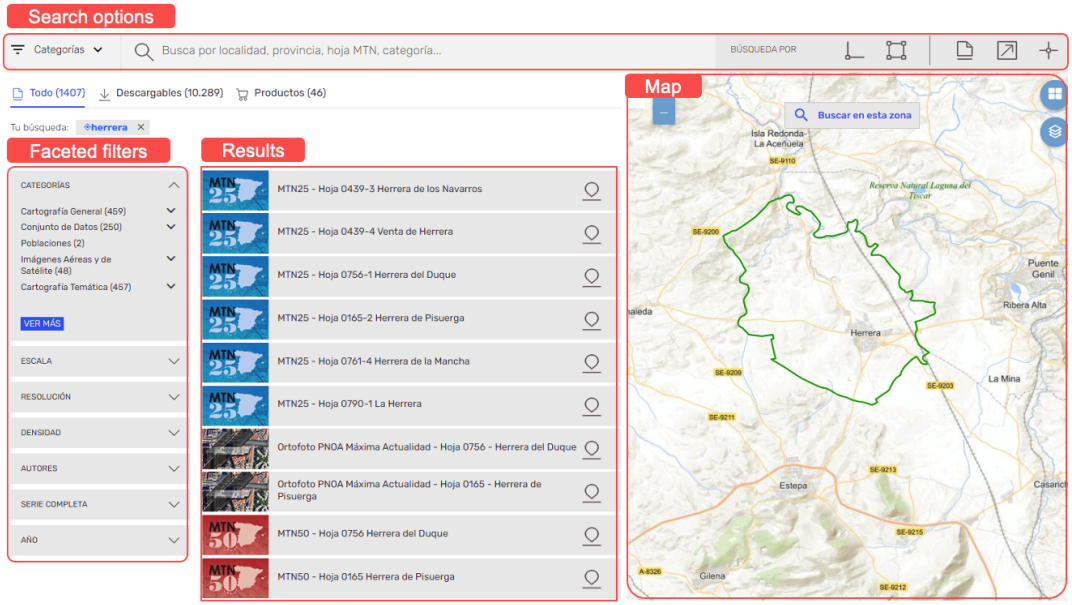


Fig. 2.2 Geospatial search engine interface.

The following subsections provide practical guidelines for the implementation of the testing design techniques in our proposed framework. In some cases, such as scenario testing, branch testing or cognitive walkthroughs, test scripts can be fully automated. For the rest of the techniques, we provide full details for the preparation of test cases and their manual execution. In all cases, we describe the obtained results after the execution of tests. There is also a code repository[5] with the implemented scripts for automated tests and some Python notebooks for the analysis of results.

[5]https://github.com/IAAA-Lab/Acceptance-testing-of-geospatial-semantic-search-engines-ODECO-CNIG

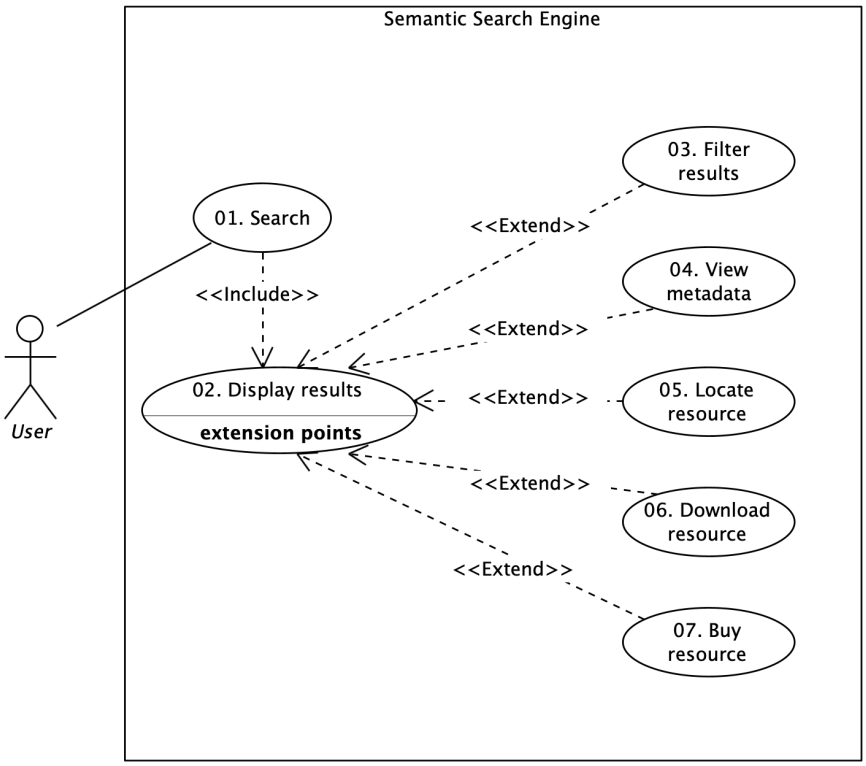


Fig. 2.3 Use case diagram illustrating all the functionalities related to the search process.

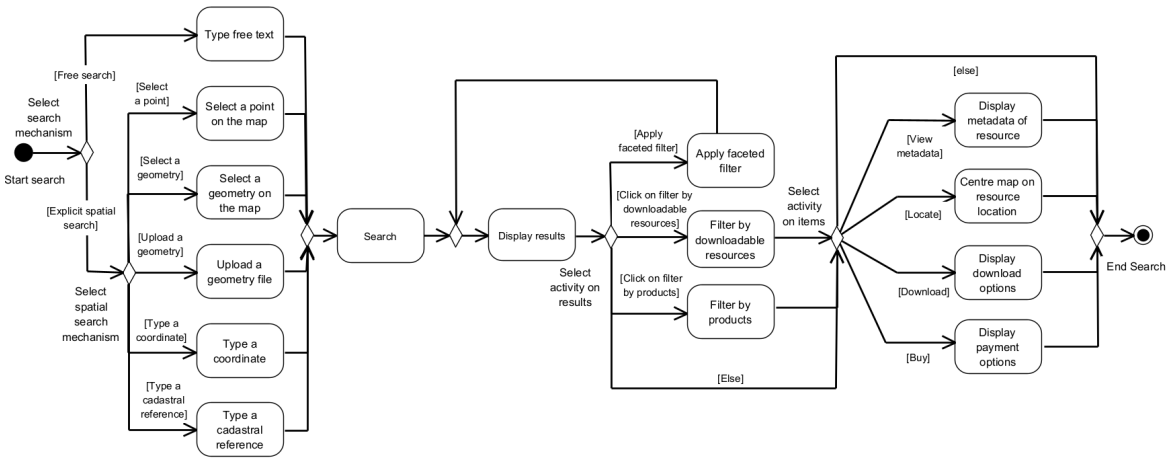


Fig. 2.4 Activity diagram illustrating the search workflow.

2.4.1 Functional testing results

2.4.1.1 Scenario testing

The use cases employed for our testing were provided by the search engine development team, who had already specified and justified them in detail at the beginning of the project in the technical

documentation. The resulting 7 use cases are the ones already shown in Figure 2.3: Search, Display results, Filter results, View metadata, Locate resource, Download resource and Buy resource.

The implementation of the Gherkin scenarios associated to the 7 features (use cases) are provided in the code repository (see *scenario_testing.feature* and the implementation of steps in these scenarios). The scenarios associated to the search feature (shown in Table 1 in Appendix 'Test scenarios for the acceptance testing of user interfaces written in Gherkin') are the most complex because they include a variety of inputs. During the execution of the test cases, we did not encounter any incidents. Based on our testing results, we can confirm that the search engine of the platform satisfies the functional requirements for which it was designed. In addition, the implementation and execution of these test cases helped us to make the implementation of the Gherkin steps as much generic as possible. This was important because the implementation of test cases derived from the application of branch testing and cognitive walkthrough testing design techniques were also expressed in Gherkin language.

2.4.1.2 Branch testing

Figure 2.5 presents a simplified version of the workflow illustrated in 2.4, which was used to derive test situations according to the test depth level N technique. In this schematic representation, the only nodes depicted, apart from the initial and final ones, correspond to the decision points in 2.4, while the edges represent complete branches or paths connecting one decision or starting node to another decision or final node in the original diagram. By reducing the detailed workflow to its essential structure, the figure provides a clearer overview of the alternative routes a user may follow during the search process. This abstraction facilitates the design of branch testing cases, ensuring coverage of all possible paths. For instance, edge labeled '2' represents the complete branch where the user selects 'free search' in the first decision node, types free text, performs the search and gets the results displayed, ending in the decision node where the user must choose among different types of filters.

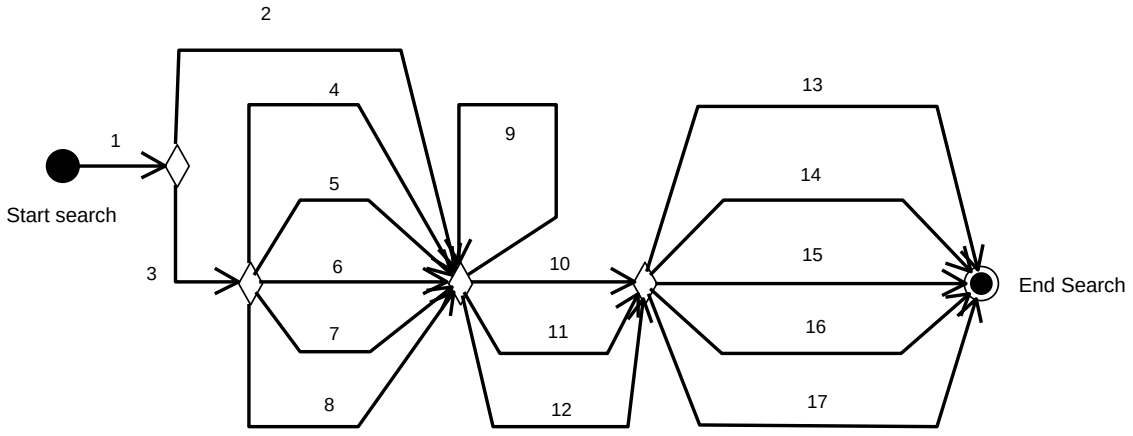


Fig. 2.5 Schematic graph with complete branches as single edges.

In our case, test depth level 2 provides a total of 46 test situations (that is, 46 pairs of consecutive branches in the graph), like '1-2' or '2-9'. These 46 test situations can be exercised in 24 execution paths or test cases. For instance, '1-3-4-9-11-17' is one of these paths and represents a test case where the user starts the search, chooses an explicit spatial search, selects a point in the map, performs the search, gets results displayed, applies a faceted filter, filters the results by products, buys and is displayed payment options.

In order to make the testing more exhaustive, we wanted to make sure that all the possible combinations of the six different types of search with the four possibilities of filtering the results and the five possible actions that can be performed with the results were also tested. Given our graph, the later can also be achieved with the test depth technique, switching to level 3. There are a total of 134 test situations in this case, that can be covered with 90 paths. In order to decrease the number of test cases, while testing all the aforementioned combinations of types of search, filters and actions, we designed again the test cases to cover all the level 2 test situations with those level 3 test situations that do not include the branch labeled with '9', that loops back to the filtering section. In this last case, we obtained a combined total of 99 test situations.

The obtained test cases cover 66% of the test depth level 3 test situations, 100% of the level 2 test situations and also 100% of the branch test situations according to part 4 of the ISO 29119 standard [59], while keeping the number of test cases in a reasonable level.

To cover the selected 99 test situations we executed a total of 66 paths. The description in Gherkin of these paths is available at *branch_testing.feature* file of the code repository, together with its associated Python implementation.

About the results of the execution of Gherkin scenarios, during the first iteration we found that 54 (82%) of the paths executed without any problems. The remaining 12 tests failed, all of which are associated with search mechanisms based on the selection of points or polygons on the side map. The instructions to interact with Selenium WebDriver were programmed to perform the selection in the center of the map, which by default shows the entire Iberian Peninsula with a portion of the ocean, and whose center falls in Portugal. We discovered that the execution failure occurred because the search engine did not have any resources indexed for that area in the download or purchase category. Therefore, any successive operations to view, locate, download or purchase resources simply could not be executed. To fix this issue, we moved the selection zone to Spain, which allowed us to confirm that the filters were working correctly. This exercise enabled us to confirm that users would eventually be able to perform the search sequences without encountering errors or crashes.

2.4.2 Relevance evaluation results

According to the proposed methodology, we calculated the *Precision@10* of the semantic search engine with some information needs. But taking into account the existence of other search engines at IGN, we also compared the measure with the one obtained by the three current geospatial search

engines existing at IGN for discovering separately resources on the Download Center, the Map Library and the Online Shop.

Previous to the computation of *Precision@10*, we had to prepare the evaluation benchmark. For that purpose, we proposed first five information needs, which are shown *information need* column in Table 2.2. Second, we compiled the first ten results returned by the four search engines with the search terms associated with these five information needs. Third, we had to annotate the relevancy of all the results with respect to the information needs. This task was performed by three experts (the judges). Finally, those results having a majority of relevant votes were considered relevant for the computation of *Precision@10*. In addition, to assess the agreement between the judges, we computed the Fleiss' kappa measure, which is an extension of Cohen's kappa measure to evaluate the agreement between two or more judges [71]. The results indicated a substantial agreement among the judges with $\kappa = .70$.

Table 2.2 Precision@10.

Information need	Semantic Search Engine	Map Library	Online Shop	Download Centre*
Discover cartographic resources of the autonomous community of **"Asturias"**	1.0	1.0	0.5	0.1
Download a trail file related to the search for the **"Way of El Cid"**	0.4	0.0	0.0	0.0
Buy the current map of the city of **"Toledo"**	1.0	0.8	0.0	0.4
Discover general cartographic resources of the region of **"Murcia"**	1.0	1.0	0.4	0.0
View the area of the **"Sierra Nevada"** National Park on the side map	0.4	0.9	0.6	0.5
Average prec@10	0.8	0.7	0.3	0.2

*It is not strictly a free-text search mechanism. The user types in the search and must necessarily choose one of the suggested terms and results are not displayed in one single list.

Table 2.2 shows the *Precision@10* measure obtained by each search engine. Overall, the semantic search engine outperformed its counterparts in both average precision and individual searches, with the exception of one case. The Library Map had a decent performance, but its relevant results were limited to historical cartography. In contrast, the Download Center and the Online Shop had notably poor performance, with some queries yielding no relevant results.

2.4.3 Usability testing results

2.4.3.1 Tests without users: Heuristics checking

Usability testing without users started with the verification of the Nielsen heuristics. Table 2.3 contains the list of violations associated with each of the ten heuristics and the assigned severity. Violations

were categorized into low, medium and high severity. Low severity violations are those that may slow down search tasks, but would not necessarily prevent the user from finding the necessary information. Medium-severity violations are those that may create significant obstacles in search tasks, which may cause confusion or delay, but do not completely prevent the user from finding or correctly interpreting the necessary information. High-severity violations are those that may prevent the user from finding or correctly interpreting useful resources that are actually contained in the system.

Upon reviewing the detected violations, we identified two distinct categories. The first category comprises violations related to search and navigation structures, such as the search bar, filters, and icons. The second category is associated with the quality of content. In addition to these specific violations, we also observed transversal deficiencies in navigation and content, such as the systematic use of jargon. Of all the heuristics used to detect violations, the one that targets the internal and external consistency of the search engine reported the highest number of violations.

2.4.3.2 Tests without users: Cognitive walkthroughs

As indicated in section 2.3.3.2, the experts of the accepting party proposed five user tasks (searching of information needs) for applying the technique of cognitive walkthroughs. The 'Test case' column of Table 2 shows the sequence of actions for these five tasks as Gherkin scenarios. The implementation of these Gherkin scenarios is provided in the code repository, see *cognitive_walkthrough.feature* and the implementation of steps in these scenarios.

The columns *Passed*, *# results* and *Ex. Time* in Table 2 in Appendix 'Test scenarios for the acceptance testing of user interfaces written in Gherkin' show the results after the execution of the test scripts. All tests met the defined criteria within execution times of less than one minute. When reviewing manually the execution of the walkthroughs contained in the test scripts, no deviations from the expected results were observed, and the results were found to be consistent with the search criteria.

2.4.3.3 Tests with users: Usability tests

For the usability test, we selected a search task that represented the projected use of the platform and was easily understood by novice and expert users, planning a trip to a popular tourist destination in Spain:

'As you plan your visit to the Sierra Nevada National Park, some information about the area is required. Use the search engine to find information, download files, or add products to the cart that could be useful for your trip'.

The usability test adhered to the Ethics Appraisal Procedure established by the European Union Horizon 2020 Program.[6] This adherence included procedures to enlist appropriate participants and an informed consent protocol. The demographics of the recruited participants for the usability test are

[6]https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

Table 2.3 Nielsen heuristics.

#	Heuristic	Notes	Severity		
			Low	Medium	High
1	Visibility of System Status	Updates of the shopping cart or download area is not noticeable	X		
2	Match between the System and the Real World	The search engine contains a wealth of jargon unknown to the general public and even to experts outside the institute			X
3	User Control and Freedom	The user can remove previous searches, but cannot edit them directly to refine them	X		
4	Consistency and Standards	Different applications of the same institute handle different icons for the same functions		X	
		The descriptions of some of the resources do not correspond appropriately to the files or products deployed			X
		Sometimes the result list does not intuitively represent the demarcated area on the side map			X
		The number of resources shown by the filters does not correspond to the result list		X	
		Sometimes the result list does not intuitively represent the demarcated area on the side map	X		
5	Error Prevention	Some search mechanisms that require the input of parameters, such as coordinates, lack default values to help the user identify the expected format		X	
		Some of the resources point to empty metadata records	X		
6	Recognitition Rather Than Recall	Most of the icons lack tooltips		X	
7	Flexibility and Efficiency of Use	Several search mechanisms cannot be combined efficiently and even combining them worsens the result			X
8	Aesthetic and Minimalist Design	The thematic category filters have dozens of options that are displayed simultaneously	X		
9	Recognise, Diagnose and Recover from Errors	No suggestions are shown when searches are unsuccessful or when resources are empty	X		
		The error messages are not informative	X		
10	Help and Documentation	The files lack sufficient description to know their contents without first downloading them		X	

shown in Table 2.4. Overall, there are no significant disparities in the composition of gender, age, and education among the three groups of participants. At the start of each session, a pre-test questionnaire was administered to confirm the classification of participants into their respective user categories.

Table 2.4 Demographics of study participants (%).

	All	I. Non-experts	II. Non-familiar experts	III. Familiar experts
Gender				
Male	17 *(57%)*	5 *(50%)*	6 *(60%)*	6 *(60%)*
Female	13 *(43%)*	5 *(50%)*	4 *(40%)*	4 *(40%)*
Age				
18-24	2 *(7%)*	1 *(10%)*	1 *(10%)*	- *(0%)*
25-34	6 *(20%)*	2 *(20%)*	2 *(20%)*	2 *(20%)*
35-44	6 *(20%)*	- *(0%)*	1 *(10%)*	5 *(50%)*
45-54	13 *(43%)*	5 *(50%)*	5 *(50%)*	3 *(30%)*
54-65	2 *(7%)*	2 *(20%)*	- *(0%)*	- *(0%)*
+65	1 *(3%)*	- *(0%)*	1 *(10%)*	- *(0%)*
Education				
High School	1 *(3%)*	1 *(10%)*	- *(0%)*	- *(0%)*
Graduate	18 *(60%)*	6 *(60%)*	5 *(50%)*	7 *(70%)*
Postgraduate	11 *(37%)*	3 *(30%)*	5 *(50%)*	3 *(30%)*
Total	30	10	10	10

During the tests, recordings were made and the moderator of the sessions took notes of the user feedback. This content was then reviewed by the accepting party, who identified three major areas for improvement that were consistently mentioned by users: a) the role of the side map needs to be rethought to make it a truly interactive visualizer that is closely linked to the search results, b) the faceted filters and categories need to be redesigned to make their meaning and operation more intuitive, and c) a better guidance must be provided explaining how to use the retrieved geographic resources. Expert users also provided specific suggestions about how to improve the information architecture of the search engine by providing examples from other platforms and previous experiences. Novice user comments were generally more limited in detail and expressiveness. These comments were also passed on to the accepting party for consideration.

The System Usability Scale (SUS) applied at the end of each of the 30 usability tests reports a value of $\alpha = .85$ for Cronbach's alpha, which means that the reliability of the questionnaire is high. Table 2.5 shows the results for item scores and overall SUS scores. A Kruskall-Wallis test was used to determine if there were significant differences between testing groups. Only items 7 and 8 showed significant differences. Post-hoc tests using the Bonferroni correction showed that non-exerts and familiar experts are different for item 7, while non-familiar experts and familiar experts are different for item 8 at statistical significance at level $\alpha = .05$.

Table 2.5 Median System Usability Scale (SUS) scores for each item and testing group.

SUS items	All	I	II	III	p
	Median score contribution (0-4)				
1. I think that I would like to use this system frequently	3	3	3	3	*0.55*
2. I found the system unnecessarily complex	2	3	2	1.5	*0.06*
3. I thought the system was easy to use	3	3	3.5	3	*0.19*
4. I think that I would need the support of a technical person to be able to use this system	3	3	4	3	*0.07*
5. I found the various functions in this system were well integrated	3	3	3	2	*0.27*
6. I thought there was too much inconsistency in this system	2	2	2.5	2.5	*0.67*
7. I would imagine that most people would learn to use this system very quickly	2	3	2.5	1	*0.01**
8. I found the system very cumbersome to use	3	3	3.5	2	*0.04**
9. I felt very confident using the system	3	3	3	2	*0.51*
10. I needed to learn a lot of things before I could get going with this system	3	3	4	3	*0.25*
SUS Score (0-100)	67.5	70	75	58.8	*0.17*

I. Non-experts, II. Non-familiar experts, III. Expert Familiar Users
* indicates the statistical significance at level a=.05

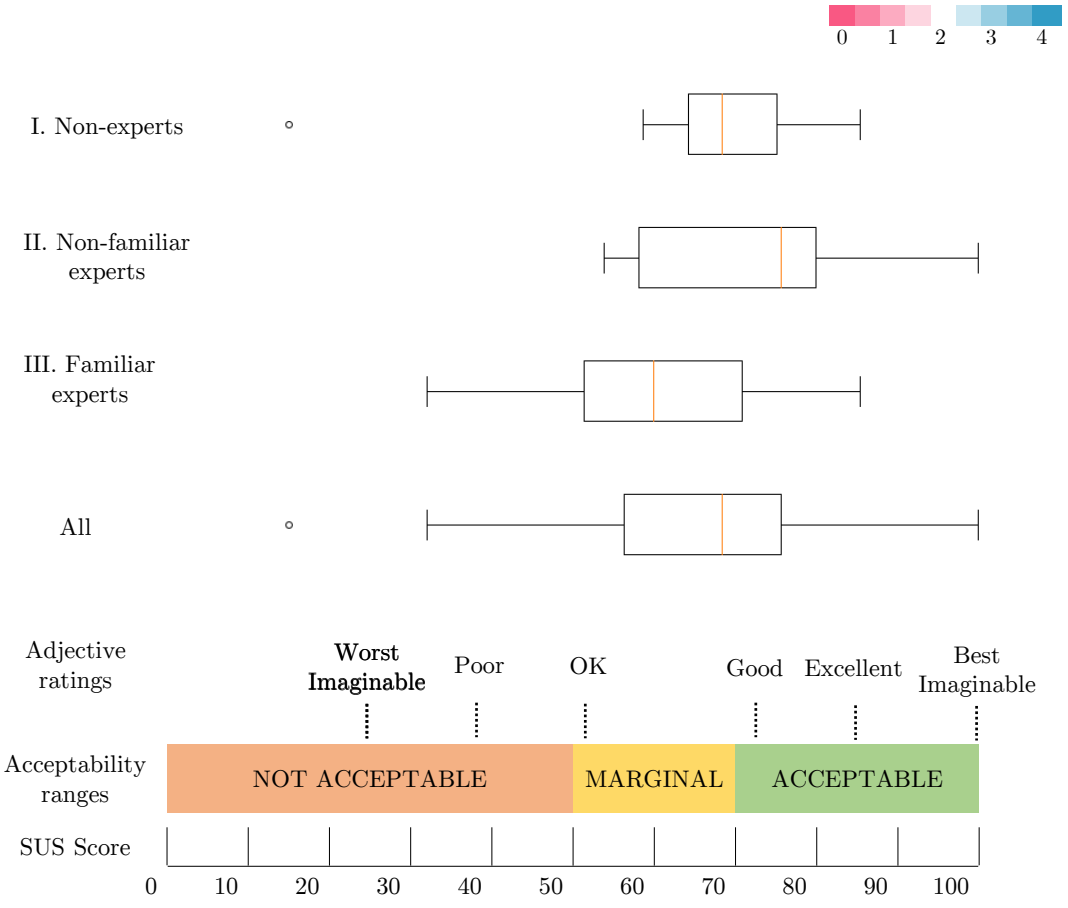


Fig. 2.6 Box plots of System Usability Scale (SUS) scores for each testing group.

While no statistically significant differences were found in overall SUS scores, Figure 2.6 presents box plots for each testing group and situates them within the adjective ratings and acceptability ranges proposed by [72]. This complementary system helps to clarify how numerical scores correspond to

qualitative judgements of usability: scores below 50 fall in the 'NOT ACCEPTABLE' range and are typically described as 'Poor' or even 'Worst Imaginable'; scores between 50 and 70 correspond to a 'MARGINAL' range (often labelled 'OK') that reflects usability which is just passable but not yet satisfactory; and scores above 70 are generally regarded as 'ACCEPTABLE', aligning with adjectives such as 'Good', 'Excellent', or 'Best Imaginable'. The overall SUS mean of 67.5 is positioned at the upper edge of the marginal range (barely acceptable, but approaching the boundary of what is considered 'Good').

2.5 Discussion

This chapter introduced a framework for the acceptance testing of geospatial search engines that aims to evaluate their functionality, effectiveness, and user-friendliness through a structured approach grounded in software testing principles. The framework builds on the TMAP methodology, distinguishing itself from other approaches that assess geographic information systems without an explicit testing process [32, 33]. By integrating various test design techniques—such as functional tests and cognitive walkthroughs—tailored to web-based interfaces, the framework offers a comprehensive strategy for evaluating search engines. Automated implementation of test scripts, when feasible, further supports the reproducibility and efficiency of the evaluation process.

Although the study did not conduct an empirical comparison with other testing frameworks, the conceptual analysis revealed notable distinctions. Compared to the SEALS approach, which emphasizes metric identification [54], the proposed framework places stronger emphasis on the integration of testing techniques within the software development lifecycle. Similarly, while the use of think-aloud protocols aligns with established practices in user interface evaluation [47], the proposed framework goes further by embedding these protocols within a structured testing process. This is particularly relevant in the geospatial domain, where similar methods have been used [34].

In applying the framework to a new semantic search engine developed by the Spanish National Geographic Institute, the case study demonstrated its practical utility. Functional testing confirmed the platform's ability to support various geographic search mechanisms. Relevance assessments indicated improvements over previous institutional platforms. However, usability testing highlighted specific limitations, particularly in the visual representation of results, the structure of filtering options, and the guidance provided to users. Interestingly, differences emerged in perceptions between novice and expert users, especially regarding the system's learnability.

We identified several potential threats to the validity of our framework and implemented measures to mitigate their effects. First, the limited size of samples used for relevance evaluation (particularly Precision@k) and usability testing can constrain the robustness and generalizability of the results. This underscores the importance of carefully selecting evaluation tasks so that they can reflect the diversity of information needs encountered in real-world use, with the same care applied to choosing representative participants for testing. Second, the constraints introduced by the moderation method based on the verbalization of actions (which, in this case, had to be employed due to security

restrictions) may have introduced a degree of artificiality that limited the validity of some results. In particular, requiring participants to articulate their actions could have slowed task performance, constrained spontaneous exploration, and disrupted the natural flow of interaction. However, it is important to emphasize that the framework itself remains independent of such contextual constraints. This limitation is best understood as an example of the kinds of adaptations that testing teams may encounter when applying the framework in their own contexts. At the same time, this unforeseen restriction, arguably an extreme variant of the think-aloud technique, may also have yielded indirect benefits. By focusing attention on verbalizing actions, users may have made more explicit some of the advantages typically associated with thinking aloud, such as the expression and capture of intentions, reasoning, and decision-making processes. Lastly, the reliance on a single geospatial search engine developed by IGN represents a limitation in terms of generalizability. Whilst this focus enabled a detailed, end-to-end demonstration of feasibility, it does not in itself confirm applicability across a wider range of systems. Nevertheless, the framework was deliberately grounded in widely adopted and domain-independent testing principles such as the TMAP life-cycle model, Precision@k for relevance evaluation, and the System Usability Scale (SUS) for usability. These principles strengthen its potential portability to other geospatial platforms.

A key consideration for the future development of this framework is its adaptability to different organizational contexts. From a scalability perspective, larger implementations would benefit from automated testing pipelines capable of managing high volumes of queries, benchmarking performance across heterogeneous datasets, and ensuring reproducible results under diverse semantic configurations. One example would be its application to cross-border infrastructures involving large-scale geospatial data. At the same time, the adoption by smaller organizations may be constrained by limited technical capacity, financial resources, and maintenance capabilities. To address these challenges, the framework could be provided with simplified test suites that minimize the need for specialist expertise while still ensuring essential coverage. In addition, modular testing procedures would allow organizations to deploy only those components most relevant to their operational context.

Ultimately, the study contributes to the broader goal of providing practitioners with complete case studies that guide interface testing from conception to result interpretation. The integration of widely recognized software engineering testing methods into a cohesive framework reinforces its relevance for developers of complex geospatial tools that require careful consideration of user interaction dynamics.

2.6 Summary

This work has presented a structured and adaptable framework for the acceptance testing of geospatial search engines. Grounded in TMAP and enriched by a variety of testing techniques, the framework addresses the need for systematic evaluations that extend beyond metric collection to include functional, usability, and cognitive considerations. The proposed methodology bridges a gap in the testing of geographic information systems by explicitly defining a test process, aligning with software

engineering best practices, and supporting automation. These features enhance its applicability and scalability across different geospatial platforms.

In direct response to the research question (RQ1) linked to this chapter, the results demonstrate that existing software testing methodologies can be effectively adapted to evaluate user interfaces in open data portals. The framework integrates functionality, effectiveness, and user-friendliness as core quality attributes, showing how each can be systematically tested using a combination of automated and user-centered techniques. By applying these methods to the IGN semantic search engine, the study demonstrates not only that the framework is operationally feasible but also that it generates actionable insights for improving platform design. While the case study is grounded in the geospatial domain, the methodological principles (structured acceptance testing, integration of multiple quality attributes, and support for automation) are transferable to other open data platforms that face similar challenges of scale and complexity. Consequently, the framework provides both a rigorous basis for testing individual systems and a generalizable contribution to the evaluation of open data interfaces.

Future work will focus on extending the application of the framework to multiple platforms, including both institutional SDI catalogues and community-driven open-source platforms, in order to confirm its adaptability and robustness across diverse technical and organizational contexts. Moreover, the observed differences in the behavior of expert and novice users, particularly in relation to feedback expressiveness, suggest that further research is needed to clarify the origins of these differences (whether they stem from cognitive overload, task complexity, or unfamiliarity with geospatial systems), their impact on user experience, and their implications for system design. One promising avenue to pursue this is through the analysis of differences in the mental models of interaction among users with varying levels of experience [73]. Additional lines of research include the automatic translation of Gherkin test cases into scripts that interact with the web application [74], the development of field testing of geographic information search systems, which will help to compensate for the limitations of think-aloud testing by capturing user interaction under more naturalistic conditions, and the investigation of how the geographic products retrieved through the system are subsequently used in real-time decision-making.

CHAPTER 3

IDENTIFICATION OF MENTAL MODELS FROM USABILITY TESTS

3.1 Introduction

The previous chapter established the value of testing frameworks and methodologies for obtaining a general overview of user interface quality in open data platforms. While such frameworks are effective for assessing broad design attributes, certain dimensions, particularly usability, require more focused and nuanced evaluation. Usability is not only central to the effectiveness of user interfaces, but also deeply intertwined with user expectations, behaviors, and mental models. Shortcomings in usability often translate into barriers to inclusivity, as complex or unintuitive interfaces can exclude less experienced users and limit meaningful engagement. This chapter therefore delves into a detailed analysis of usability experiments with the aim of uncovering how users conceptualize and interact with the system (RQ2). By comparing these mental models with the underlying conceptual model that guided the design, we seek to identify mismatches and derive insights that can inform iterative improvements in interface design.

The design of user interfaces for data discovery poses a major challenge [75]. Searching is a dynamic user-driven process, where the mental models of users must fit with the functionalities of the system to deliver satisfactory experiences. Although it is well-established that user mental models play a crucial role in the design of user interfaces [76], it is difficult for developers to anticipate during the design and development phases what will be the most effective information architecture or how the system will be used in practice.

In many cases, the usage models envisioned by designers are not aligned with the true mental models of users potentially diminishing user experience. Usability testing lets development teams identify design problems in digital products by collecting qualitative and quantitative information. Nevertheless, this technique alone is often not able to provide a panoramic view of the interaction with the system. Titus et al. [77] point out that usability testing, like other popular evaluation methods, has limitations in providing a complete picture of usability. They identified an opportunity to introduce methods that complement traditional approaches, contributing to a more comprehensive understanding of the user experience. In particular, this chapter investigates the hypothesis that the application of

process mining techniques can derive an appropriate representation of the mental models inferred from the interaction of users with the system and reveal mismatches with the original user interface design.

Understanding user behavior is a central concern not only during the early stages of a web platform —as is the focus of this research— but throughout its entire lifecycle. This ongoing interest has motivated numerous research efforts [78–80]. Although the research literature contributes with valuable insights for improving user experience design, a common and persistent challenge remains: how to coherently and usefully organize the multitude of findings that emerge from something as dynamic and complex as human behavior. In this context, both the well-established theory of mental models in human-computer interaction and the more recent developments in process mining provide promising concepts and tools that can help to articulate and structure these insights more effectively.

Continuing with the case study of the geospatial search engine, this chapter proposes a framework for actively integrating process mining into the usability testing of open data platforms. One of the great benefits of this integration is the availability of visualization tools that allow to draw up inferences about the mental model of the users and deviations from the initial design concept. The input data in the case study was compiled through a usability testing experiment where twenty-one participants, ranging from novice to expert users, were recruited to perform a search task using the geospatial search engine. Their interactions were recorded as event logs and subjected to analysis using process mining techniques, descriptive and inferential statistics. The findings revealed that the mental model of users leans towards the archetype of a regular search engine rather than fully utilizing the geographic functionalities provided by the platform, as intended by its designers.

This chapter reflects on the potential and challenges of integrating process mining into a usability test, both before, during and after the test. This extension also incorporates other analytical dimensions (usability scores, control flow, textual search, filters, search results) needed to understand the information search experience more holistically. In addition, the principles discussed in the analyzed case study are directly applicable to design of other search engines with similar features. The availability of case studies that provide comprehensive evaluations of user interfaces in specific domains can provide valuable guidance to other practitioners with the challenge of designing new satisfactory user interfaces in related domains [81, 82].

The remainder of this chapter is organized as follows. Section 3.2 presents an overview of the current state of the art regarding usability testing, process mining, mental models in user interaction, and modeling of information-searching processes. Subsequently, Section 3.3 outlines the proposed framework for integrating process mining into usability testing in order to infer mental models of users. Section 3.4 presents the outcomes derived from applying the framework to the new geospatial search engine of IGN. Section 3.5 discusses the implications of incorporating process mining into usability testing and particular findings from the case study. Finally, Section 3.6 offers concluding remarks and highlights potential future directions for research.

3.2 Related work

This section introduces the conceptual building blocks behind the goal of integrating process mining tools in usability testing to identify potential gaps between the design model and the mental model of users within an information search scenario. Firstly, we review the state of the art of usability testing, one of the most widespread research tools to evaluate user experience. Secondly, we present process mining as a discipline that provides a realistic view of "as-is" processes from event data using a variety of visual representations. Thirdly, we explain the concept of mental models, which is one of the cornerstones of user experience design. Finally, we describe some of the main theories to explain information search processes that will be useful for understanding the case study that will be developed later.

3.2.1 Usability testing

Usability is the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" [83]. The concept of usability is directly related to that of user experience, which includes "perceptions and responses that result from the use and/or anticipated use of a system, product or service" [83]. The specialized field dedicated to improving the usability of interactive systems is known as usability engineering [84]. The various evaluation methods for assessing whether a product is usable and how users experience its use have undergone significant evolution since the early 1980s to the present day [85, 86]. These methods can be classified into those that do not require direct user participation, such as usability inspections (heuristic evaluation and cognitive walkthroughs), and those that actively involve users in the evaluation process. The latter category encompasses usability testing, as well as various interrogation methods such as surveys, interviews, and focus groups, along with field methods such as ethnography. Despite having a higher cost, multiple studies have pointed out the capacity of usability testing to detect more severe, recurrent, and global problems than other methods such as heuristic evaluation [87].

Usability testing employs a moderator to guide participants through specific tasks using a given user interface [88]. During these sessions, the moderator closely observes participant behavior and attentively listens to their feedback. This process serves to unveil issues in product design, reveal opportunities for enhancements, and provide insights into the behavior and preferences of the target user group. When it comes to the execution of usability testing, thinking aloud stands out as the most popular approach [89]. In this method, participants are requested to interact with the system while verbalizing their thoughts in real-time as they navigate through the user interface.

Usability testing can be categorized into qualitative and quantitative [90]. Qualitative usability testing focuses on collecting insights on how participants interact with the product or service based on narrative evidence. This approach is particularly effective for discovering problems in the user experience. Conversely, quantitative usability testing focuses on the gathering of metrics that describe

the user experience, enabling benchmarking and objective evaluation. In practice, the integration of both quantitative and qualitative data proves valuable, leading to studies of a mixed nature. The blend of qualitative and quantitative information can vary based on the testing stage. Tests administered during early development stages often emphasize reliance on quantitative data, whereas those conducted in later stages tend to incorporate a greater emphasis on qualitative insights [91].

Generally speaking, there is a broad consensus when describing the procedure for the execution of usability tests [92–95]. A standard procedure starts with a planning phase where purpose, objective, product to be tested, test groups, and relevant metrics are defined. This is followed by a preparation phase where participants are recruited and the test environment is set up. The critical moment comes with the execution of the test where the user interacts with the product and data is collected. Finally, the collected data is analyzed and the corresponding reports are prepared.

In the quantitative dimension, there are consolidated measures of effectiveness, efficiency, and satisfaction provided by user-based testing [95]. For effectiveness, the task completion rate computes the number of participants who successfully complete the task divided by the total number of participants. To measure efficiency, metrics such as task execution time are suggested. Finally, for satisfaction, the results obtained in the Smiley scale [96] or in satisfaction questionnaires such as the System Usability Scale (SUS) [97] are examples of representative metrics. On the other hand, the qualitative dimension seeks to analyze narrative evidence and user feedback in verbal and non-verbal formats that help explain why a metric is the way it is [90].

Multiple studies have addressed the usability of GIS applications. GIS applications exhibit a distinctive attribute wherein their usability is notably influenced by the map, setting them apart from conventional user interface elements [98]. Kurniawan et al. [99] conducted a systematic literature review to explore the landscape of usability evaluation in geographic information systems, encompassing the methods employed and prevalent usability challenges. Usability testing emerged as the most widely utilized method, with inquiry methods relying on self-reported experiences second. In this second category, the use of standardized questionnaires, such as SUS, is a common practice [100–102]. Inspections, grounded in expert knowledge and experience, were the least frequently employed method. In addition, eye-tracking and mouse-tracking techniques have been used to assess the usability of the GUI map design [35]. Recurring usability issues in GIS applications include challenges associated with user guidance, tool use, and interface design. Finally, this literature review identified the need for further development of methods to analyze GIS interactions and workflows [98].

3.2.2 Process mining

Process mining is an emerging discipline that connects the models and event data about the process. That is why this field can be viewed as the missing link between process science and data science [103].

There are three basic types of process mining [104]: (a) process discovery to automatically learn process models from event logs; (b) conformance checking to compare observed behavior with prescribed behavior; and (c) model enhancement to repair or extend existing models. In addition to

the three core types of mining, there are various orthogonal dimensions: the control flow dimension focusing on the sequence and order of activities within a process; the organizational dimension focusing on the resources and actors involved in the process and how they are interconnected; the case dimension dealing with the distinctive properties of individual cases within a process; and the time dimension focusing on the timing and frequency of events within the process. It is important to recognize that the dimensions of process mining are neither mutually exclusive nor exhaustive. Rather, they are designed as flexible constructs to adapt to dynamic and evolving use cases [105]. Together, they provide a multidimensional view of the process.

Regardless of the type of process mining or dimension one wants to analyze, these techniques are impossible to apply without adequate event logs. Van der Aalst [103] mentions four basic assumptions about event logs: (a) a process is made of cases; (b) a case is made of events and each event refers exclusively to one case; (c) events within a case are displayed in order; and (d) events can have attributes. As a result, a typical event log has basic attributes such as case, activity, and timestamp to which can be added optional attributes such as resource, status, and cost, among many others.

In terms of results, the role of human judgment in detecting and interpreting patterns makes visual process presentations one of the flagship products of any process mining analysis [106]. Among the visual designs commonly used by process mining, we can mention among others directly-follows graphs, Petri nets, process trees, dotted charts, variant diagrams, and process matrices.

Process mining is applied to all kinds of domains and the field of usability is not an exception. The term usability mining refers to integrating process mining and usability engineering [107, 108]. Thaler [107] has delved into the automated generation of software usage reference models, aiming to quantify the usability of business information systems. Meanwhile, Dadashnia et al. [108] have contributed to the field with multiple usability mining studies focused on mobile policing applications developed in Germany.

3.2.3 Mental models in human-computer interaction

According to Rouse and Morris [109], mental models are the mechanisms through which individuals construct descriptions of the goals and design of a system, formulate explanations of its operational dynamics, and generate predictions about its future states. The concept of a mental model has been widely recognized in the field of human interaction since the 1980s [110]. Despite its acceptance, the measurement, representation, and use of mental models continue to pose challenges, especially with the advancement of complex software applications [111].

During the design of information systems, it is essential to differentiate between the mental model of designers, often referred to as the conceptual model, and the mental model of users [112]. In the case of the mental model of the designer, it involves having a conceptual representation of the intended system and transforming those ideas into a tangible implementation. On the other hand, the mental model of users acknowledges their actual knowledge about the system, which is influenced by their cognitive abilities, past experiences, problem-solving strategies, and individual variances. According

to Nielsen [113], designers frequently possess intricate mental models of their own creations, which can lead them to believe that every feature is intuitively understandable. Conversely, mental models of users are often more limited, resulting in a higher likelihood of mistakes and finding the system more challenging to use.

There are multiple conceptualizations of mental models and their representations. De Kleer and Brown [114] differentiate between component models and causal models. Component models primarily focus on the structure of the system, while causal models aim to explain system functioning in terms of cause-effect relationships. Carroll and Olson [112] categorize mental models into four types: surrogates, metaphors, crystal boxes, and networks. Surrogate models imitate the input/output behavior of systems. Metaphorical models directly compare the target system with another system familiar to the user. Crystal box models are a blend of metaphors and surrogates. Finally, network models encompass system states and user actions to transition the system to a different state.

Understanding the mental model of users has two primary applications within the realm of human-computer interaction: interface design and user training [112]. Designing interfaces that align with the prevailing mental model of users can facilitate them to learn and operate the system more easily, resulting in fewer errors during the interaction. An inadequate mental model can prevent users from fully harnessing the capabilities of the system. In such cases, user training and guidance can facilitate the acquisition of appropriate conceptual models.

Jakob's Law strongly correlates with the concept of mental models. The law establishes that users form expectations regarding design conventions based on their prior experience with other websites [115]. Consequently, users are inclined to anticipate that a website will adhere to familiar design patterns and conventions they have encountered before. This suggests that users may encounter difficulties when confronted with new or unfamiliar designs.

According to Zhang [116], three methods are commonly used to elicit mental models. The first method involves eliciting verbal accounts from participants [117]. This can be achieved by asking users to describe a system, explain its mechanisms, provide analogies or metaphors, or engage in thinking aloud while performing search tasks. Transcripts of these accounts are then carefully analyzed to develop representations and evaluations of the mental models about the system under study. The second method entails drawing, wherein participants are prompted to create visual representations, such as pictures or diagrams, to depict their mental images of a system [118]. The third method involves observing errors occurring during typical tasks of the evaluated system to identify gaps in their mental models of the system. This method is often combined with think-aloud protocols when the objective is to represent mental models [119]. Although the research on mental models in human-computer interaction has been ongoing for several decades, recent years have witnessed an increasing diversity in its fields of application and the resulting need for further conceptual and methodological development [120].

3.2.4 The information-search process

Mental models are a concept that extends across a wide range of technological artifacts originating from the design process. In the context of this case study, this subsection delves into the theoretical models related to information search. Hearst [121] makes a comprehensive compilation of these models. Within these, the standard model and the dynamic model are appropriate for explaining short-term search tasks.

With slight variations, several authors have described the standard model of information searching as a cycle of interaction comprising several key stages. Initially, an information need is identified, followed by activities such as specifying the query, examining retrieval results, and potentially reformulating the query, all aimed at achieving a satisfactory set of results [122]. This model aligns with Marchionini's assertion [123] that information-searching resembles a specialized form of problem-solving, encompassing problem recognition, search planning, execution, result evaluation, and iterative processes if necessary. Sutcliffe and Ennis [124] offer a comprehensive four-phase cycle, including problem identification, articulation of information needs, query formulation, and result evaluation, associating distinct search strategies with each phase. Similarly, Shneiderman et al. [125] outline four steps: query formulation, action (query execution), result review, and refinement. Marchionini and White [126] describe a seven-step process, adding elements like recognizing the information need, accepting the challenge to act, and using the results to the information-searching journey.

The traditional model of the information-searching process assumes a static user information need, where users refine their queries until they retrieve the documents relevant to their original need. However, real-world observations of information searchers reveal a dynamic and evolving process [127]. Information needs change as users interact with search systems, learning about the topic as they review retrieval results and suggestions, and formulating new sub-questions as earlier ones are answered. This challenges the idea that the primary goal of the search process is to achieve a perfect match with the initial information need. Bates [128] introduced the dynamic (berry-picking) model, which emphasizes two key points. First, information needs and queries continually shift during the search, influenced by the information they encounter. Second, satisfaction is not derived solely from a final set of documents but from a series of selections and bits of information acquired along the way.

When discussing mental models within the context of the information search process, studies across diverse fields have observed the prevalence of the "Google-like" effect [129–131]. This phenomenon is characterized by a preference towards generic and direct search engines for accessing and disseminating information resources, often at the expense of more sophisticated or specialized functions offered by alternative types of digital libraries.

3.3 Methodological framework

Our framework aims for a seamless integration of usability testing, mental modeling, and process mining. This integration is the most distinctive factor of our proposal when compared to the other efforts presented in the state of the art. In terms of usability testing, we start from a conventional design that incorporates both quantitative and qualitative elements. From this basic structure, the critical task is identifying where and how to organically incorporate process mining for mental model analysis without overcomplicating the test.

As depicted in Figure 3.1, the proposed framework involves four distinct phases. Firstly, the target system is examined to identify its conceptual model and the specific activities to be mined. Secondly, a search task is designed and carried out with representative users, during which their interactions with the system are recorded. Third, the mined interactions are analyzed to identify recurring patterns and behaviors that can be used to infer the mental models of the users. Last, the conceptual model is compared with the inferred mental model of users to identify mismatches. The ultimate objective of this process is to improve the design and development of interfaces, ensuring that they become increasingly user-friendly. Regarding the mental model elicitation methods outlined in Section 3.2.3, this proposal aligns with the category of observational methods used along with a think-aloud protocol.

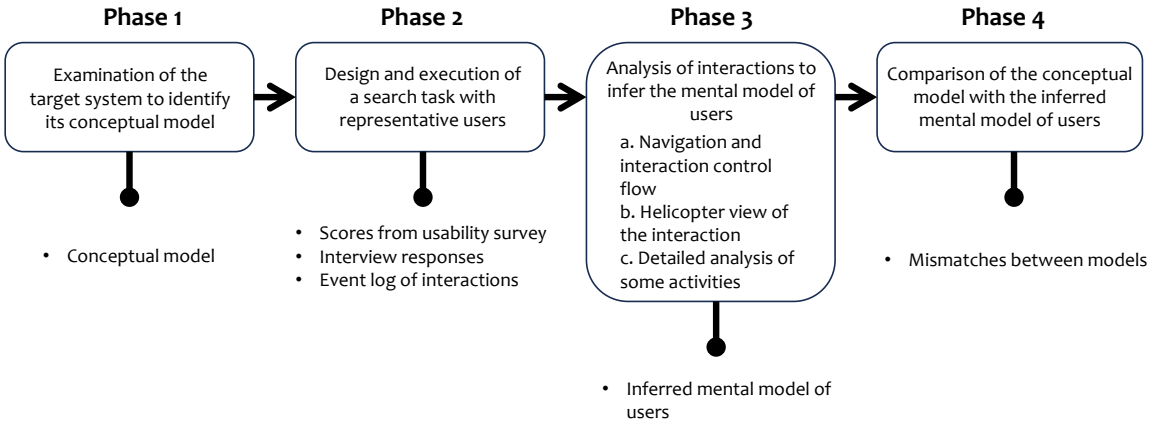


Fig. 3.1 Methodological framework at a glance.

3.3.1 Examination of the target system to identify its conceptual model

Understanding the conceptual design of a geospatial search engine involves both interviewing members of the product team and the development team, as well as examining the technical documentation (feasibility study, analysis guide, design guide, construction guide and implementation guide) and user interface of the geospatial search engine.

The conceptual model of the geospatial search engine is usually defined as a mixed model, integrating a search engine and a geographic information system (GIS), as depicted in Figure 3.2. This conceptual model draws upon the metaphorical notion of mental models, comparing the system

to archetypal systems. This GIS influence enables users to gain a comprehensive understanding of the geographic context of the search results. However, designing such a system goes beyond simply adding features from each archetype. It is essential to provide users with the flexibility of a search engine in exploring search results while also delivering a comprehensive geographic context. The integration of the two systems must be seamless and intuitive to ensure a meaningful and efficient search experience for users.

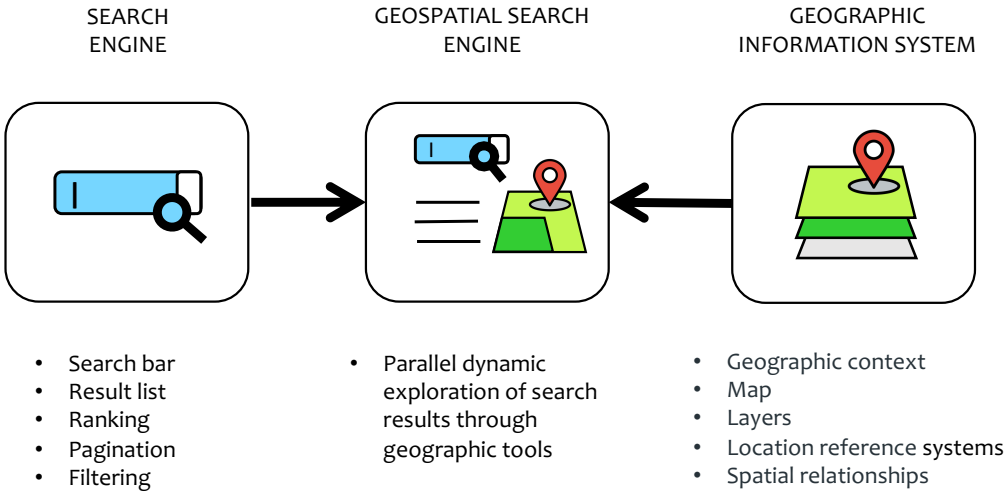


Fig. 3.2 Conceptual model of the geospatial search engine.

3.3.2 Design and execution of a search task with representative users

As presented in Figure 3.3, the sessions involve a sequence of pre-test, test, and post-test activities. To begin, all participants sign an informed consent form and completed a pretest questionnaire to confirm their assigned category. The informed consent encompasses essential elements, including the research objectives and methodologies, the process for withdrawal, and the confidentiality measures in place to safeguard data and information provided to the research team.

Participants are recruited and evenly divided into three categories:

I. Novice users: This category includes people with no academic background or relevant professional experience in geography or related disciplines and who are not regular users of the geographic information platforms of the institute.

II. Expert unfamiliar users: This category includes people with relevant academic training or professional experience in geography or related disciplines who are not regular users of the geographic information platforms of the institute.

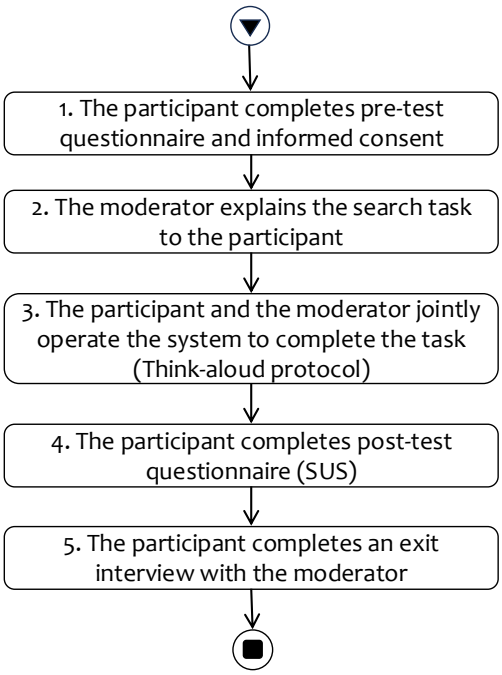


Fig. 3.3 Flowchart representing the test workflow.

III. Expert familiar users: This category includes people with relevant academic training or professional experience in geography or related disciplines who are regular users of the geographic information platforms of the institute.

The categorization is driven by the hypothesis that domain expertise and familiarity with specific conventions, such as those embedded in the platforms of a particular data publisher, can influence behavior and mental models. The recruitment of participants is carried out by means of a referral sampling method [132].

The experiments must adhere rigorously to the Ethics Appraisal Procedure established by the European Union Horizon 2020 Program.[1] This adherence encompasses crucial components, including precise procedures to discern and enlist appropriate research participants and an informed consent protocol that ensures complete disclosure to all human subjects regarding their participation and the subsequent management of their data.

The sessions finish with the System Usability Scale (SUS) questionnaire and a brief unstructured exit interview to capture the opinion that users have of the interface. The SUS questionnaire is widely recognized in the usability field [97]. It consists of 10 items, each assessed on a 5-point Likert scale. This evaluation yields a final score between 0 and 100.

[1] https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/ethics_en.htm

3.3.3 Analysis of interactions to infer mental models of users

After the completion of the sessions, an event log is created with three attributes: case (session identifier), activity, and timestamp. The event log is stored in a comma-separated value file. Once the interactions are mined, we propose a combination of quantitative methods for analysis.

Standard process mining techniques, descriptive statistics, and inferential statistics are used to explore the data. The main objective of the analysis is to visually represent the interaction process observed in the usability tests. For this purpose, we follow a typical process discovery workflow [103], progressively incorporating various perspectives to form an integrated understanding of the observed interaction. The initial step is the analysis of the process from a **control flow perspective**, which is usually the main focus of process discovery [103]. To describe the transitions between browser pages and the execution of activities, a series of direct-follows graphs are generated. This allows us to comprehend how users navigate through the platform. Given that certain actions can be performed on multiple pages, our goal is to identify if there are particular pages where such actions might be preferred.

The control flow perspective is followed by a comprehensive overview of the interactions contained within the event log, which can be defined as a **helicopter view** of the process [103]. To achieve this dimension, we employ dotted and variant explorer charts. Dotted charts represent each event as a dot on a two-dimensional plane, with the horizontal and vertical axes representing the time and class of the event, respectively. On the other hand, a variant explorer provides a visual representation of all the distinct paths taken by a specific process across observed cases. We then briefly delve into the use of search terms and filters used by the participants in the study. In addition, given the nature of the geospatial search engine, we show a **detailed analysis of some activities** which includes querying, filtering and interacting with search results.

Throughout these previous stages, results segregated by each of the three user categories to identify potential differences among them are provided. Where applicable, the analysis shows the result of an ordinary least squares regression to evaluate the potential impact of user category on specific study variables (SUS scores, number of interactions by type of activity, session duration, and perceived relevance of results). Regression produces coefficients describing the linear relationship between quantitative independent variables and a dependent variable. Gender, age, and educational level are introduced as control variables in this assessment. Given that these variables are categorical in nature, they are transformed into a one-hot encoding representation. Novice users are used as the reference category for user groups. Whenever a significant difference is detected in relation to the other two groups, the corresponding statistical significance level is denoted using asterisks in the respective figure or table.

3.3.4 Comparison of the conceptual model with the inferred mental model of users

In this last phase, we compare the findings obtained regarding the conceptual model of the system — as identified through internal documentation and interviews with the development team — and

the mental models inferred from user interactions, behavioral patterns, and qualitative feedback. The conceptual model represents the intended design logic, functionality, and structure as envisioned by the creators, while the mental models of users reveal how they perceive and make sense of the system during real-world use. By comparing these two models, we assess the degree of alignment between design intentions and user understanding, uncovering areas of congruence as well as potential mismatches. These mismatches may manifest as incorrect assumptions, misunderstandings, or gaps in mental representations of how the system behaves or should behave.

Such discrepancies are particularly important because they often underlie usability issues, task errors, or inefficiencies in user interaction. For example, if users interpret a feature differently from how it was intended, they may fail to use it effectively, or worse, lose trust in the system. Therefore, this comparative analysis not only helps to diagnose problems, but also provides a foundation for targeted design interventions. These may include improving system transparency through better feedback and visibility, enhancing learnability by aligning interface metaphors with user expectations, or refining onboarding materials to bridge conceptual gaps. Ultimately, this reflective process informs iterative refinements to the platform that are grounded in both the theoretical architecture of the system and the lived experiences of its users, contributing to a more intuitive, user-centered, and robust design.

3.4 Experiments and Results

This section describes a case study applying our proposed method for the evaluation of a new geospatial search engine that has been developed by the National Geographic Institute of Spain.

3.4.1 Examination of the target system

As already indicated in Chapter 2, IGN has developed a new geospatial search engine, which provides access to a vast collection of two million geographic resources previously scattered across various platforms within the IGN [36]. One of the fundamental objectives of the new engine is to democratize the access to geographic information resources by appealing to a broader user base in contrast to its existing portals that are aimed at specialized audiences. Figure 3.4 illustrates the user interface, showcasing the three distinct navigation levels available within the search engine: a) “Quick search”, b) “Advanced search & results”, and c) “Metadata” pages.

Within each level, we have mapped out the user activities that can be performed. In the “Quick search” page, which acts as a gateway to the site, the central element is a text search bar. Additionally, users have access to a button that allows them to explore advanced geographic search options. These advanced options accept various inputs such as points, polygons, geometry files, coordinates, and cadastral references. Thematic category filters are also available to refine the search. Upon conducting a search on the “Quick search” page, users are directed to the “Advanced search & results” page, which serves as the core of the platform. The search options from the “Quick search” page remain accessible

at the top of the page. The center of the page presents a list of search results, with each result offering multiple interactive options such as viewing, downloading, purchasing, or locating the corresponding geographic resource. To further refine their search results, users can employ a list of faceted filters located on the left side of the page. On the right side of the page, a map is displayed, providing a visual representation of the search results. The users can interact with the map by adjusting the zoom level, panning, and adding various additional geographic information layers. When a result is selected, the corresponding perimeter is highlighted on the map. Upon selecting a specific resource from the list of results, the users are redirected to a "Metadata" page that offers more detailed information about the selected resource. This page also facilitates actions such as downloading, purchasing, or locating the corresponding resource. As it can be seen, the geospatial search engine provides some flexibility in the sense that certain activities can be executed on one or the other page of the search engine.

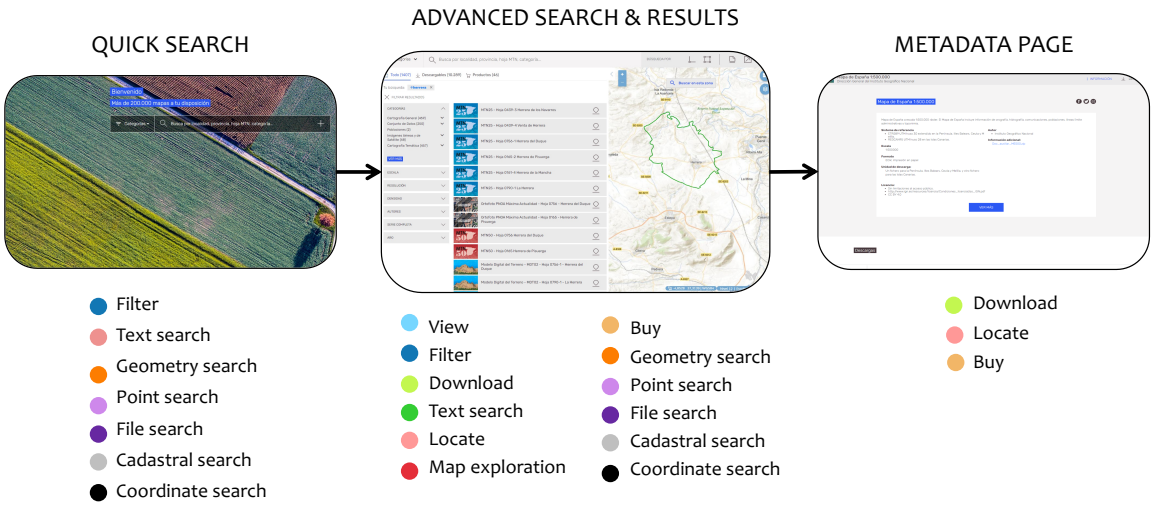


Fig. 3.4 Geospatial search engine interface.

The design model outlined in Section 3.3.1 is made explicit in the "Advanced search & results" page. The left part of the interface is influenced by the typical search engine archetype, characterized by two key attributes: a type-keywords-in-entry-form and view-results-in-a-vertical-list approach [121]. On the other hand, the right part of the interface is influenced by the typical GIS archetype, where the map is the dominant element, occupying a significant area of the interface (often more than 90% in pure GIS interfaces) [133].

3.4.2 Design and execution of a search task with representative users

The design and execution of the search task were described in the Section 2.4.3.3 on usability testing in the previous chapter. The evaluation focused on how novice and expert users interacted with a semantic geospatial search engine by completing a task that simulated a real-world scenario—planning a trip to

Sierra Nevada National Park. User feedback was collected through session recordings and moderator notes, highlighting three recurring areas for improvement: making the side map more interactive and better integrated with the search results; redesigning the filters and categories to improve their intuitiveness; and providing clearer guidance on how to use the retrieved geographic resources. Expert users offered more detailed and specific suggestions than novices. SUS scores fell within the marginal or barely acceptable range, with familiar expert users rating the system significantly lower than other groups.

3.4.3 Analysis of interactions

We have created a GitHub repository[2] with the Python notebooks for the analysis of results, both in terms of process mining and descriptive and inferential statistics. The logs were processed using PM4PY,[3] a process mining library written in Python. The process mining charts were created using PMTK,[4] a front-end solution built on top of PM4PY. Disco, a proprietary process mining toolkit under academic licence [134], was also used to generate the directly-follows graphs.

3.4.3.1 Navigation and interaction control flow in the geospatial search engine

In this section, we analyze the logs generated during sessions using directly-follows graphs [135]. Figure 3.5 describes the navigation through the search engine pages according to the type of participant. Each vertex denotes the three pages and each directed edge denotes transitions between them. Inside each vertex we show the average percentage of the total session time that the participants spent on each page. The thickness of the edges is proportional to the frequency of transitions that is shown next to the edges. Regardless of the group, the sessions started on the "Quick search" page and shortly moved to the "Advanced search & results" page, where most of the session time is spent iterating multiple times with the "Metadata" page.

In the directly-follows graph of Figure 3.6, in addition to the pages (green vertices), the actions (blue vertices) are added. The thickness of the edges and the color saturation of the vertices are proportional to a higher frequency of events. This graph has been filtered to show only the most frequent input and output edges. Often the directed follow graphs resulting from a process analysis have a spaghetti-like appearance that makes them difficult to interpret so vertex and edge pruning is common practice.

The graph allows us to derive a main navigation path that starts with the text search on the "Quick search" page and continues on the "Advanced search & results" page, where the user views and filters the items in the results list. The interactions with specific resources such as download, locate, and buy typically occur on the "Metadata" page for each resource. The search cycle is repeated between the "Advanced search & results" and "Metadata" pages without returning to "Quick search". Although the

[2]https://github.com/IAAA-Lab/Applicability-of-process-mining-in-usability-tests-ODECO-IGN

[3]https://pm4py.fit.fraunhofer.de

[4]https://pmtk.fit.fraunhofer.de

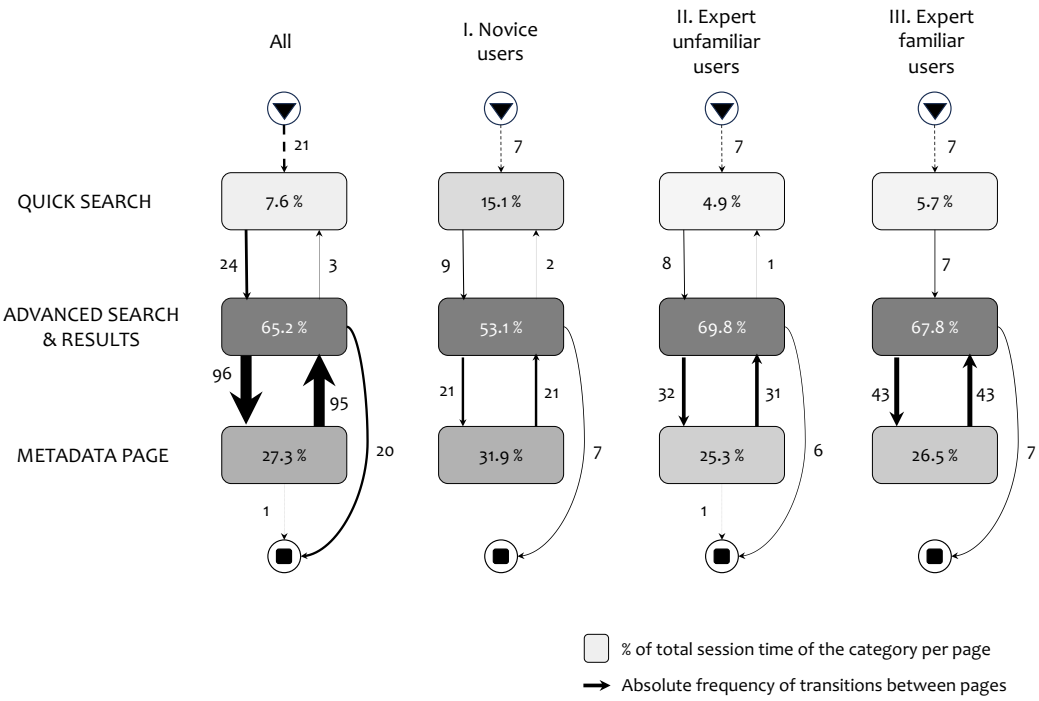


Fig. 3.5 Navigation process map.

figure only shows the graph for the total number of participants, it must be noted that the subgroups showed this same pattern of navigation.

3.4.3.2 Helicopter view of the interaction

Figure 3.7 shows a dotted chart representing the events corresponding to each session, where each line represents a session. The horizontal axis represents the duration of the session in elapsed minutes. The color of each dot corresponds to the 10 types of activity executed by the participant. To analyze the data, we initially grouped the traces by participant type and then sorted them in ascending order based on the total session duration. Our analysis revealed significant variations in session duration ($mean = 22'18''$, $median = 20'42''$, $sd = 10'18''$, $max = 44'06''$, $min = 7'42''$) and the number and sequence of interactions ($mean = 15.9$, $median = 16$, $sd = 7.3$, $max = 30$, $min = 6$). Our regression analysis detected significant differences between user groups. The novice users exhibited shorter session duration and performed fewer interactions compared to the other two categories, as evident from the graph.

Figure 3.8 presents the average number of interactions for each activity type, sorted in descending order based on the total average. The left section of the figure displays the values categorized by activity type and participant group. The right section of the figure showcases dot plots of the total number of interactions for each activity, organized in three parallel lanes representing the three participant groups: novice users in pink, expert unfamiliar users in light blue, and expert familiar users

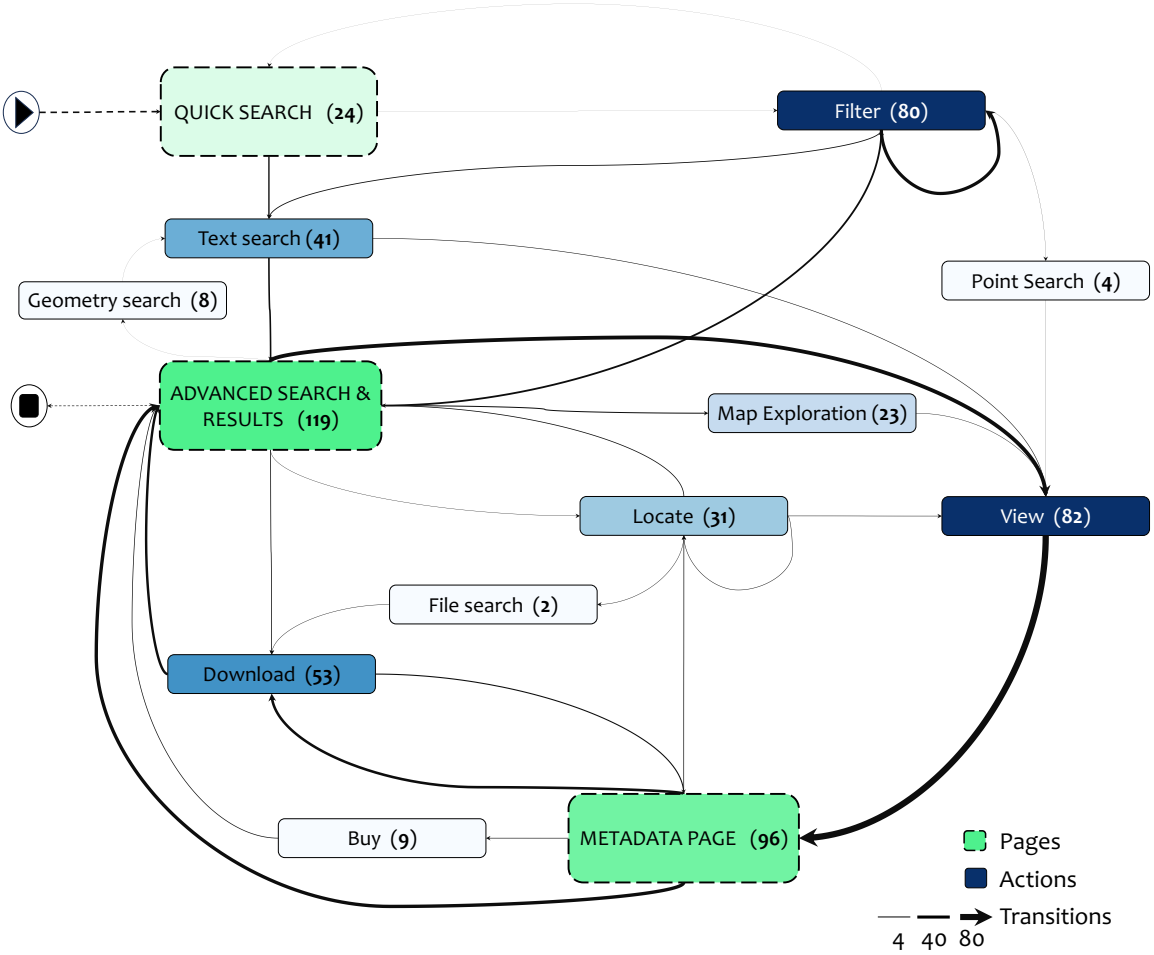


Fig. 3.6 Search process map including pages and actions for all participants.

in dark blue. The horizontal axis in this section represents the duration as a percentage of progress in the session.

A regression test did not detect significant differences between groups for any activity ($p < 0.05$), except for view where novice users saw fewer resources than expert familiar users. When considering the mean of the total number of interactions as a reference to profile the behavior of a representative user, we can observe certain patterns. The "view" and "filter" activities were the most frequently executed activities, occurring four times each. Following them, the "download" activity was performed three times, and text search was performed twice. Lastly, the "map exploration" and "locate" activities were typically executed once in the search process. On the other hand, activities such as "geometry search", "point search", and "file search" were not executed in a typical session. Additionally, neither the "cadastral search" nor the "coordinate search" activities were used by any user in this sample. The distribution of points on the dotted chart suggests that there may be variations in the timing of sessions where certain types of activities are more likely to be executed.

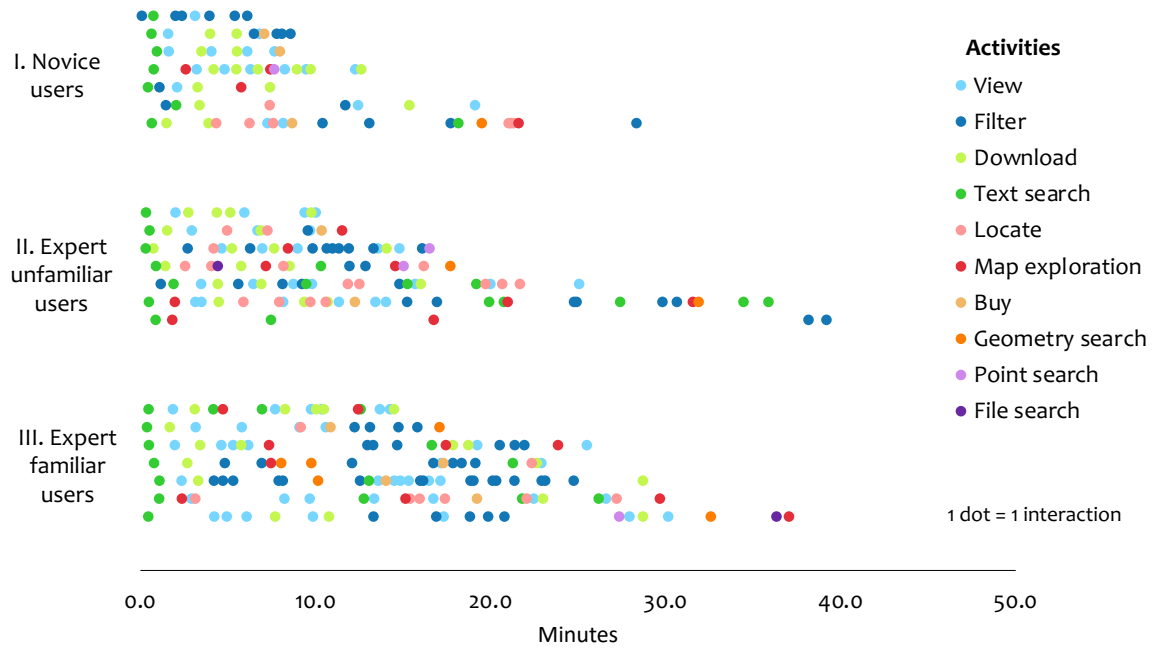


Fig. 3.7 Dotted chart distribution of the events over absolute time.

Activities	All	I. Novice users	II. Expert unfamiliar users	III. Expert familiar users	
	Average number of interactions				
View	4	3	4	6	**
Filter	4	2	4	5	
Download	3	2	3	3	
Text search	2	1	2	2	
Locate	1	1	2	1	
Map exploration	1	1	1	1	
Buy	0	0	0	1	
Geometry search	0	0	0	1	*
Point search	0	0	0	0	
File search	0	0	0	0	
Total	15.9	10.7	16.9	20.0	**
Duration (min)	22`18	15`21	24`14	27`20	**

*, **, *** significant at p ≤ 0.10, 0.05 and 0.01 respectively.

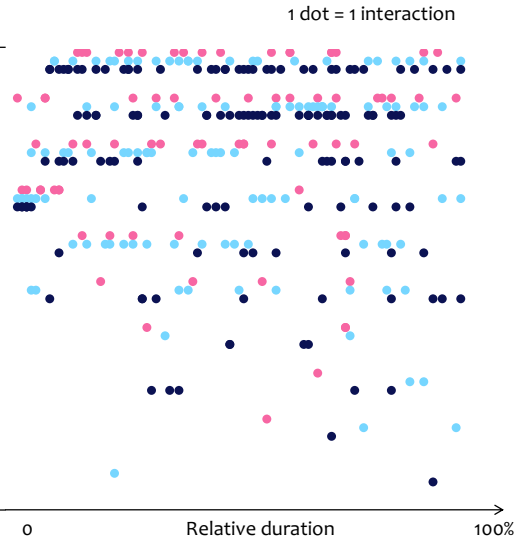


Fig. 3.8 Dotted chart distribution of the events over relative time.

To demonstrate the diversity and structure of the process, Figure 3.9 is presented. Whereas the left side of the figure showcases a variant explorer graph, the bottom right displays a relative timeline indicating the expected time of occurrence for the various activities. The variants are displayed in

descending order based on the number of interactions; the times of occurrence appear from earliest to latest.

The variant explorer graph reveals significant variability in the number and structure of interactions observed across the sample of sessions. There are no repeated sequences, and few generalizable patterns are evident. However, it is worth noting that most sessions initiated with a text search, sometimes followed by a filter. In the "Quick search" page, none of the users executed advanced search activities. Furthermore, several traces concluded with filter sequences (represented by light green frames), and these sequences were not consistently followed by resource exploration activities.

The timeline presented in Figure 3.9 offers valuable insights into the temporal distribution of activities across sessions, indicating that certain activities may be more prevalent at specific stages of the search process. Based on the timeline, it can be observed that the "text search" activity commonly took place at the beginning of the session and was typically the first activity performed. Following this, activities such as "view", "locate", and "download" were more frequently observed near the end of the first half of the session. On the other hand, activities like "map exploration", "filter", and the remaining activities tended to occur somewhat later, after the middle of the session.

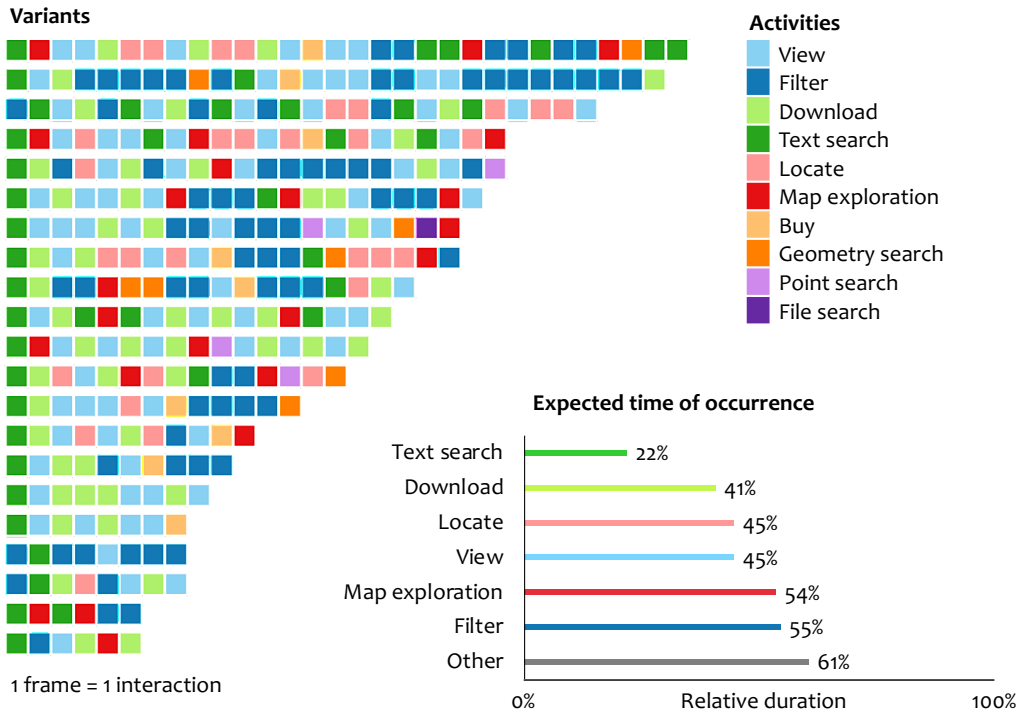


Fig. 3.9 Variants explorer sequence of events.

3.4.3.3 Detailed analysis of some activities

This section examines the particular parameters of the activities directly involved in the search process: query terms, filtering of results and an analysis of the hits.

With respect to the analysis of query terms, we analyzed the inputs from 41 textual user searches. When presented with the text box, all users opted for keyword-based searches. 83% of these searches explicitly mentioned the national park in question. Only one novice user, along with a couple of familiar expert users, attempted searches using place names in proximity to the park without mentioning it directly. No attempt was made to formulate queries in a natural language style.

Regarding the filtering of results, the search engine offers seven options: thematic category, year, scale, format, series, downloadable content, and products. Among the 80 interactions with these filters, 78% were directed towards the thematic category filter, with smaller proportions observed for the remaining filters. This consistent pattern of behavior was observed across all the three user categories. Regarding the thematic category filter, it is worth noting that several participants expressed difficulty in using it, citing the simultaneous display of dozens of options during their sessions.

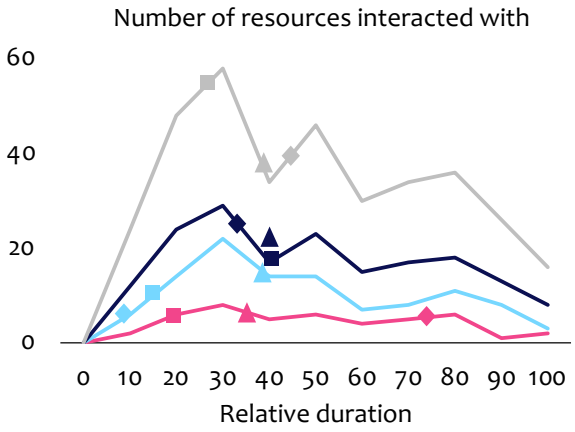
Finally, we analyzed search results. Considering the spectrum of activities encompassing "view", "locate", "download", and "buy", participants engaged with a diverse array of resources, summing up 46 distinct items, which include maps in various formats. It is important to note that there was no indication about time constraints for executing the task, nor an explicit definition of the list of the successful results that should be discovered.

Figure 3.10 offers a diagram with some indicators regarding the interaction of the three distinct interest groups with geographic resources. In the upper left section, timelines illustrate the frequency of resource interactions by participants and subgroups by corresponding time deciles. The expert user categories exhibit a peak in resource interaction early on, typically occurring within the initial 20-30% of the session progress. The novice users, on the other hand, display a more evenly distributed pattern of resource engagement throughout their sessions.

The top right section of the diagram presents metrics about the quantity and quality of search results. These metrics encompass the following key aspects: the number of resources per category, the average score to the question whether the results given by the search engine were useful and relevant (1- Total Disagree, 5- Total Agree), and the number of users who clicked on the first result of their search. The novice users engaged with a smaller number of resources and were less inclined to click on the first result compared to the familiar expert users. Despite this, the novice users perceived a significantly higher degree of relevance in the search results than familiar expert users did.

The wide-ranging nature of the results obtained makes it difficult to categorize them as either relevant or irrelevant with respect to the specific search scenario of planning a trip to a national park. The obtained items include: resources directly related to the national park; resources about municipalities, regions, or towns that intersect with the boundaries of the national park or are in close proximity to it; general national-level information, such as transportation routes, which may still hold relevance for trip planning; and resources that appear to have a remote connection to the search scenario, making their relevance to trip planning less evident.

The top three resources that gathered the highest attention were the following: "Sierra Nevada National Park routes", "Maps for mobiles Sierra Nevada National Park", and "Sierra Nevada National Park (8 maps + guide)". These resources are directly related to the search scenario and we have ranked



	All	I	II	III	
Total number of resources interacted with					
	176	45	62	69	
Users clicking on the first result					
	15	4	4	7	
Perceived relevance of results (1-5)					
	3,6	4,1	3,6	2,9	**

	Top 3 resources	Users who interacted with the resource			
▲	Sierra Nevada National Park routes	18	5	6	7
■	Maps for mobiles Sierra Nevada National Park	17	5	5	7
◆	Sierra Nevada National Park (8 maps + guide)	8	2	1	5

*, **, *** significant at p ≤ 0.10, 0.05 and 0.01 respectively.

Fig. 3.10 Interaction with resources found by participants.

them as the most relevant of the whole set. To provide a detailed breakdown of user interactions with these key resources, we have incorporated an icon associated with each of them. These icons serve to pinpoint the anticipated time of interaction with each resource within the session timelines. It must be noted that the interactions with these valuable resources occurred predominantly during the initial half of each session.

3.4.4 Model comparison

The primary inferences regarding mental models are predominantly at the metaphorical level. On the one hand, the conceptual model promotes a seamless integration of GIS and search engine paradigms. On the other hand, the mental model inferred from the users suggests a dominance of the search engine archetype over the GIS interface archetype. This aligns with findings observed in studies on information search systems [129–131] (see Section 3.2.4), reinforcing the prevailing tendency where the search engine archetype eclipses alternative metaphors for system representation.

The results indicate a gradient of alignment where expert users exhibit a mental model more akin to conceptual design, while novice users show a greater deviation. Although limited by the small sample size of the subgroups, the integration of process overview, control flow and resource interaction analysis allow us to trace similarities and differences between the groups in a comprehensive way. Among the similarities, we highlight the tendency of the navigation of the groups to reflect the stages

of the standard model of information searching the download and purchase activities. Although these activities can be performed before, they are actually registered on the "Metadata" page after the user has evaluated the result. Probably due to the focused nature of the task, dynamic searching behaviors (see Section 3.2.4) such as exploring toponyms found in previous searches were rare. With respect to differences, the most noticeable one is the tendency of the expert user groups to interact more frequently and intensively with the search engine. The same applies to the interaction with resources, where expert users browsed a greater number of items thereby more closely aligning with the hybrid model envisioned by the designers, although not with the ease initially expected. In contrast, novice users often exhibit linear search patterns and underutilized spatial tools, suggesting a more search-centric mental model. However, interaction with the most relevant resources reveals that most users, regardless of their group, managed to find material relevant to the search scenario.

The behavior of users, and in particular beginners, suggests a clear distinction between frequently used functions and infrequently used ones in a typical search scenario. This observation aligns with the Pareto principle [136], where a limited number of functions account for most user interactions. This finding has significant implications for UX design, as it allows the team to prioritize their efforts, allocate resources efficiently, and maximize impact. It is important to note that the infrequently used features should not be automatically discarded. Instead, they should be carefully examined to understand their potential contribution to the overall user experience. This is particularly relevant for geographic features that exhibited lower interaction frequencies compared to simpler search and exploration activities. In addition, these features were considered at a later phase of the session after users had tested simpler search strategies. This is where the main gap between the design model and the usage model can be seen: the parallel dynamic exploration of search results through geographic tools that the design team expected does not happen spontaneously. For instance, despite being a powerful tool that enhances the exploration of geographic resources and occupies a substantial area of the interface, the map received less attention than anticipated by the product team. This finding prompted designers to reconsider the role and presentation of the map in the overall search experience. Similarly, other underused functions such as file, cadastral, or coordinate search may need to be reevaluated and potentially removed from the initial design, if general purpose searches are to be prioritized.

Regarding the perception results of the SUS questionnaire and the assessment of the relevance of the results, a clear distinction is observed between the familiar expert users and the other two groups. The former provided significantly lower scores compared to the others. This differential evaluation could have been influenced by factors outside the test and interaction with the geospatial search engine, such as differences in self-perception of own competencies across varying levels of experience. This finding may serve as a trigger for working hypotheses in future research.

3.5 Discussion

Our discussion explores the benefits and challenges associated with integrating process mining tools into usability testing. In addition, some limitations of the study are described.

When looking at the purpose of usability testing and process mining, the two approaches appear to complement each other. The goal of usability testing is to provide direct feedback on how real users interact with a system, while process mining is designed to capture the actual behavior of a process through data analysis and visualization. The rich event-driven nature of usability testing seems well-suited for integration with process mining. For instance, the observed contrast between the results of process mining techniques and usability perception questions reflects this complementarity. The typical metrics reported in a usability test are limited in explaining how and why an interaction is performed [90], something that process mining can help to address. This broad perspective is beneficial when testing products where the user has great freedom of action, as it has been the case with the geospatial search engine.

In this context, it is essential to consider the challenges and adaptations required when incorporating process mining into usability testing—before, during, and after the test. One of the steps that requires more time and effort compared to a standard usability test is identifying the product to be tested (see Section 3.3.1). As our case study has evidenced, preparing for a process mining exercise involves more than just selecting a product and configuring it for testing. It is necessary to elucidate the assumptions regarding its use from a design perspective and map the potentially usability-relevant activities.

The mapping process involves the challenge of determining the appropriate level of granularity to match the interests of stakeholders aiming to improve the product. It is critical to keep in mind that mapping a set of activities is not definitive; rather, it prepares the researcher to be receptive to relevant activities that may arise during the usability test but were not initially considered. It is precisely this step of making explicit relevant design assumptions that enables this enriched approach to detect possible gaps between the mental model of designers and users.

During the execution of the test, an important consideration is to establish the most appropriate mechanism for recording events. In our case study, tests were video-recorded, and these videos were subsequently manually annotated by a researcher. While this approach offers the advantage of providing a detailed understanding of user interactions with the product, it is not optimal. It is significantly time-consuming and more error-prone due to its manual nature. The research of Dadashnia et al. [108] also highlights the practical challenges arising from the manual transcription of videos in identifying key activities performed by users.

While standard usability testing does not usually involve the collection of massive data sets [137], analysis of the results requires a research team with knowledge of data manipulation and familiarity with process mining techniques. Fortunately, the increasing availability of user-friendly open-source and proprietary data mining tools[5] simplifies this task. Among the tools employed in this study,

[5]https://processmining.org/

PM4PY and PMTK serve as examples of open-source resources, while Disco represents a proprietary alternative.

Our case study has several limitations that should be taken into account. Although the number of participants was suitable for usability testing, which aims to gain an overall understanding of user behavior, it may be insufficient for drawing statistically robust conclusions. This limitation is particularly important when detecting differences between novice and expert groups, as a larger sample size would be necessary to obtain sounder results. Usability studies often involve a relatively limited number of participants, primarily due to budgetary constraints. Nielsen [88] and Virzi [138] assert that identifying the majority of usability issues can be accomplished with as few as 5 users when a think-aloud approach is employed. Additionally, they argue that testing with a larger number of participants typically results in only marginal enhancements, assuming that the recruited participants adequately represent the intended user base. However, it is important to note that this topic remains a subject of ongoing discussion [139–141].

The potential impact of the moderator on participants is a crucial factor that should not be overlooked when interpreting usability test results [92]. The presence of a moderator might have resulted in longer execution times and more complex operations than they would typically perform, especially for expert users. Nevertheless, despite these potential disadvantages, the think-aloud method proved to be valuable for the researcher in understanding the reasoning behind the behavior of participants. This information was used later for modeling and interpreting the observed behavior. By verbalizing their thoughts, the users reveal their perception of the system and highlight what they do or do not observe.

A careful consideration must be given to the impact of the selected search task. Although it might align with the main intended usage, it may have a substantial impact on the observed outcomes. For instance, opting for a holiday tourism scenario inherently lacks an organic link to cadastral references. Consequently, it is not surprising that the observed frequency of searches using cadastral references is minimal.

3.6 Summary

Along the work presented in this chapter, we conducted an exploration of how process mining techniques can be used to uncover the mental models of users during the design phase of a geospatial search engine. These mental models serve the purpose of adjusting the system to align with user expectations and educating users about unfamiliar functionalities. To carry out our investigation, we employed a case study approach based on a newly developed geospatial search engine created by the National Geographic Institute of Spain. The case study involved recruiting representative users who participated in usability tests of the new portal. The sessions were recorded, processed, transformed into event logs, and subsequently analyzed using process mining techniques, as well as descriptive and inferential statistics. The findings from our study indicate that the observed mental model places a higher emphasis on the features typically associated with a standard search engine, rather than those

related to a geographic information system. This deviation in user behavior differs from the mixed model envisioned by the creators of the system. In direct response to the research question (RQ2) linked to this chapter, the results confirm that usability testing, when extended with process mining, provides a systematic way to identify user mental models, assess their alignment with the conceptual model, and derive insights that guide iterative improvements in interface design.

This study has identified several lines for future research. Firstly, although the current implementation of an automatic event logging tool did not fully accomplish its goal of supporting event log annotation, there is still significant potential to automatically record and model user behavior [79]. An important challenge lies in accurately mapping user clicks and keystrokes to activities that hold meaning from both a process and business perspective. Secondly, this work focused on bridging the gap between the conceptual model and the mental model of users solely from a process discovery perspective. However, there is an opportunity to deepen this understanding from a conformance-checking standpoint with appropriate metrics such as fitness [142]. Lastly, personalized and real-time support mechanisms can be developed to enhance the experience of users with inadequate mental models, based on their usage patterns of the platform.

Finally, in line with the emergence and popularization of new user interfaces, we hope that our work will not only help to promote further experimentation with process mining techniques by usability testing practitioners, but also stimulate discussion on the development of quantitative and qualitative tools for the analysis and visualization of mental models in human-technology interaction.

CHAPTER 4

IDENTIFICATION OF COLLECTIVE INTELLIGENCE IN OPEN DATA ECOSYSTEMS

4.1 Introduction

Open data ecosystems strive to be not only user-centered, but also circular and inclusive, fostering participation from a wide range of stakeholders. In such ecosystems, end users are no longer passive consumers of data; they also play an active role as contributors, helping to expand and refine the shared data commons. This dual role is particularly important for enabling collective intelligence, which emerges from the collaboration, coordination, and problem-solving efforts of distributed contributors. Understanding how this collective intelligence forms and functions is useful for evaluating the effectiveness and sustainability of open data platforms. The objective of this chapter is to propose a methodology for analyzing collective intelligence within open data ecosystems by examining patterns of user behavior and interaction. As a case study, we focus once again on an ecosystem centered around geographic information; this time shifting our attention to OpenStreetMap (OSM) and the Humanitarian OpenStreetMap Team (HOT) Tasking Manager, a platform built on top of OSM that organizes and coordinates large-scale volunteer mapping efforts for disaster response.

Intelligence is not an exclusive property of individuals. It also arises in groups of individuals such as families, nations, companies or other human, non-human and hybrid conglomerates. Numerous proposals have been developed in recent decades to define the concept of collective intelligence [143]. In general, these definitions converge on the idea of an emerging capacity within groups to solve problems, make decisions or achieve results more effectively than individuals working alone. The concept of collective intelligence finds applications in various fields, such as biology, sociology, political science and economics [143]. However, its influence is particularly noticeable in computer science, especially in the fields of human-computer interaction and computer-assisted cooperative work [143, 144].

The popularization of the Internet, and more recently artificial intelligence, has taken the idea of collective intelligence to a new level where interconnected groups of people and computers collectively do intelligent things in multiple domains. This is the case of the production model known as

crowdsourcing, in which the work traditionally performed by a designated agent is outsourced to a large, undefined group of people, usually in the form of an open call [145]. Examples of crowdsourcing include Wikipedia and open-source software development, where the collective contributions of numerous individuals result in robust, high-quality products [146]. This research focuses on crowdsourced geographic information, specifically the phenomenon of Volunteered Geographical Information (VGI). VGI describes the efforts of individuals and communities to address digital geographic information gaps, often arising due to the absence of comparable commercial platforms [147].

Prior to examining the specific case, it is important to briefly describe the main characteristics of the open data ecosystem being analyzed: humanitarian mapping initiatives in OSM. In recent years, humanitarian mapping missions have become one of the major focus areas of VGI initiatives, including the OSM community [148]. This type of mapping is transforming the disaster response landscape in situations where conventional geographical data sources are inaccessible or outdated [147, 149, 150]. Humanitarian mapping differs substantially from other forms of voluntary mapping in aspects such as its purpose, geographic areas of impact, mapping interfaces, editing and validation dynamics, and resulting footprint [148]. For instance, while overall mapping in OSM tends to be heavily concentrated in regions with a very high Human Development Index, humanitarian mapping shows a distinct pattern, primarily targeting regions with medium and low human development. Furthermore, the study of the effects that micro-tasking introduces into the dynamics of OSM peer production is a promising field for research [151]. Despite the growing role of humanitarian mapping in disaster response and its particularities, the mechanisms underlying its success remain poorly understood. Questions persist about how these projects harness collective intelligence to achieve their goals, and what lessons they offer for broader applications in peer-production systems.

This chapter analyzes humanitarian mapping projects as a collective intelligence system. To achieve this, we drew on two key approaches. First, we used the collective intelligence framework proposed by Malone et al. [143] to formulate relevant research questions that shed light on the dynamics of the humanitarian mapping system. Second, we leveraged the extensive dataset provided by HOT[1], the leading humanitarian VGI initiative operating on the OSM platform, to gather evidence to answer these research questions. More specifically, these data relate to the technological tool used for the coordination of humanitarian mapping, called the HOT Tasking Manager (HOT-TM).[2]

Malone et al. [143] argue that new information technologies have fostered novel forms of collective intelligence, leading to the emergence of a distinct field of study. To address this evolution, they propose a definition of collective intelligence that encompasses these new realities while remaining consistent with previous interpretations of the concept. In their revised definition, collective intelligence is characterized by three key elements: "(1) groups of individuals (2) acting collectively (3) in ways that seem intelligent". Each of these three components is described below along with the research questions derived from them.

[1]https://www.hotosm.org/
[2]https://tasks.hotosm.org/

From the perspective of new collective intelligence systems, a **group of individuals** can consist of both human and computational agents [152, 153]. While groups consisting of a single person and a computer are at the periphery of collective intelligence, the core lies in the combination of multiple individuals and computer systems working together to collectively address problems.

Decades of research on human groups reveal that individual characteristics significantly shape group dynamics, influencing both team performance and outcomes [154]. Effectively managing group composition, the configuration of the attributes of its members, involves identifying and prioritizing key attributes. When a particular attribute becomes salient, it can be a determining factor in the structure and interactions of the group [155].

According to Jiao et al. [156], the success of crowdsourcing is closely related to individual attributes. In this field, the experience of the group members is often the most highly regarded attribute. Evidence consistently shows that members with knowledge and experience in a specific domain produce higher quality results than non-experts on a wide range of tasks [157]. This evidence reinforces the significant emphasis placed on estimating, identifying, and managing the expertise of contributors to enhance task allocation in crowdsourcing settings [158–160].

As far as the non-human component of the groups is concerned, Malone [152] recognizes that computer agents can participate in a group in various ways, depending on their relationship with human agents. The most common role is that of a tool in which computers serve to enhance human capabilities. Like other tools, a computer requires direct instructions from a human to perform tasks. A step further is the role of assistant. Unlike a tool, an assistant has more autonomy and can take the initiative to help humans achieve their goals. Further up the hierarchy, computers can act as peers, demonstrating an autonomy comparable to that of human group members. Finally, computers can take on managerial roles. In this capacity, they perform tasks such as determining the sequence of tasks needed, predicting which contributor is best suited for each task, automatically assigning tasks to appropriate contributors, and evaluating their work.

Given the importance of group characterization in peer production environments, we formulate the following research question:

> RQ1: What characterizes the group of individuals participating in HOT-TM mapping projects?

By **acting collectively**, collective intelligence implies that the behavior of the group is characterized by the relationships between the activities of its members. This does not mean that all members share identical objectives or cooperate at all times but emphasizes the presence of interdependencies between their activities. According to Suran et al. [161], the processes of collective action can be analyzed and categorized based on two key dimensions: the type of activities (creating or deciding) and the type of interactions (independent or dependent). In creating activities, contributors produce something new (e.g. ideas or artifacts), while in deciding activities, contributors evaluate and select alternatives. To fulfil their missions, organizations usually need both creation and decision capabilities. Creation capabilities require decision capabilities to select the best outcomes, and decision capabilities

need creation capabilities to provide evaluation options. Since both activities can be carried out by individual actors as well as by groups, they can also be considered as dependent or independent interactions. The intersection of these two dimensions, types of activities and interactions, gives rise to four combinations:

1. Collection (Create + Independent): In these activities, individuals contribute independently, each offering unique inputs to the system based on their work.

2. Collaboration (Create + Dependent): These activities involve multiple individuals working to generate interrelated or interdependent solutions.

3. Individual Decision (Decide + Independent): Decisions are made by individuals acting independently, with outcomes that may vary from person to person. Sometimes these decisions may be influenced by information shared by others.

4. Group Decision (Decide + Dependent): Decisions are made collectively by a group, resulting in a consensus that affects the group as a whole. Important variants include voting, consensus, averaging, and prediction markets.

Understanding and describing these processes, along with identifying the dominant dimensions of collective action, are useful for effectively managing crowdsourcing efforts and designing more efficient workflows. McDonald [162] outlines an itinerary for monitoring a range of key aspects in crowdsourcing initiatives. This itinerary begins with examining the type, time and effort involved in micro-tasks and extends to analyzing the interaction, both positive and negative, between the actions of the involved actors. Classic concepts within the field of computer-assisted cooperative work such as articulation work, user roles and division of labour also contribute to enriching the debate on collective action [163, 164].

This context leads us to the following research question:

RQ2: How is collective action characterized in HOT-TM mapping projects?

By using the term **seem intelligent**, the definition acknowledges that what is considered intelligent can vary depending on the perspective of the observer [143]. Characterizing something as intelligent can be challenging due to the multifaceted nature of intelligence. We can detect intelligence in a system by identifying cognitive processes such as reasoning, consciousness, planning, abstraction, and learning. Alternatively, we can observe typical outcomes of intelligence, such as adaptive behavior, problem-solving, and artifact creation, which relate to goals and interactions with the environment [143]. According to Riedl et al. [165], collective intelligence has the potential to predict group performance across a wide range of tasks.

Collective intelligence is not by default a universal property of collaborative groups [166]. In this regard, it is important to highlight what differentiates intelligent action from mere collective action. While intelligence cannot always be directly or linearly related to success in task performance, the concept of collective intelligence suggests that groups, under certain conditions, can outperform individuals working alone. This idea is closely aligned with the notion of the “wisdom of crowds” [167],

an emergent property in which groups of individuals may be smarter than the smartest individuals within those groups. However, collective action alone does not guarantee the emergence of superior group performance, a highly desirable outcome in any crowdsourcing initiative. On the contrary, an ineffective organizational model for coordinating collective efforts can lead to the opposite effect, the "madness of mobs" [168]. This distinction underpins the following research question:

RQ3: What evidence of intelligent action can be identified in HOT-TM mapping projects?

In addressing these questions, our research aims to contribute to several key areas. First, it allows us to better understand the dynamics of participation in humanitarian mapping. As it will be discussed in section 4.2, while this topic experienced important developments in its early days, subsequent research has not kept pace with the major transformations in the field. This gap has resulted in a lag behind studies of other peer production environments, especially in terms of how contributors collaborate and interact effectively. Furthermore, this study helps to identify specific opportunities to improve the mapping process by focusing on ideas that strengthen, rather than disrupt, productive and sustainable collaborative dynamics.

The remainder of this chapter is organized as follows. Section 4.2 reviews the existing HCI research in the field of VGI, offering insights into humanitarian mapping efforts through the lens of collective intelligence. Section 4.3 introduces basic notions about HOT and HOT-TM. Subsequently, Section 4.4 explains the process followed to answer the research questions on how the concept of collective intelligence manifests itself in HOT-TM projects. Section 4.5 presents the development of the case study following the methodology proposed for that purpose. Section 4.6 focuses on reflecting on how the current collective intelligence arrangement contributes to or hinders the humanitarian mapping process. Finally, Section 4.7 offers concluding remarks, opportunities for improvement and potential future directions for research.

4.2 Related work

Over the years, the OSM community and its humanitarian applications have served as a rich basis for research on cooperative work practices. In this section, we aim to synthesize previous findings to inform our reflection and discussion on the three components of the working definition of collective intelligence adopted in this chapter. Some of these studies share the particularity of using the same data source as our study, the HOT-TM API.

4.2.1 Crowdsourced Work in OSM

We begin by exploring studies that provide insights into the characteristics of the groups of individuals involved in OSM. As reported by Anderson et al. [169] most efforts to study the social organization of the OSM community have relied predominantly on qualitative research methods, such as participant observation and interviews. These studies have offered insights into the demographics and motivations

of contributors. For instance, Choe et al. [170] highlight the critical role of group composition in shaping group dynamics in OSM, especially conflicts. Their study emphasizes the impact of differences that often arise during interactions between various sub-groups of mappers. These sub-groups are often distinguished by factors such as gender, geographical location, relationship between mappers and the areas they map, level of experience and professional affiliation.

Several studies employing more quantitative approaches have explored the role of factors such as mapper experience in shaping the dynamics of OSM. According to Yang et al. [171], a small proportion of contributors with extensive experience in geo-data editing, proficiency in professional software, and a high level of enthusiasm and concentration are responsible for the majority of contributions. This same research suggests that higher experience consistently produces high quality geographic data. Begin et al. [172] identified that most OSM contributors are part of an inactive majority who do very little editing, in contrast to a small group of prolific contributors who are highly active. According to this study, it typically takes several years, with 4.5 years being the norm, for a user to become an advanced OSM contributor. Urrea and Yoo [173] specifically examine the effect of experience on HOT projects. The results of analyzing 5,162 HOT projects show that the project completion rate improves, albeit at a decreasing rate, with the experience of contributing volunteers. Furthermore, the effect of experience on the project completion rate is influenced by the urgency of the project. In terms of retention, the results indicate that volunteers feel incentivized to return to an online volunteering platform more quickly when they are close to reaching a new rank based on experience. Dittus et al. [174] conducted an analysis of HOT-TM contributor retention using behavioral data from 1,570 first-time contributors in 99 projects. Their findings indicate that most first-time HOT contributors tend to work at a fairly steady pace, while contributors with previous OSM experience tend to work faster and remain active for longer. In addition, they found that more complex task requirements can demotivate first-time contributors, regardless of their previous OSM experience.

Another key attribute which has received a great deal of attention is the geographical location of the contributors. This factor raises important questions, such as the importance of local knowledge in mapping and the influence of mappers in certain regions on outcomes in other areas. Even the initial conceptualization of VGI reflects the singular attention attached to the local factor. According to Goodchild [147], “the most important value of VGI may lie in what it can tell about local activities in various geographic locations that go unnoticed by the world’s media, and about life at a local level”. This intuition appeals to a greater knowledge of the terrain and context on the part of nearby contributors. Locally-produced VGI tends to be associated with higher quality, richness, diversity and utility [175–177]. From the earliest academic studies on HOT, the role of local mappers has been emphasized. Soden and Palen [178] highlight the need to move from the initial showcase of crowdsourcing efforts following the 2010 Haiti earthquake (the first major humanitarian response by the OSM community) to the development of sustainable, locally-owned community mapping ecosystems in at-risk regions around the world. However, studies indicate that the goal of capturing local knowledge of the terrain often falls short in VGI initiatives, leading to a significant amount, or even the majority, of VGI content being non-local [179–181].

It is also noteworthy that there are quantitative and qualitative works characterizing the OSM group of contributors by developing archetypes that reflect broad patterns of behavior [182, 183]. Such is the case of Zhang et al. [183] who developed a quantitative approach to identify the emerging editor roles in OSM based on temporal behavior, change set type, feature diversity and geographical diversity. They identified two clusters particularly associated with humanitarian mappers through a clustering approach: humanitarian enrichers and humanitarian creators. Humanitarian creators show a notable inclination towards creating new mapping features in OSM, demonstrating a strong interest in humanitarian mapping efforts but showing a low propensity to revise existing mapping features to enrich or improve them. In contrast, humanitarian enrichers actively engage in humanitarian mapping efforts, focusing on enriching existing map features by adding detailed attribute information rather than creating new features. In general, humanitarian mappers have lower retention rates than the average among active mappers, with enrichers having higher retention than creators.

4.2.2 Collective Action in OSM

We now turn to studies that explore collective action within VGI initiatives. Mapping, by its very nature, is a creation activity, in which contributors enrich the collective map by adding elements such as nodes, pathways and relationships. These contributions are made through various map editing interfaces. Humanitarian mapping shows a distinct mapping footprint compared to general OSM activities [148]. It is characterized by being more intensive in mapping buildings than in mapping roads. In addition, edits in humanitarian mapping are more often focused on creating new elements, with less emphasis on modifying or updating elements. This divergence may reflect the unique priorities and objectives of humanitarian mapping efforts, which often aim to rapidly generate essential geographic data for disaster response and relief operations.

Mooney and Corcoran [184] sought to determine whether this creation in OSM involves genuine collaboration or whether it consists primarily of individual tasks performed in isolation with minimal interaction between contributors. Using a case study that examined the entire mapping history of London in OSM between 2005 and 2011, the authors used object co-editing as an indicator of collaboration between mappers. Their findings revealed that collaboration between contributors was limited, suggesting that much mapping activity occurs independently rather than through coordinated efforts. Building on their earlier research, Mooney and Corcoran [185] subsequently extended their study of co-editing networks to include seven major cities, focusing specifically on the activity of very frequent contributors. The results were mixed. On the one hand, these contributors were found to do a great deal of mapping work independently, presumably related to objects that did not attract sufficient interest from other mappers to collaborate in their editing or development. On the other hand, frequent contributors also collaborated with the less frequent contributors by editing or updating their work.

The study of socio-behavioral phenomena in OSM lags behind research in other peer-production environments, such as Wikipedia, where such phenomena have been extensively analyzed [186, 169]. The challenges posed by the manipulation of OSM data [169] and the difficulty of tracking the

influence of OSM discussion channels on the immediate mapping process [186] are pointed to as possible explanations. Kogan et al. [186] delved into the analysis of user interactions in OSM, while critically addressing the limitations of defining collaboration solely through the co-editing of objects. These authors argue that this narrow definition does not capture the nuanced and multifaceted nature of collaboration as understood in the fields of CSCW and HCI. In their investigation of the dynamics of collaboration during the 2010 Haiti earthquake, they conducted a two-phase study. In the first phase, they employed network analysis techniques to identify high-value segments within the extensive OSM dataset for a more focused qualitative examination. In the second phase, they conducted a detailed content analysis of selected mapping activity and interviewed participants. This approach allowed the researchers to uncover detailed mapping practices that reveal variations in temporal, spatial and interpersonal interactions. The study also recognized the influence of HOT-TM, introduced after the Haiti earthquake, for future research examining the nature of interactions within the OSM community.

4.2.3 Intelligent Action in OSM

Research has also been carried out to identify the presence of intelligent action within VGI initiatives, with the aim of distinguishing it from mere collective action. Spielman [187] reviewed the existing literature and evidence to assess whether OSM fosters the conditions necessary for the emergence of collective intelligence. The review begins by highlighting the need for map quality metrics and then identifies a recurring tension in relevant literature between two key quality perspectives: credibility and accuracy [188–191]. Spielman concluded that the system has mixed conditions. On the one hand, low barriers to participation enable widespread contributions but also make it difficult to assess their credibility. This creates the risk that high quality or highly credible data will be degraded when combined with lower quality contributions, which could hinder the development of an intelligent collective outcome. To mitigate this, the OSM community has implemented data validation strategies that rely on the authority of contributors or the sequence in which contributions are made. However, reliance on user authority in crowd-based systems introduces a major drawback: it can become self-reinforcing. This dynamic can lead to overemphasizing the work of certain contributors and undervaluing or marginalizing others, potentially limiting the diversity and adaptability of the collective intelligence process. Yin et al. [151] used HOT-TM data to evaluate the effects of a micro-tasking intervention on contribution dynamics. According to their research, micro-tasking can effectively increase both the number of contributors and the contribution rates. However, it may also deepen the concentration patterns commonly observed in such settings, where a small group of contributors accounts for the majority of contributions and efforts. They argue that in HOT-TM, as well as in other peer-to-peer production environments such as Wikipedia [192], there is a trade-off between productivity and equity.

Finally, it is worth mentioning emergent research on the effects of artificial intelligence on VGI activities. AI has a transversal impact on the collective intelligence system, as artificial agents, in their various roles, could influence the composition of the group of contributors, modify the organization of work and may ultimately affect the quality of the final mapping products. In order to assess the

impact of AI on volunteer productivity, Tipnis et al. [193] studied the effect of introducing Rapid[3] (an OSM editor designed to add buildings and roads that were identified by AI from aerial imagery) into HOT-TM. Using a difference-in-differences event study design, they found that volunteers who were introduced to AI-powered mapping reduced their weekly contributions to the platform by an average of 8.1% compared to those who were not introduced. Moreover, this effect varied depending on the volunteer experience. While the negative impact on productivity persisted for the least experienced volunteers, those with the highest levels of experience increased their contributions after being introduced to Rapid.

A critical reflection on previous research reveals a significant opportunity to advance studies on collective intelligence in the context of humanitarian mapping, an opportunity that this study aims to seize. On the one hand, the foundational efforts of OSM and humanitarian mapping attracted a great deal of scholarly interest. However, since that initial research there have been a number of noteworthy developments, including the introduction of tools such as HOT-TM and broader societal changes, such as the ongoing impact of the 2020 pandemic on perceptions of remote and distributed work. The continuation of previous research is clearly warranted.

In addition, previous research has highlighted a lag in the study of OSM compared to other peer production environments, particularly in terms of insights into contributor interactions. There is a notable appetite for research that delves into how contributors collaborate and interact to complete mapping tasks but few studies have taken full advantage of the rich data provided by tools such as HOT-TM. Those that have explored these processes have done so only superficially, leaving significant gaps in understanding the dynamics of micro-tasks and the underlying factors that influence their success. This study addresses these gaps by incorporating a collective intelligence framework, which offers a holistic perspective on the humanitarian mapping system, by examining interactions at the micro-task level and identifying factors associated with the successful completion of micro-tasks.

Looking ahead, the buoyant interest in AI and its applications in a variety of fields underscores the importance of improving our understanding of collaborative mapping. The findings of this study can serve as a basis for thoughtful integration of AI tools into humanitarian mapping workflows, ensuring that they enhance rather than disrupt productive and sustainable collaborative dynamics.

4.3 Background

4.3.1 Understanding the Context of Humanitarian Mapping and HOT-TM

As a predominantly quantitative study, a variety of qualitative methods were employed at the outset of the project to better understand the context of humanitarian mapping. These included in-depth interviews with field experts, participant observation and inspection of the HOT-TM user interface. Although the information collected during this phase was not analyzed in depth, it served as a valuable complement to designing, interpreting, validating and discussing quantitative research. We consider it

[3]https://rapideditor.org/

constructive to highlight this aspect of the research process, as access to rich data sources, such as those provided by the HOT-TM API, could lead to assume that analysis can be carried out without a nuanced understanding of the context. However, this omission could limit the potential depth and value of the analysis.

The semi-structured interviews were conducted in two rounds. The first round focused on open-ended questions on mapping objectives, data use, challenges, bottlenecks in the mapping process, communication and collaboration practices during mapping and validation. In this round, we interviewed separately a support specialist from the HOT Hub team for Latin America and the Caribbean (HOT-LAC) and a logistical advisor from the GIS team at Médecins Sans Frontières (MSF) for just over an hour each. In the second round, we analyzed the data from the tasking process in HOT-TM, generated preliminary results, and presented these findings to the same interviewees from the first round. For the presentation to the HOT-LAC team, three additional members joined the discussion: a manager, a data quality associate and a technical product owner. The research team annotated the presentation slides to gather feedback on the interpretation of the specific results shown in tables and figures. Interviewees were then asked to identify any inconsistencies in the results, highlight areas for improvement and share their assessments of the quantity, quality and possible improvements of the collaboration during the mapping and validation. In both rounds, one researcher acted as interviewer while the other researcher recorded the responses. These notes were periodically reviewed by the research team to identify supporting arguments and ideas that could enrich the discussion and interpretation of quantitative results.

Interviewees agreed that validators are a scarce resource in projects. They also highlighted the coexistence of two approaches to humanitarian mapping. The first approach prioritizes speed and agility to provide basic geographic information to emergency beneficiaries as quickly as possible. The second approach focuses on more elaborate objectives, such as emergency preparedness, prevention, and forecasting, and emphasizes higher-quality and more detailed information, such as building types, water bodies, and health facilities.

For the participant observation, three members of our research team attended an online Mapathon organized by HOT. One researcher, with experience in humanitarian mapping missions and OSM, took part in the task validation, which is only allowed to expert contributors. The other two researchers, who were new to the process, first received a half-hour introduction to HOT-TM before starting basic mapping operations. The activity lasted two hours in total.

4.3.2 Understanding the HOT Tasking Manager

In this section, we introduce basic notions to understanding HOT and the HOT Tasking Manager (HOT-TM), the main technological tool for coordinating humanitarian mapping. This description focuses particularly on identifying connections with categories of collective action presented in the introduction, which will serve as a basis for the development of later sections. Figure 4.1 shows a typical project page in HOT-TM which contains a map with the different tasks and their current states.

As can be seen, the project area is divided into tasks following a grid. The process followed by the mapping tasks can be divided into the Mapping and Validation phases [194]. As the tasks go through these two phases, they pass through various states registered by the HOT-TM system (see Figure 4.2).

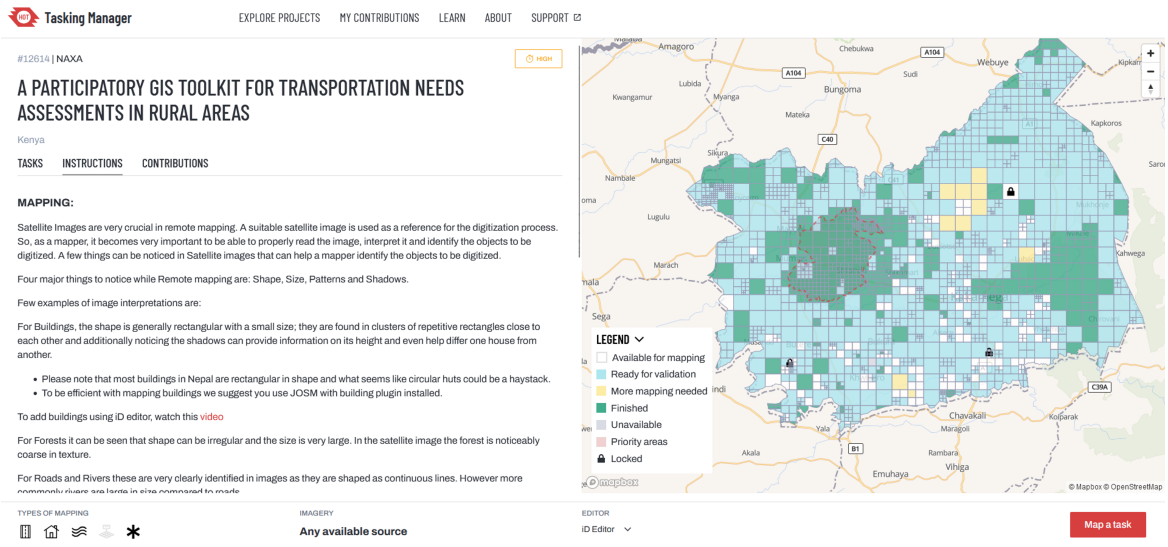


Fig. 4.1 Project page in HOT-TM: the left section displays information about the project, including a list of task states, instructions, and contributor metrics. On the right, a map visually represents the task states. Contributors can select a random task to map using the button at the bottom right, or they can manually select it. Once a task is selected, contributors can choose between mapping or validate modes. Source: https://tasks.hotosm.org/projects/12614/tasks

In the mapping phase, contributors select a task or receive a random one. A built-in editor opens in HOT-TM (although contributors can also choose to use a desktop editor). In this editor, contributors must draw missing elements as specified in the project description, such as buildings or roads, making sure not to draw outside the marked area. Note that they correspond to creation activities. Then, contributors decide individually whether the task is fully mapped, not fully mapped, in need of splitting, or its imagery is poor. Splitting a task divides it into four smaller tasks. This step can be repeated several times. Since the contributions of successive mappers are directly influenced by the work of previous mappers on the same task, this implies that mapping activities allow for some degree of asynchronous collaboration. The characteristic states of this phase registered by HOT-TM are as follows:

- **LOCKED FOR MAPPING:** Tasks currently being edited by mappers are locked to prevent overlapping editions. The system records the duration of this state. In the end, the mapper decides if the area is completely mapped, or if more mapping is needed.
- **AUTO-UNLOCKED FOR MAPPING:** Tasks locked for mapping are automatically unlocked after a timeout period of 2 hours. The mapper may have done some editing during the period which would still be preserved.

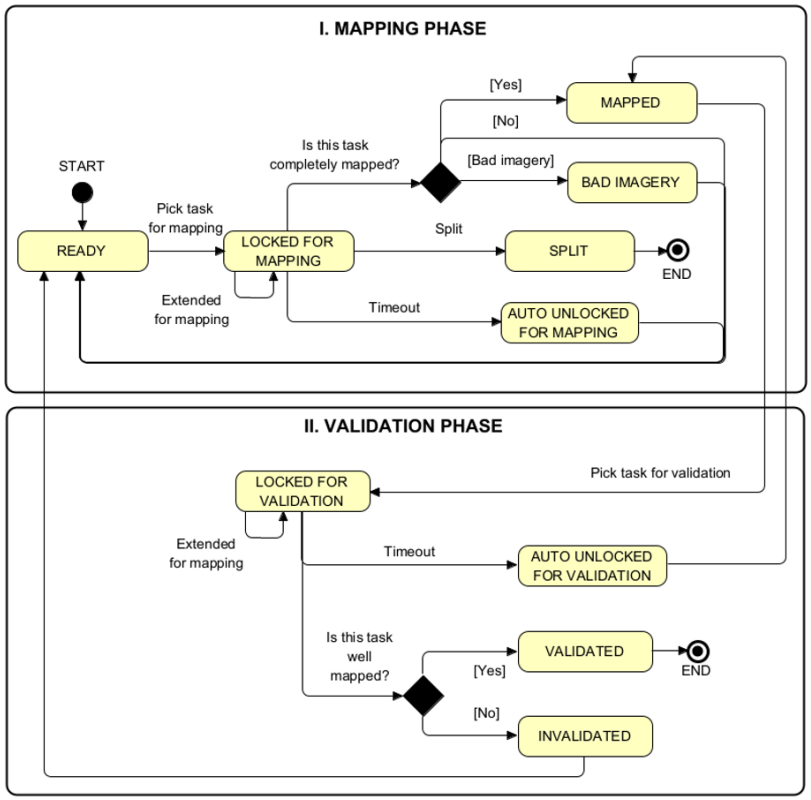


Fig. 4.2 Task state diagram showing the mapping workflow in HOT-TM. [194]

- **MAPPED:** Once the mapping job is complete as judged by the last mapper, then the task becomes complete and is ready for validation in the next phase.
- **SPLIT:** The mapper can split the task into smaller parts by assigning this state when the task is dense or complex. When a task is split, four new tasks are created.
- **BAD IMAGERY:** This state is acquired when, in the opinion of a contributor, the image of the task is not of sufficient quality for proper mapping.

In the validation phase, contributors review the mapped tasks. An editor similar to the mapping phase opens offering multiple options. The contributor then decides individually whether the task is properly mapped and moves it to the validated state or, if not complete, to the invalidated state. Those tasks that are invalidated return to the mapping phase for subsequent work. The characteristic states of this phase registered by HOT-TM are as follows:

- **LOCKED FOR VALIDATION:** Mapped tasks acquire this state when a contributor selects them for validation, preventing concurrent validation. Note that the validator has the ability to self-correct the task. Therefore, the validation phase may also include creation activities. The system records the duration of this state.

- **AUTO-UNLOCKED FOR VALIDATION:** Tasks locked for validation are automatically unlocked after a timeout period of 2 hours. The validator may have done some work during the period it was locked but did not ultimately record its validation decision.
- **VALIDATED:** Tasks correctly mapped according to the mapping instructions are marked as validated.
- **INVALIDATED:** Tasks incorrectly mapped according to the mapping instructions are marked as invalidated and return to the mapping phase.

There is an additional state that can appear in both the mapping and validation phases called EXTENDED FOR MAPPING. After two hours have elapsed, the contributor can extend this locked period to continue mapping or validating the task, avoiding the auto-unlock.

4.4 Methodology

To meet our objective, we proposed a two-stage methodology: (1) data collection, and (2) data analysis (see Figure 4.3). In addition, we have created a code repository[4] containing Python and R notebooks that provide a detailed description of the procedures used for data collection, preprocessing and analysis in this study.

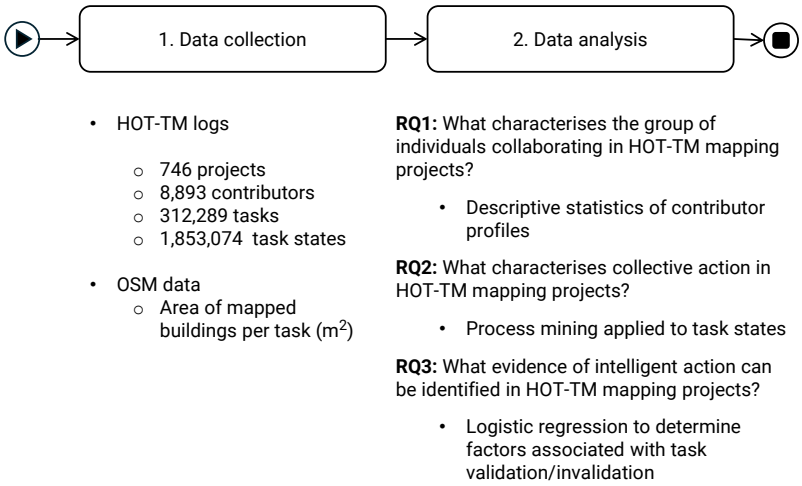


Fig. 4.3 Methodology at a glance.

[4]https://github.com/IAAA-Lab/Collective-Intelligence-in-Humanitarian-Voluntary-Geographic-Information-ODECO-HOT-TM

4.4.1 Data Collection

This subsection describes the procedures used to collect the quantitative data used for the study. The HOT-TM API[5] provides access to numerous data on humanitarian mapping projects. The API operations are grouped using tags representing mapping-relevant concepts (e.g. *projects*). Each invoked operation returns items with associated data fields (e.g., *action*, *actionDate*, *actionBy*), which we then mapped to answer the research questions.

The *users* tag was revised to address the question of group composition. This tag includes an operation (*queries*) for retrieving attributes describing each OSM user, which was queried with those participating in each humanitarian project. Among its fields, *mappingLevel* and *projectsMapped* provide information on their mapping experience. Similarly, the *country* field provides information on the location of contributors. As explained in the related work section, these attributes contribute substantially to understanding group composition. In addition, secondary fields such as *photo*, *slack*, *facebook*, *linkedin*, and *twitter* were collected.

For the collective action question, we sought data describing the tasks executed in a project and the individuals responsible for those tasks, with the aim of analyzing types of activities and interactions. The *projects* tag was relevant to this purpose, particularly its *activities* operation, which retrieves the history of states through which tasks in a project pass. The key fields are *action*, which specifies the type of task state (see Section 4.3); *actionDate*, which records the date of the state change; and *actionBy*, which identifies the person responsible for the state change. From the same tag, we also used the *statistics* operation to retrieve detailed info for all projects. The extracted fields included *priority*, *difficulty*, *perc_validated* (the percentage of total tasks validated), *created* (the date the project was created), *total_mappers*, *organization* and *country*. Difficulty and priority are attributes that potentially influence all tasks in the same project and will, therefore, serve as recurring variables in subsequent analyzes.

Regarding the question of intelligent action, previous research underlines the importance of having map quality indicators to study the spatial collective intelligence of a VGI system [187]. Although the available API operations lack direct indicators of map quality, the *activities* operation of the *projects* tag provides basic task performance evidence through the task validation phase. In particular, the label field indicates the result of the task validation as *VALIDATED* or *INVALIDATED*. These results serve as indicators to assess whether the mapping process has been executed correctly.

The resulting dataset comprises 746 HOT-TM projects. We selected projects created between December 1st, 2021, and the collection date (December 1st, 2023) and, from those, all the complete and archived projects. In these projects, 38,893 contributors completed 312,289 mapping tasks. In addition, we collected data on the 1,853,074 states that the mapping tasks went through in the projects.

For this exercise, we mainly relied on the data provided by the HOT-TM API to maintain manageability. However, we considered it necessary to include a basic control indicator associated with the workload of the tasks. To this end, we complemented the analysis with OSM data to calculate

[5]https://tasks.hotosm.org/api-docs

the observed building area for each mapped task. Task grids were used as the main input for the Bunting Labs API[6] to retrieve buildings for each mapped task. Before calculating their areas, both the task grid and building datasets were reprojected from geographical coordinates to the Universal Transverse Mercator (UTM) projection.

The data collected from the APIs were stored in files for further processing. The primary goal of data pre-processing was to prepare an event log, enabling the application of the relevant quantitative analysis techniques. For more details on the data collection process and pre-processing, the code repository includes a Python script that explains step-by-step how the HOT-TM and Bunting Labs APIs were accessed.

4.4.2 Data Analysis

This subsection outlines the analytical strategy used to answer the research questions. Before elaborating on the strategy, it is important to note that the analysis began with the development of a profile of the projects under study. This profiling was done to provide context on key project variables, such as difficulty and priority, which can significantly influence the execution of mapping tasks. These variables will be mentioned frequently in the subsequent analysis and presentation of results.

4.4.2.1 RQ1: What characterizes the group of individuals collaborating in HOT-TM mapping projects?

We developed a profile of the composition of the group based on two key attributes: the experience and location of the mappers, using the available data. For this initial profile, we consider as a group unit all 38,893 contributors who participated in any of the 746 mapping projects included in the dataset. Regarding expertise, HOT-TM uses a mapping level system to classify contributors into three categories: beginners (less than 250 OSM changesets), intermediate (250-499 changesets) and advanced (500 or more changesets).[7] Unlike other labels, this classification is available to all users of the system. We analyzed the frequency distribution of mapping levels and their association with the number of HOT-TM projects in which contributors participated. In addition, we use complementary fields (name, city, country, photo, Slack, Facebook, LinkedIn, Twitter) to examine the completeness of individual profiles according to their mapping level. Similarly, to profile contributors location, we used data on the country from which they reported and the countries in which humanitarian projects were implemented. This allowed us to describe the distribution of contributors according to the origin and destination of contributions. In addition, since HOT organizes its activities through regional hubs serving specific geographic areas, we also grouped contributors using these broader geographic units. It is important to note that location-based frequency calculations were only performed for

[6]https://buntinglabs.com/solutions/openstreetmap-extracts
[7]https://github.com/hotosm/tasking-manager/blob/v4.7.4/backend/config.py

contributors who reported their country, which represents 29% of the total. Finally, we calculated the group composition of these two attributes according to the type of project.

4.4.2.2 RQ2: What characterizes collective action in HOT-TM mapping projects?

We conducted a quantitative characterization of collective action using the activity and interaction framework proposed by Suran et al. [161] and the crowdsourcing activity tracking approach outlined by McDonald [162]. In this analysis, it is important to distinguish between states that directly reflect creation activities (e.g., "Locked for mapping") and those that indicate decision-making activities (e.g., "Mapped", "Validated", "Invalidated").

Our findings are presented in three parts. The first part characterizes the mapping process by analyzing the control flow and duration of task states in a typical mapping task. To improve readability, details on the frequency and duration of mapping states are provided in the appendix. The second part explores work articulation and contributor roles throughout the mapping process. We analyzed how task states are distributed based on mapping levels and the geographic location of contributors, distinguishing between mapping and validation phases. Additionally, we examined how these patterns vary with project difficulty and priority. The final part focuses on interaction analysis, recognizing that while mapping tasks can be completed individually or collaboratively, key decisions are always made individually. We began by quantifying the number of mappers involved in task execution across different scenarios. Next, we assessed the frequency of collaboration during the mapping phase of a standard task. Finally, we examined how collaboration relates to the mapping level of contributors by identifying common combinations of collaborators through a variant explorer and applying the "handover of work" concept [103]. This concept posits that the more frequent an individual "x" performs an activity that is causally followed by an activity performed by "y", the stronger the relationship between "x" and "y" is.

To guide the analysis of this question, we took advantage of process mining techniques [103]. They use event logs to gain insights into processes from multiple perspectives. In this case, the logs consist of state changes in mapping tasks provided by the HOT-TM API. Specifically, we selected bupaR,[8] an open-source R package for business process data analysis. This choice was motivated by its extensive range of analytical capabilities and well-documented functionality.

4.4.2.3 RQ3: What evidence of intelligent action can be identified in HOT-TM mapping projects?

To address this question, we adopted the approach that tracks the footprint of intelligent activity on system outputs [143], in line with the recommendation to use a quality indicator to study spatial collective intelligence [187]. As noted in the data collection section, the main indicator available

[8] https://bupar.net/

for this purpose in HOT-TM is the field that indicates the validation result of each task as either "Validated" or "Invalidated". This result serves as evidence of whether the underlying mapping process has been executed correctly. Based on this reasoning, we performed a logistic regression analysis. This technique produces coefficients that describe the relationship between quantitative independent variables and a binary dependent variable, VALIDATED (0) / INVALIDATED (1) states.

The selection of independent variables included indicators that capture the collective intelligence factors discussed in the previous sections. To account for the effect of mapper experience, we included the mapping level of the mapper who marked the task as mapped, as this mapper is assumed to have the most influence on decision-making about the task. To account for the effect of mapper location, we included the location of the mapper who marked the task as mapped. The result of this factor should be interpreted with caution given the high level of profiles with incomplete localization and the bias of more experienced users to complete their profiles. To capture the effect of mapper interactions during task execution, we included an indicator specifying whether the task was completed by a single mapper or whether it required the contribution of multiple mappers. In addition to these variables, we included control factors such as the difficulty and priority of the project, as well as the area of buildings mapped per task, reflecting the workload and complexity associated with each task. We also included the relative validation time to account for the potential impact of corrective work performed by validators. Difficulty, priority, involvement of multiple mappers and the mapping level and location of whoever declared the task as finally mapped were coded as dummy variables. Relative validation time was calculated as the proportion of the total processing time devoted to validation, while the area of mapped buildings was logarithmically transformed to account for scale effects.

In order to have an appropriate dataset for the experiment, we trimmed the event log to take into account only the states prior to the first validation. Tasks with states of SPLIT, AUTO-UNLOCKED FOR MAPPING and AUTO-UNLOCKED FOR VALIDATION were also excluded to remove the effect of duplicate states and because of the impossibility of determining the effective time that contributors spent on processing the auto-unlocked tasks.

More details on the analysis procedure can be directly explored on the project repository, which contains Python and R notebooks describing the step-by-step process.

4.5 Results

This section begins by describing the main characteristics of mapping projects and then develops the findings for the three research questions.

4.5.1 Profiling of Humanitarian Mapping Projects

The study includes 746 mapping projects, for which the system records key variables such as difficulty, priority, number of tasks and number of collaborators, along with additional descriptors such as

location and coordinating organization. These variables are characterized according to their relevance in shaping the context of task execution.

The left panel of Figure 4.4 shows the distribution of the projects according to their difficulty, priority, number of tasks and contributors who participated in them. Approximately two thirds of the projects belong to the easy difficulty category, one third to moderate and only a few projects fall into the challenging category. Just over half of the projects have a low priority, almost a third have a moderate priority and the remaining fifth are divided between high priority and urgent projects. The number of tasks and contributors per project exhibits a noticeable dispersion and a positive skew due to some large observations.

The right panel of the same figure elaborates on the relationship between the above variables using a series of scatterplots segregated by level of difficulty that display the number of tasks per project on the horizontal axis and the number of contributors on the vertical axis. Priority levels are encoded using different colors. Between the two main categories of difficulty, there is a difference in the proportion of low priority projects, which tends to be higher for easy projects than for moderate difficulty projects (60% vs. 31%). The majority of urgent projects are of moderate difficulty (69%). There is a positive correlation between the number of tasks and the number of contributors to the projects regardless of the level of difficulty. One fifth of the projects of moderate difficulty were completed by a pair of contributors, presumably one in the role of mapper and the other as validator. No strong patterns are apparent with respect to the association between the priority level of projects and their number of tasks and contributors. Although in moderate projects, more urgent cases seem to recruit more contributors than less urgent projects given an equivalent number of mapping tasks.

Other notable features of the projects are their location and the coordinating organization. HOT-TM projects are mainly organized around four regional hubs serving specific geographical areas of particular interest (Asia-Pacific —AP—, East and Southern Africa —ESA—, West and North Africa —WNA—, Latin America and the Caribbean —LAC—). ESA is the hub with the highest number of projects (41.4%), followed by LAC (26.0%), WNA (17.3%), and AP (11.0%). Projects outside the hubs are infrequent (4.3%). The membership of certain countries in a given hub is a factor that will be taken into account later in the analysis of the geographical origin and destination of contributions. Finally, each project is associated with one of 56 organizations. The organizations with the highest number of projects are HOT (24.9%) and Médecins Sans Frontières —MSF— (17.2%). The other organizations report marginal frequencies.

4.5.2 Data Analysis

4.5.2.1 RQ1: What characterizes the group of individuals collaborating in HOT-TM mapping projects?

With the aim of profiling the members of the group according to their level of experience, Table 4.1 shows the breakdown of the total number of unique contributors who participated in the humanitarian

Projects		
Difficulty	nº	%
Easy	501	67.2
Moderate	239	32.0
Challenging	6	0.8

Projects		
Priority	nº	%
Low	377	50.5
Medium	232	31.1
High	92	12.3
Urgent	45	6.0

	# Tasks	# Contributors
Avg.	418.6	91.2
Std.	487.3	227.5
Min.	8.0	1.0
25%	107.3	14.0
50%	286.0	38.0
75%	564.0	86.8
Max	4,483.0	3,608.0

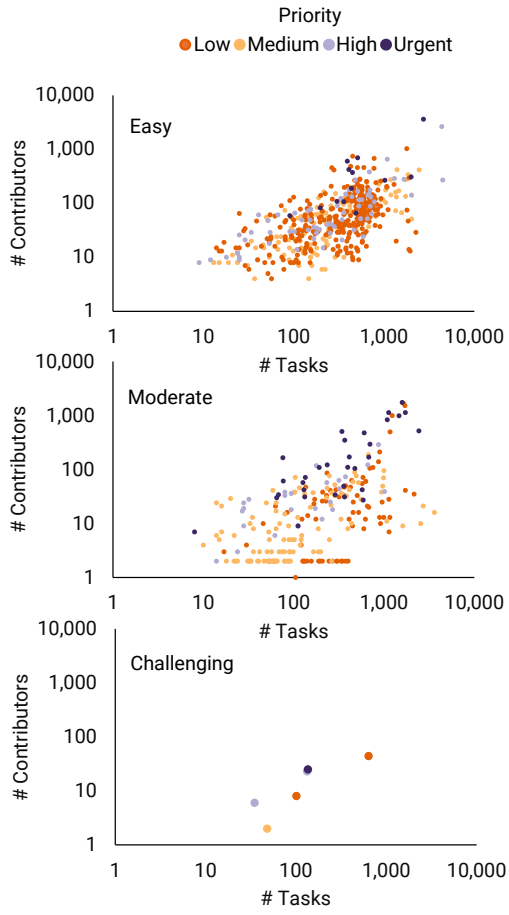


Fig. 4.4 Analyzed projects at a glance

mapping projects analyzed according to their mapping level. It can be seen that the vast majority (approximately nine out of ten) of contributors are beginner OSM mappers. Only a small proportion of contributors are advanced and intermediate OSM mappers. The increase in the mapping level is also followed by a higher frequency of participation in HOT-TM humanitarian mapping projects. Most novice users have participated in a single humanitarian mapping project, while intermediate and advanced contributors are repeat participants.

Mapping Level	Contributors		Nº HOT projects		
	nº	%	Mean	Median	Std.
Beginner	35,387	91.0	1.9	1.0	3.4
Intermediate	980	2.5	12.2	8.0	13.3
Advanced	2,526	6.5	35.7	12.0	79.4
TOTAL	38,893	100	4.4	1.0	22.3

Table 4.1 Overview of total unique contributors by mapping level

Differences in mapping levels are also evident in the completeness of profile fields, as shown in Figure 4.5. Beginners tend to have substantially more incomplete profiles than intermediate and advanced contributors, who have similar completeness rates for most attributes. The exception is the profile picture field, which advanced contributors complete more frequently than intermediate contributors.

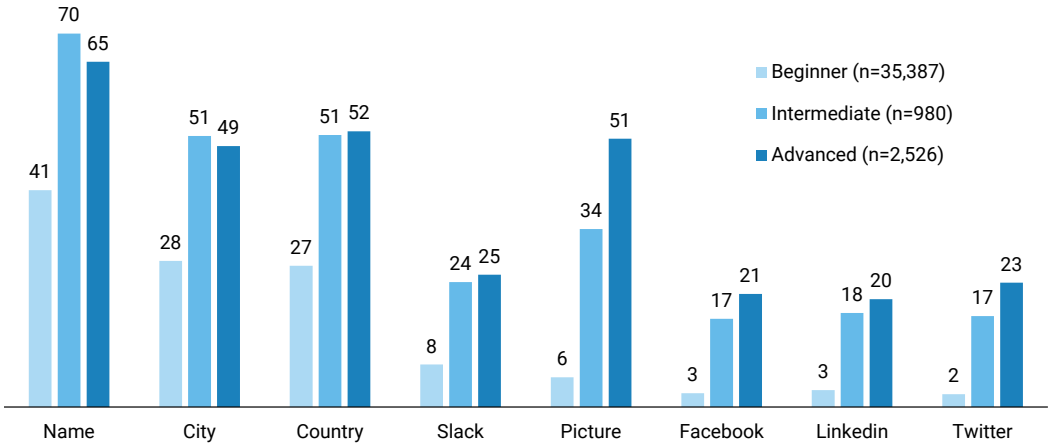


Fig. 4.5 Completeness of the contributor profile by mapping level - % of total contributors-

Another important attribute to characterize the members of the group is their location. In this respect, the reported countries give a clue about the geographical origin and destination of contributions in HOT-TM, as shown in Table 4.2. The table presents the weighted proportion of contributors according to their declared location, categorized by the hubs in which projects are located. The upper part of the table breaks down contributions by hub of origin, while the lower part of the table highlights contributions from the 10 most active countries overall. It must be noted that the calculations exclude contributors who did not report their location, which, as mentioned above, constitute the majority. Approximately two-thirds of the contributions with identifiable origins come from outside the HOT hubs, a trend that holds for all reference regions. Generally, the second largest group of contributors within a hub is those located in the same hub. Looking at the results by country, the United States, the United Kingdom and the Netherlands together account for about one-third of the total contributions, both overall and within each destination hub. This reflects a mapping dynamic in which contributors are still mainly located in the Global North, while most projects are in the Global South. When examining the distribution of users between the different hubs according to their mapping level, some differences can be observed, although they are not drastically pronounced. The proportion of experts ranges from 15.5% in the LAC Hub to 29.8% in the WNA Hub.

To complete the overview of the composition of the groups, Figure 4.6 shows the distribution of contributors according to their mapping level and location for different kinds of projects. The values shown correspond to the weighted average percentage of contributors for each project category. The calculation of the proportion of users by location (same country as the project, another country in

Contributor location		Hub where project is located					
		TOTAL	ESA	LAC	WNA	AP	OTHER
Hub	OTHER	63.6	63.5	54.5	69.6	56.5	74.8
	ESA	10.6	20.1	2.9	7.0	5.7	5.8
	LAC	2.2	1.6	5.0	1.0	1.6	1.8
	WNA	6.5	5.8	3.4	16.9	4.8	5.3
	AP	17.	8.9	34.2	5.5	31.4	12.2
Top 10 countries	United States	15.8	16	18.3	11.2	15.4	16.4
	United Kingdom	11.7	12.1	12.0	11.4	11.3	11.2
	Netherlands	5.2	4.9	1.9	11.0	5.3	5.3
	Philippines	4.7	2.2	16.1	0.7	2.0	2.5
	India	4.5	1.8	11.2	1.3	7.6	2.8
	Nepal	4.5	2.3	4.9	1.7	13.5	3.4
	Germany	3.4	3.6	2.5	4.1	3.3	3.8
	France	3.2	3.0	2.0	6.2	2.4	3.6
	Kenya	2.9	5.7	0.9	0.9	1.1	2.1
	Nigeria	2.6	2.6	1.2	7.3	2.1	1.6

Table 4.2 Origin and destination of HOT contributions - weighted proportion of contributors from project locations.

the same hub as the project and countries outside the hub) excludes contributors who do not report their country. A breakdown of the association between the type of project and the composition of the mapping level of the participants shows that as the difficulty assigned increases, the proportion of advanced users increases substantially to the detriment of beginners. Priority does not seem to be a determining factor in the mapping level mix of a project. Projects with a lower number of tasks are associated with a higher proportion of advanced contributors. The breakdown of proportions by contributor location suggests that the bulk of activity is carried out by contributors outside the country and the hub where the projects are located regardless of the category of project in question. However, the proportion of mappers located in the same country tends to be higher for lower priority projects with fewer tasks.

4.5.2.2 RQ2: What characterizes collective action in HOT-TM mapping projects?

a) The Mapping Process: How Does the Work Happen?

Figure 4.7 illustrates the typical lifecycle of a mapping task, highlighting the various states it progresses through. This directly-follows graph, which represents the frequency and duration of mapping task states, was derived from the 85% most common traces. In terms of frequency, the nodes indicate the absolute number of state instance executions, while the edges represent the absolute number of times the source and target states were executed sequentially. Both the thickness of edges and vertices are proportional to the frequency of occurrences, with higher frequencies resulting in

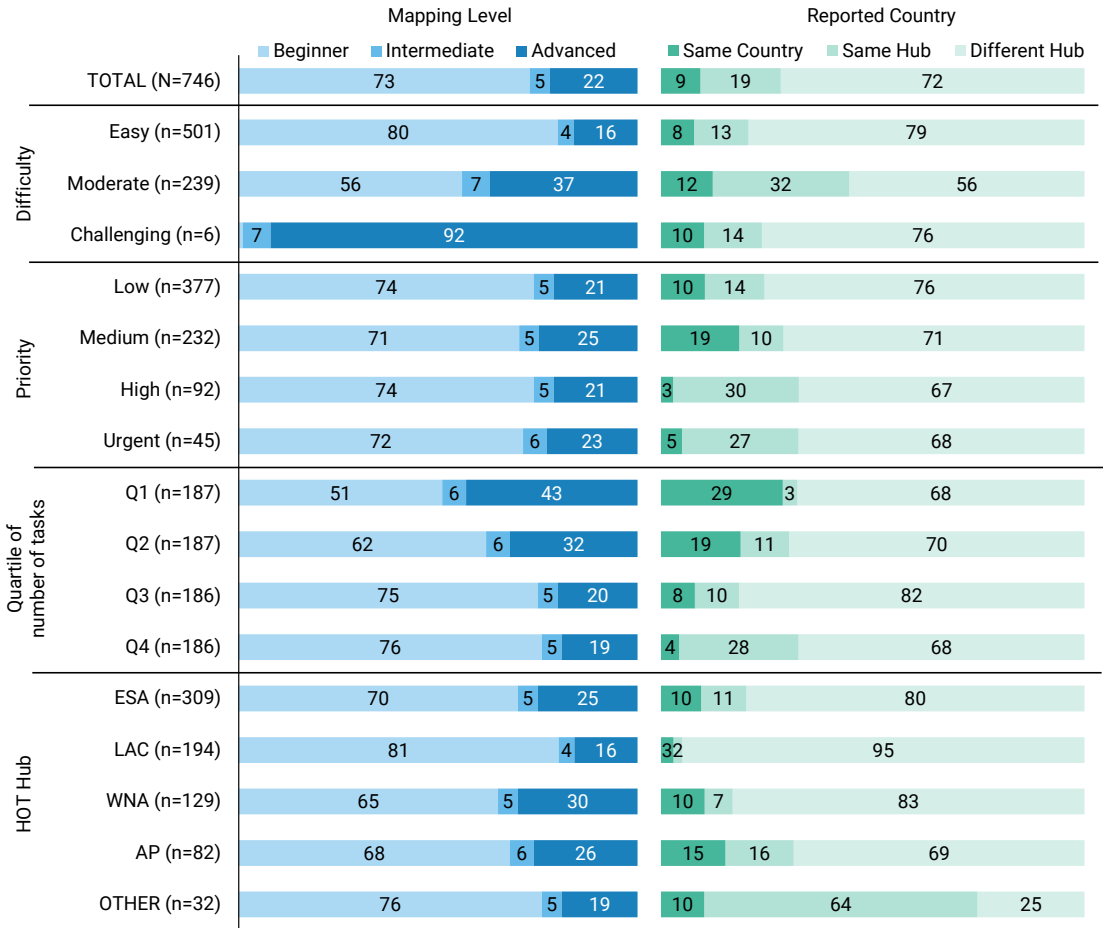


Fig. 4.6 Participation structure in projects - weighted proportion of contributors per project-

thicker edges and vertices. The median duration of states or transitions is shown just below each frequency count, either in minutes, hours or days as appropriate. Those transitions whose median duration was longer than one day are highlighted in orange.

The flow reveals a primary pathway for mapping tasks that follows a straightforward progression. Typically, contributors lock a task for mapping and complete one or two mapping cycles before it is marked as MAPPED. The task then moves to the LOCKED FOR VALIDATION state and is usually declared VALIDATED on the first attempt. However, alternative, less frequent pathways also exist, as indicated by states such as SPLIT, AUTO-UNLOCKED FOR MAPPING, and INVALIDATED. It is important to note that the frequency of these deviations from the standard mapping process varies considerably depending on the project type, increasing for projects with higher difficulty and priority (see Table 3 in Appendix 'A Closer Look At the Mapping Process'). This variation underscores how project characteristics shape the complexity of task execution and decision-making processes.

Shifting focus to the temporal dimension of the flow, we find that the median duration of a mapping cycle is just 2.1 minutes. However, if a task is not declared MAPPED at the end of a cycle, it faces a median wait time of one day before re-entering another mapping cycle. Once a task is marked

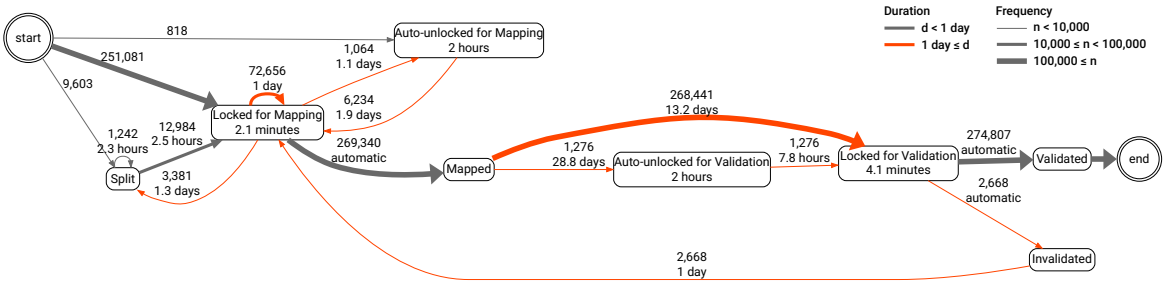


Fig. 4.7 Frequency and duration map of task states and transitions (85% of most frequent traces)

as MAPPED, it remains in a queue for a median of 13.2 days before being LOCKED FOR VALIDATION, a process that itself takes a median of 4.1 minutes. Similarly, if a task is INVALIDATED, it typically waits another day before re-entering the mapping phase. Notably, the actual processing times for mapping and validation are brief—just a few minutes—whereas the transitions between these stages introduce significant delays, often measured in days. This contrast underscores how waiting time constitutes one of the main "costs" of decision-making. The bottleneck becomes particularly evident when a mapper hesitates to mark a task as MAPPED or when a validator decides to INVALIDATE a task. Another key observation, which will be explored further, is that the median time a task remains in the LOCKED FOR VALIDATION state is twice as long as in the LOCKED FOR MAPPING state. Once again, the duration of states and transitions varies significantly by project type (see Table 4 in Appendix 'A Closer Look At the Mapping Process'). For instance, tasks in low-priority projects can wait nearly a month for validation after mapping, whereas for other projects this delay is typically less than a week.

b) Roles and Responsibilities: Who Does What in Mapping?

It is now time to address the issue of role configuration, with a particular focus on identifying who assumes the burden of creation and decision-making within the system. To this end, Table 4.3 shows the breakdown of the states according to the mapping level of the contributors responsible for their execution, depending on whether they were beginners or advanced contributors. Intermediate contributors, reporting a generally marginal presence, were omitted for the sake of readability. The first row of results takes the weighted average number of contributors per project for each mapping level as a benchmark. The remaining rows show the proportion of the total instances of each state that was performed by one or another contributor profile according to the type of project.

The first key finding concerns the creation activities during the mapping phase. According to the proportion of LOCKED FOR MAPPING states, advanced mappers take on a significantly higher mapping workload compared to their proportional representation within the group of mappers across all project types. However, this distribution depends on the difficulty of the project. In terms of overall volume of activity, beginners carry the main mapping load on easy projects. This trend is reversed in projects of higher difficulty, where the mapping load is concentrated on advanced mappers. Advanced mappers also play a dominant role in the decisions made during the mapping phase, as

		TOTAL N=312,289		Easy n=222,729		Moderate n=88,468		Challenging n=1,092	
		Beginner	Advanced	Beginner	Advanced	Beginner	Advanced	Beginner	Advanced
Proportion of contributors mapping phase		76	20	82	13	58	34	1	91
MAPPING PHASE	Locked for mapping	51	42	61	32	29	65	0	99
	Mapped	44	50	55	39	17	77	0	99
	Auto unlocked for mapping	74	20	79	16	65	29	0	100
	Split	25	71	35	60	15	82	0	100
	Bad imagery	52	44	82	12	14	85	0	100
Proportion of contributors validation phase		4	91	7	89	1	94	0	97
VALIDATION PHASE	Locked for validation	1	98	1	98	0	99	0	100
	Auto unlocked for validation	2	96	2	95	0	96	0	100
	Validated	0	98	0	98	0	99	0	100
	Invalidated	2	96	3	96	0	96	0	100

		Low n=166,911		Medium n=80,768		High n=38,119		Urgent n=26,491	
		Beginner	Advanced	Beginner	Advanced	Beginner	Advanced	Beginner	Advanced
Proportion of contributors mapping phase		77	18	74	22	76	19	73	21
MAPPING PHASE	Locked for mapping	53	41	44	49	47	45	57	35
	Mapped	47	47	41	53	37	56	43	50
	Auto unlocked for mapping	71	23	68	24	74	21	77	17
	Split	27	69	27	69	18	79	25	71
	Bad imagery	75	22	34	64	76	11	41	48
Proportion of contributors validation phase		6	90	7	89	2	95	0	92
VALIDATION PHASE	Locked for validation	1	99	0	98	0	96	0	98
	Auto unlocked for validation	2	96	2	97	1	98	1	91
	Validated	0	99	0	98	0	96	0	98
	Invalidated	6	89	1	97	0	98	0	99

Table 4.3 State execution based on the contributor mapping level - % of states-

suggested by a substantially higher than expected frequency of MAPPED and SPLIT states in all project categories. In contrast, the proportion of AUTO-UNLOCKED FOR MAPPING states is mainly associated with beginners, which could indicate a higher likelihood of interface-related misunderstandings, lack of confidence in making decisions or deviations from recommended mapping practices. In the validation phase, all states are executed almost exclusively by advanced contributors. This is due to the participation criteria set by project managers at the beginning of the project, which usually exclude beginners.

Incomplete data from the location field limits a detailed analysis of the distribution of states by mapper location. However, to provide an overview of the intensity of mapper contribution as a function of their location with respect to the project, Figure 4.8 presents the distribution of total MAPPED states by project priority and project size (measured by number of tasks). These two variables showed the greatest variation in the concentration of national mappers, as highlighted in Figure 4.6. As a benchmark for assessing the volume of activity, the first row presents the total weighted proportion of users in each location category across all projects. It is important to note that this calculation excludes MAPPED states executed by users with unknown location, and interpretation should take into account a possible bias towards expert users, who tend to have more complete profiles. The results indicate that national mappers tend to contribute proportionally more than their representation within the

overall group of contributors. In addition, their relative activity appears to be concentrated on lower priority projects with a lower volume of tasks. Projects with these characteristics are more suitable for preventive, preparatory and follow-up mapping activities.

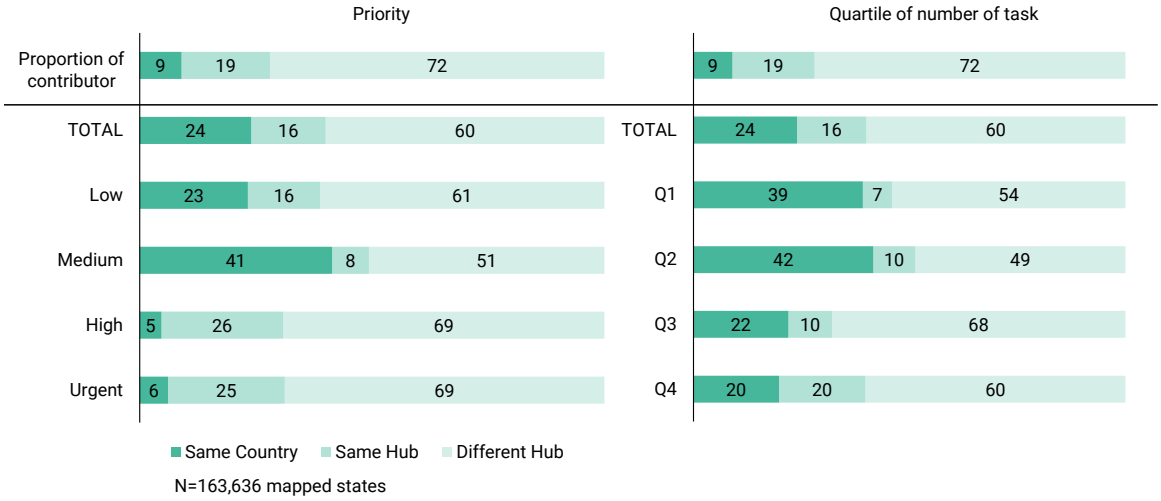


Fig. 4.8 Execution of mapped states based on contributor location - % of mapped states where the location of the mapper is known-

c) Collaboration in Mapping: How Do Contributors Interact?

After analyzing the nature of the work and the configuration of roles based on mapping states, the next step is to examine the interactions between the mappers in completing the tasks. To do this, Figure 4.9 elaborates on the number of contributors involved in the completion of a task. We take as a basis the tasks that do not report frictions such as splits or invalidations, which are the majority. Then we show the case of tasks with splits, invalidations or both states. For each task, a distinction is made as to whether such participation occurred for the total number of states or for the states corresponding to the mapping and validation phases. Descriptive statistics such as average number, standard deviation, and quartiles are displayed for each scenario.

If we look at the averages, the number of contributors needed to carry out the mapping operations of a task multiplies almost 4-fold when SPLIT, and 5-fold when INVALIDATED states occur. In the case of the validation phase, the presence of these states also reports an increase, albeit more discrete, in the number of validators. This suggests that, in addition to the waiting time costs discussed above, decisions to perform a split or an invalidation must also be assessed in terms of the number of mappers needed to process the task.

Having realized that tasks that suffer disruptions such as splits or invalidations by nature require a combined effort between several contributors, the following results focus on the analysis of interactions in ordinary tasks, which comprise the vast majority of cases. We then concentrate on states in the mapping phase corresponding to tasks that have not undergone splits or invalidations to analyze

N=312,289		Avg.	Std.	25%	50%	75%
No split/No invalidated n=263,165 (84.3%)	Total	2.4	0.9	2	2	2
	Mapping	1.4	0.9	1	1	1
	Validation	1.0	0.2	1	1	1
Split n=42,223 (13.5%)	Total	7.2	6.2	3	5	9
	Mapping	5.9	5.8	2	4	8
	Validation	1.5	1.0	1	1	2
Invalidated n=19,750 (6.3%)	Total	10.0	7.4	5	7	13
	Mapping	8.1	7.0	3	6	10
	Validation	2.3	1.1	2	2	3
Split & Invalidated n=12,849 (4.1%)	Total	12.5	7.9	7	11	16
	Mapping	10.5	7.3	5	9	14
	Validation	2.6	1.2	2	2	3

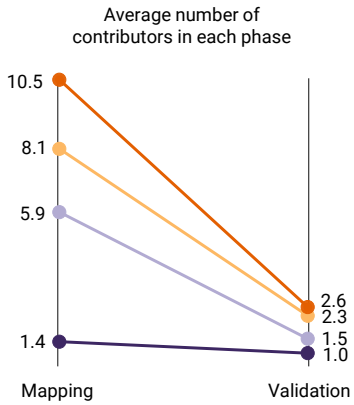


Fig. 4.9 Number of contributors per task, depending on the occurrence of Split or Invalidated states (average number, standard deviation, and quartiles)

interactions. As explained previously, the HOT-TM micro-tasking scheme allows several mapping operations of the same or different mappers on a task before declaring it as mapped, prior to its validation. These multiple operations, if they occur, are not done simultaneously, as the task is locked for mapping. Each subsequent edit builds on the work done by the previous mappers, as the map is done on the shared OSM database.

Figure 4.10 shows the average percentage of tasks per project whose mapping states were executed by more than one mapper, broken down by difficulty and priority. It is important to note that, proportionally, tasks are predominantly carried out by individual mappers rather than through collaborative efforts in all project categories. However, there are variations. Higher priority levels mean an increase in the number of tasks with multiple mappers. In terms of project difficulty levels, easy projects report the highest proportion of collaborative tasks, followed by projects of moderate difficulty and challenging projects.

In the next step, we focus on understanding the most common combinations of mapping levels that are observed in collaborations in the mapping phase. Figure 4.11 presents a variant explorer of the five most frequent combinations of mapping levels involved in LOCKED FOR MAPPING and AUTO-UNLOCKED FOR MAPPING states before the first validation of a task. Next to each combination, the percentage of the total collaborative tasks it represents is shown. It can be observed that the 5 most common combinations alone account for approximately two thirds of the total collaborative tasks for almost all project types. Most of these popular combinations are binary and to a much lesser extent trinary. The combination involving two beginners is the most common, except for projects of moderate and challenging difficulty and high priority. An equivalent analysis based on the location of mappers was not carried out due to noise caused by incomplete user profiles.

Figure 4.12 elaborates on the dynamics of collaborative tasks using the "handover of work" concept [103]. In this figure, the nodes identify the mapping levels and the weights of the arcs between them are based on the proportion in which a handover occurs from one mapping level to another,

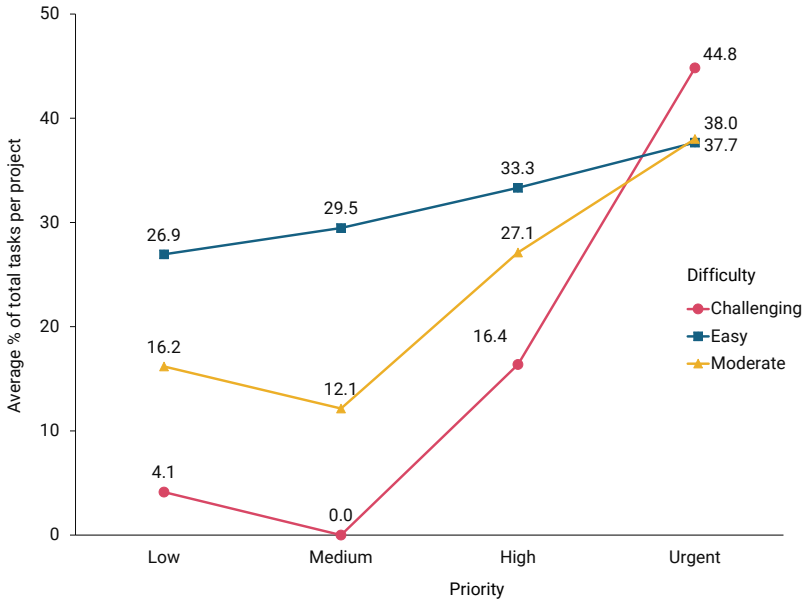


Fig. 4.10 Tasks mapped by more than one contributor by project difficulty and priority

100% of the handovers leaving each mapping level to the others. The shaded area of the nodes shows the proportion of the total number of mapping states executed by each mapping level that belong to collaborative tasks.

As can be seen, most of the mapping states of novice contributors occur in collaborative tasks and that proportion decreases considerably for intermediate and advanced contributors. For the latter, task mapping is mostly solitary. Regardless of the mapping level, the proportion of collaborative mapping increases progressively with the urgency of the projects. In terms of difficulty, the proportion of collaborative tasks increases for beginners when they move from easy to intermediate projects, while for advanced users the proportion of collaborative tasks decreases with increasing difficulty. The arcs suggest that the handover from beginner and advanced users is to contributors of the same mapping level. For intermediate contributors, the results are mixed.

The frequency of handovers between groups may simply reflect the overall volume of mapping activities carried out by each group within the respective project categories, rather than indicating a greater or lesser tendency for interaction between groups. To account for this, the Table 4.4 compares the observed handover frequencies with the expected values based on the proportional distribution of mapping states between groups. These expected values are estimated using the combined frequencies of the LOCKED FOR MAPPING and AUTO-UNLOCKED FOR MAPPING states of each group, as presented in Table 4.3. It can be observed that handovers between novice users are substantially higher than expected to the detriment of handovers from an advanced user to a novice user which is consistently lower than expected.

The overall picture not only reflects a scenario where mapping is primarily driven by individual contributions (collection) but also highlights a limited level of meaningful collaboration. By limited

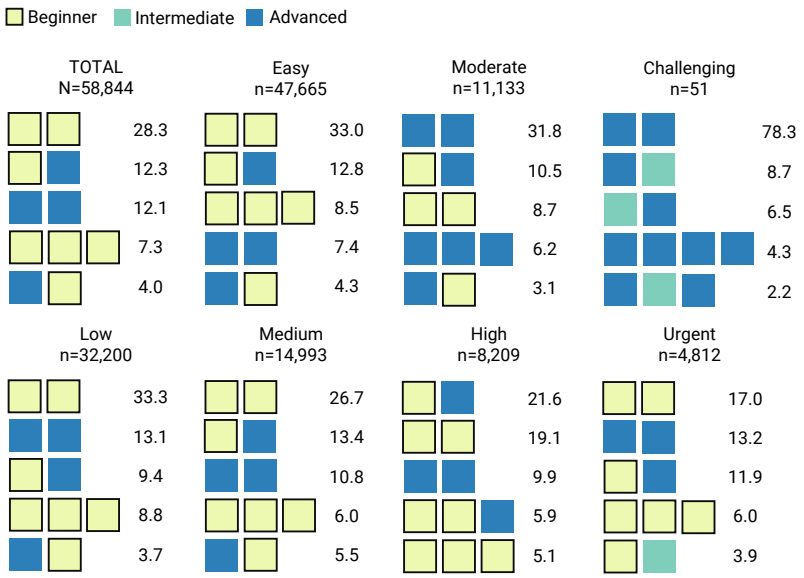


Fig. 4.11 Variant explorer representation of mapping level profiles across mapping states for tasks completed by more than one contributor -% of tasks-

we do not only refer to the lower frequency of collaborative task execution but to the nature of these interactions. Collaboration seems to be the result of novice mappers leaving tasks incomplete or being unsure whether their work is sufficiently accurate, thus handing over responsibility for completion and decision to another mapper. This handover often comes at the cost of delaying the progression of the task by one or more days. This is in stark contrast to the behavior of advanced users who try to complete tasks individually, helping to dispatch tasks quickly and without friction. In this context, collaboration (defined as the involvement of several mappers to complete a task) seems to indicate inefficiency rather than representing a desirable behavior.

4.5.2.3 RQ3: What evidence of intelligent action can be identified in HOT-TM mapping projects?

So far, performance considerations have focused on aspects such as the complexity of the mapping process in terms of sequence, duration and interaction, as reflected in task states. However, these indicators alone do not directly reveal whether mapping activities have been carried out properly. Therefore, the following section aims to identify factors that contribute to the success of mapping outcomes, providing potential evidence of intelligent collective action.

We performed a logistic regression to determine the effect of the task characteristics on the first validation result, that is, to determine if there are factors that make a task more likely to be validated or invalidated. As mentioned in the methodology, the dataset includes states prior to the first validation.

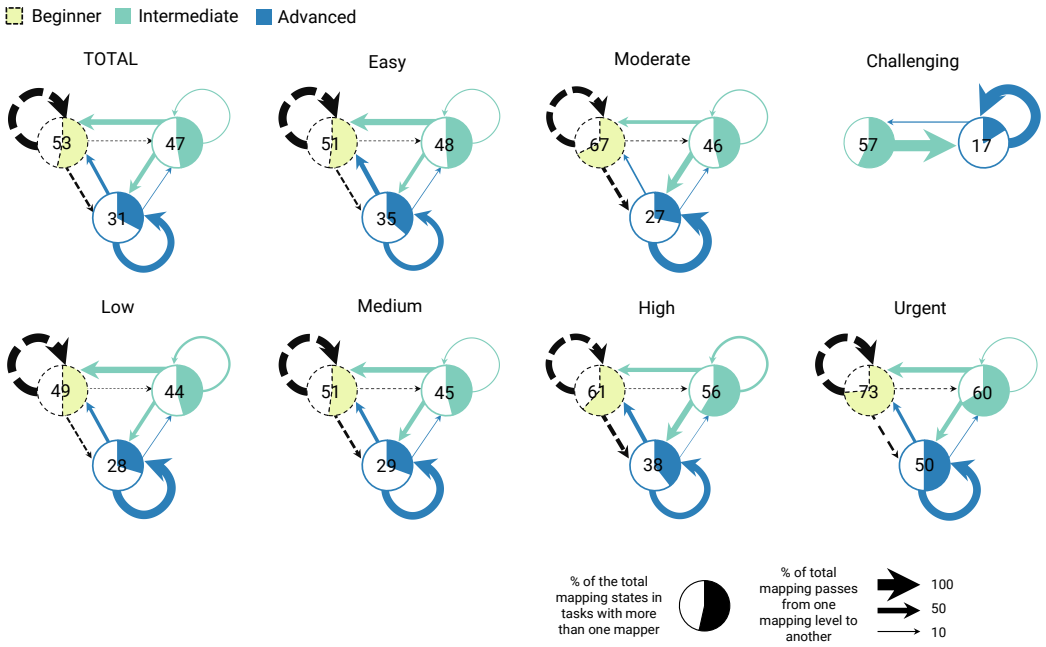


Fig. 4.12 Handover of mapping tasks

Tasks with SPLIT, AUTO-UNLOCKED FOR MAPPING and AUTO-UNLOCKED FOR VALIDATION states were discarded to remove noise. This results in a total of 281,653 tasks.

Table 4.5 lists each of the independent variables along with their regression coefficient, standard error, z statistic, and p-values. Larger regression coefficients indicate higher probabilities of invalidation. However, since these numbers are often not very intuitive, they are accompanied by odds ratios (OR). The odds ratios represent an exponential transformation of the regression coefficient, implying a multiplication factor of the dependent variable for each one-unit increase in an independent variable. For each categorical variable where hot encoding was applied, the proportion of the total number of tasks belonging to that category is shown in parentheses, along with the average invalidation rate of the group, including the reference categories. For the numerical variables, Relative Validation Time and Building Area Mapped, the invalidation rates for tasks below and above the median for each indicator are displayed.

Starting with the effect of the mapping level of the mapper who marked the task as mapped, used as an indicator of experience, tasks are less likely to be invalidated when declared as mapped by more advanced mappers. Continuing with the effect of mapper location, the highest performance is observed among mappers from different hubs, followed by national mappers. However, this result should be interpreted with caution, as the unknown category comprises an unidentified mix of origins, and the location data is skewed toward expert mappers due to their more complete profiles. As for the effect of the involvement of several mappers, used as an indicator of interactions, the probability of task invalidation increases. Among the control factors, higher difficulty levels, higher priority levels

	TOTAL N=97,847			Easy n=77,508			Moderate n=20,288			Challenging n=51		
Handover	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.
Beginner - Advanced	16	21	-6	16	19	-3	14	20	-6			
Beginner - Beginner	49	28	21	55	39	16	25	11	14			
Beginner - Intermediate	4	4	1	5	4	0	3	2	1			
Intermediate - Advanced	3	3	0	2	2	0	4	4	0	8	1	6
Intermediate - Beginner	4	4	0	4	4	-1	3	2	1			
Intermediate - Intermediate	1	0	0	1	0	0	1	0	1			
Advanced - Advanced	15	16	-1	9	9	-1	39	38	1	82	97	-15
Advanced - Beginner	7	21	-14	7	19	-12	7	20	-13			
Advanced - Intermediate	2	3	-1	1	2	-1	4	4	0	10	1	8

	Low n=49,058			Medium n=24,470			High n=13,868			Urgent n=10,451		
Handover	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.	Obs.	Exp.	Diff.
Beginner - Advanced	13	22	-9	17	22	-4	26	21	4	15	20	-5
Beginner - Beginner	54	29	25	45	21	25	41	25	16	45	37	7
Beginner - Intermediate	4	3	1	4	3	1	4	4	0	5	4	1
Intermediate - Advanced	2	2	0	3	3	0	4	3	1	3	2	1
Intermediate - Beginner	3	3	0	4	3	1	3	4	-1	4	4	-1
Intermediate - Intermediate	1	0	0	1	0	0	1	1	0	1	1	0
Advanced - Advanced	15	16	-1	15	23	-7	11	18	-7	18	10	7
Advanced - Beginner	7	22	-15	8	22	-13	8	21	-13	8	20	-12
Advanced - Intermediate	2	2	-1	2	3	-1	2	3	-2	2	2	0

Table 4.4 Observed and expected frequency of handovers -% of handovers-

and larger building area mapped are associated with higher invalidation rates. Finally, longer relative validation time tends to correspond to lower probabilities of task invalidation.

This result aligns with the findings from previous sections, suggesting that the outcome of the system is not casual but is directly shaped by the group composition and the dynamics of collective action. Advanced mappers are not only more active, report less friction in the process and lead mapping and validation decisions but their contributions tend to better mapping results. This is evidenced not only by higher validation rates but also by longer validation times associated with fewer invalidations that are likely to reflect corrective work on their part. These results also highlight the challenges posed by the contribution dynamics of beginners. It is not just that their validation success rates are lower. In easy and low-priority projects, where their contributions on mapping are concentrated, the relatively longer time invested by validators (presumably for corrective work) seems to be necessary to compensate for the performance of beginners. Finally, evidence of the shortcomings of collaborative mapping, defined as the involvement of several mappers to complete a task, is reinforced. Collaborative mapping not only slows down the process but also is associated with further invalidation.

	N=281,653	% Inval.	β	OR	S.E	z	p value	
	(Intercept)	-	-2.30	0.10	0.03	-83.21	<2e-16	***
Experience of mappers	Mapped [Beginner] yes (vs. no) (44.7%)	5.9	-	-	-	-	-	-
	Mapped [Intermediate] yes (vs. no) (5.9%)	4.9	-0.59	0.55	0.04	-14.52	<2e-16	***
	Mapped [Advanced] yes (vs. no) (49.3%)	3.1	-1.15	0.32	0.02	-46.46	<2e-16	***
Location of mappers	Mapped [Unknown] yes (vs. no) (41.9%)	5.5	-	-	-	-	-	-
	Mapped [Same Country] yes (vs. no) (14.5%)	4.4	0.23	1.26	0.03	7.27	3.7e-13	***
	Mapped [Same Hub] yes (vs. no) (7.8%)	6.8	0.35	1.42	0.03	10.40	<2e-16	***
	Mapped [Different Hub] yes (vs. no) (35.8%)	2.9	-0.36	0.70	0.02	-14.96	<2e-16	***
Collaborative mapping	Multiple Mappers no (vs. yes) (75.3%)	3.0	-	-	-	-	-	-
	Multiple Mappers yes (vs. no) (24.7%)	8.8	0.44	1.55	0.02	21.72	<2e-16	***
Difficulty	Difficulty [Easy] yes (vs. no) (72.0%)	3.7	-	-	-	-	-	-
	Difficulty [Moderate] yes (vs. no) (27.6%)	6.4	0.66	1.93	0.02	27.47	<2e-16	***
	Difficulty [Challenging] yes (vs. no) (0.3%)	12.9	1.30	3.67	0.10	12.38	<2e-16	***
Priority	Priority [Low] yes (vs. no) (54.5%)	3.1	-	-	-	-	-	-
	Priority [Medium] yes (vs. no) (26.9%)	4.0	0.21	1.23	0.02	8.37	<2e-16	***
	Priority [High] yes (vs. no) (11.9%)	5.6	0.46	1.58	0.03	15.48	<2e-16	***
	Priority [Urgent] yes (vs. no) (6.7%)	15.5	1.16	3.19	0.03	40.48	<2e-16	***
	Building Area Mapped log([Square meters of buildings]+1)	<M: 3.5, >M: 5.4	0.06	1.06	0.00	23.72	<2e-16	***
	Relative Validation Time % Validation time/ (Validation time + Mapping time)	<M: 7.4, >M: 1.5	-0.03	0.97	0.00	-74.18	<2e-16	***

*, **, *** significant at p ≤ 0.10, 0.05 and 0.01 respectively.
M: median.

Table 4.5 Logistic Regression for validation state of a task (Validated vs. Invalidated)

4.6 Discussion

Our discussion focuses on synthesizing and combining the findings on the collective intelligence system in HOT-TM projects with arguments from related work to present actionable insights and contributions to the field. It also reflects on possible ways to address the limitations identified in the current system set-up.

4.6.1 Towards a More Sustainable Group Composition of Humanitarian Mappers: Developing Novice Mappers and Enhancing Local Participation

The results of our study provide an overview of the group composition of HOT-TM mappers based on the key attributes of experience and location. Broadly speaking, this characterization is in line with the profile of VGI participants derived from previous research, although previous studies tend to address this issue in a more fragmented way. By incorporating insights into the processes and outcomes of the collective intelligence system, our study sheds light on how group composition both influences and is influenced by these two key components of the system.

Starting with the attribute of expertise, which appears to have the greatest impact on the current scheme, HOT-TM reproduces the pattern of a group characterized by a small proportion of advanced contributors outperforming a majority of beginners both in the relative quantity and quality of their contributions and in the stability of their participation over time [171, 173, 172, 195]. Our findings on how decision-making functions are articulated in the mapping, together with validation rates on projects of varying levels of difficulty and priority, suggest that the system is designed to capitalize on the 'wisdom' of advanced contributors. This is achieved through deliberate decisions (such as restricting the validation and mapping of more complex projects to experienced mappers) or through presumably more spontaneous behavior, such as corrective actions by validators to address the contributions of beginners. Meanwhile, beginners are largely confined to easier, lower priority projects and have limited involvement in decision-making processes. Our study reveals that this design has at least one major cost: the slowness of the validated mapping results. This is evident from our analysis of the duration and limited availability of the advanced users. Given that some of these data are used for active disaster response efforts, this delay represents a substantial inconvenience rather than an insignificant trade-off.

Our research also complements the work of Yin et al. [151] by providing more detailed findings. While they use macro-level project statistics to suggest that the HOT-TM micro-task model reflects the dependence on user authority and self-reinforcement characteristic of OSM communities [187], our study offers a more operational perspective on these dynamics. By following this approach, the system does not encourage long-term retention or support the skills development of beginners. The marginal presence of intermediate mappers, both in terms of their proportion within the group and their overall contribution to the mapping process and its outcomes, that we detected is further evidence of this lack of investment in the development of a cadre of more experienced mappers.

In light of the above, fostering the development of entry-level mappers through a strategy that supports their retention and progression to higher levels of performance presents a promising opportunity to build a more sustainable community. This strategy could take the form of a structured 'career path' that integrates short and long-term actions.

The tendency of newcomers to disappear after contributing to a single project, combined with signs of uncertainty and hesitation as they map, suggests that an unsatisfactory user experience could contribute to their lack of return. In this respect, an onboarding process designed to create a more

rewarding first mapping experience could play a key role. Regular mapathons (both physical and virtual) that facilitate interaction between mappers and foster a sense of community provide an ideal framework for welcoming newcomers. In addition, the user interface should guide users through a navigation flow that encourages exploration of relevant sections and ensures a satisfactory mapping experience. Simple actions, such as asking users to complete their profile information, could foster a greater sense of identity as contributors. However, the most significant impact lies in empowering these users to gain confidence in completing tasks and making mapping decisions. HOT-TM already provides comprehensive materials on how to edit maps effectively. These resources could be reinforced with contextual hints to help novice mappers discern when a task is ready to be marked as complete, when it requires splitting, or when it should be reported due to poor imagery.

In the medium term, gamification elements could be explored to encourage novice users to return and gradually move on to more complex mapping tasks, as suggested by the work of Watkinson et al. [196]. This is in line with the findings of Urrea and Yoo [173] who highlight the positive effect of reaching new levels of experience, especially for novice users, and suggest that further segmentation of ranks based on experience could improve retention. Detailed data on mapper profiles (including demographics, preferences and performance) can help implement a reporting system that not only assigns online volunteers to the projects where they are most productive but also matches their career progression. In terms of long-term retention strategies, it is important to recognize the central role of altruism as a motivator for participation in OSM [182], and in humanitarian mapping in particular [197]. It is therefore convenient to actively communicate to mappers the tangible impact of the projects to which they contribute and, wherever possible, to highlight their individual contributions to relief and disaster prevention efforts.

Continuing with the locality factor, it is important to note that our work can be seen as a checkpoint in assessing progress towards the goal of developing sustainable, locally owned community mapping ecosystems in at-risk regions around the world, as proposed by Soden and Palen [178] at the beginning of academic research on HOT. Our findings indicate that the majority of contributions come from mappers outside the local hub, located in the Global North. In second place, are contributions from national mappers. It further reinforces the contradiction (already noted in other mapping contexts [179–181]) that locality, the founding principle of the VGI concept [147], seems to be relegated to the background.

As already mentioned, the main initiative to decentralize the organization and incorporate local expertise in project management has been the creation of Open Mapping Hubs.[9] However, these initiatives are relatively recent, as the current four hubs were created between 2021 and 2023. Presumably, these hubs are still in the development phase, so their impact may not yet be fully reflected in the mapping results. In this context, the results presented here can serve as a reference for future assessments of the impact of the hubs as they become more mature. However, there are promising signs. An example of this is the fact that mappers from the hub in the area of interest are

[9]https://www.hotosm.org/hubs/

the second largest group, after mappers not affiliated to any hub, in almost all regions. This highlights the presence of a critical mass of local mappers that could be tapped to further develop HOT hubs. Additionally, the contribution of national mappers far exceeds their proportion of the group, and is particularly pronounced in the lower priority projects and projects with fewer tasks. This pattern suggests a possible specialization in preventive, preparatory and follow-up mapping activities.

In addition to actions at the organizational governance level, technical solutions must also be part of the approach. The current design of HOT-TM is optimized for desktop use but is not well suited to field operations. However, new tools for localized mapping are being developed, such as the Field Mapping Tasking Manager (FMTM),[10] a tool for organizing ground mobile survey mapping, which is currently in beta. These recent innovations point in the right direction but it is still too early to assess their full impact, which warrants continuous monitoring and evaluation.

4.6.2 Towards a More Meaningful Collective Action in Humanitarian Mapping: Facilitating Effective Collaboration Among Mappers

In examining the dimensions of collective action, HOT-TM seems to heavily priorities collection over collaboration at the level of the mapping task. Indeed, if we adopt a loose definition of collaboration (i.e. the mere contribution of several mappers to complete a task [184, 185]) it tends to be more of an anomaly than a beneficial factor in achieving mapping goals. On the contrary, it tends to lead to rework, slows down the progress and reveals systemic problems in the user experience for beginners. In addition, collaborative instances tend to show worse validation performance. Thus, the expected improvement effect associated with the 'wisdom of crowds' [167] concept does not seem to emerge, at least at the task level. If the main objective is to cover large territories quickly, a contributor should ideally complete a task independently within a few minutes, without the need for subsequent intervention by other mapper.

The analysis shows that the interactions between novices and experts are mainly characterized by corrections made by the latter as validators of the work of the former, a pattern that is also observed in the general dynamics of OSM [185]. The correction process in OSM usually occurs discretely [186], which means that novice mappers receive little or no direct information about the specific reasons for a correction. In HOT-TM, contributors are notified when a task they have mapped is validated or invalidated, provided that they have enabled notifications. However, the notification alone does not guarantee that users receive meaningful feedback on their mapping performance, especially if the validators take on the correction on their own. This lack of guidance may hinder their development as mappers and reinforce existing group dynamics.

These results support the concerns raised by Kogan et al. [186] regarding the limitations of relying solely on co-editing to study interactions between mappers. In this sense, the lack of visibility in discussion channels in OSM [186] is also a blocking factor present in HOT-TM. Although HOT-TM includes a comments section for each task, its visibility is somewhat limited, and it may not always

[10] https://fmtm.hotosm.org/

offer clear guidance to subsequent mappers. As a result, understanding why a previous contributor left a task incomplete or identifying aspects that still require attention can be challenging. Enhancing the system with a more actionable task history could support more effective collaboration by helping mappers complete tasks more efficiently and enrich the mapped elements.

The challenge, therefore, lies in fostering interactions between novice and advanced mappers that facilitate knowledge transfer. While this type of mentoring interaction often occurs organically during face-to-face mapping events [186], there is potential to further develop mechanisms that provide constructive and accessible feedback in response to validation outcomes. These mechanisms should avoid intimidating styles that could discourage future contributions and instead create a supportive environment that encourages continued engagement and growth.

It is also worth considering whether the intrinsic nature of current mapping tasks requires and encourages collaboration during their execution. Where tasks are limited to adding missing elements, such as buildings or roads, it is understandable that the involvement of additional mappers would not significantly enrich the outcome. In this context, diversification of task types could offer valuable improvements. For example, enrichment tasks could complement creation tasks, in line with the specialization patterns identified by Zhang et al. [183]. This approach could support a layered model in which co-editing adds significant value to map content, helps to close the detail gap that distinguishes humanitarian mapping from other types of OSM mapping [148], and further emphasizes the importance of local contributions. Note that national mappers now appear to be significantly active but this is not necessarily associated with greater success in validation.

4.6.3 Towards More Intelligent Collective Action in Humanitarian Mapping: Fostering Greater Equity Without Compromising Productivity

As noted since the introduction of the concept of collective intelligence, defining intelligence can be challenging, as it often depends on the objectives of the observer [143]. For example, if we focus solely on the goal of productivity (covering territory and validating tasks), this perspective fits well with the urgent needs of first response efforts during an ongoing emergency. Although no formal benchmark exists, an overall task invalidation rate of 6.3% and 15.5% for urgent projects suggest that the current system is relatively efficient in producing an initial mapping output from mapping operations that typically take only a few minutes. However, the need for corrective work by validators and the waiting period of several days for validation highlight potential areas for improvement in the quality of this initial output.

If we shift the focus from productivity to equity-related objectives, the limitations of the current approach become more apparent. Equity-oriented objectives could include fostering the growth of the mapping community, promoting the integration of local communities, and, most importantly, avoiding unsustainable mapping outcomes—such as the creation of low-quality data that are not maintained and quickly lose relevance [148, 151]. Note that these other objectives are better suited to the needs of prevention, preparedness and long-term monitoring of emergencies.

Our results and observations align with and extend the findings of Yin et al. [151], who highlight that in HOT-TM, as in other micro-tasking tools used in peer production, there is an inherent trade-off between productivity and equity. This may raise questions about how to address the tension between productivity and equity objectives. In this sense, it would be beneficial to allow project managers to define the most appropriate mapping objectives (whether focused on productivity or equity) before the project is launched, depending on the specific needs and timing of the emergency. These objectives could then be linked to tasks aimed at creating or enriching the map, as appropriate. In addition, targeted recruitment efforts to assign online volunteers to projects where their contributions will have the greatest impact could further support this approach, as suggested by Urrea and Yoo [173].

The review of objectives must be accompanied by a corresponding review of system metrics, especially given the strong influence these metrics often have on system incentives. As Spielman [187] suggests, if there is any kind of spatial collective intelligence, it should be reflected in map quality measures focused on the credibility and accuracy of the output. In HOT-TM, the most direct indicator of the success of a task is achieving validation. This is a credibility metric, since it comes from the status of the validators. However, this metric aligns with the dynamics of authority and self-reinforcement, as discussed above. According to Spielman [187], ideally, credibility and accuracy complement each other. This opens up an opportunity to introduce new metrics that are less dependent on the reputation of validators and better aligned with evolving mapping objectives.

When discussing the concept of intelligence, it is pertinent to consider opportunities to better harness the 'wisdom of crowds' [167]. We have identified a validation bottleneck linked to the relative scarcity of validators, as this task is mainly limited to the most experienced users. In addition, the decision-making process is always individual rather than collective. A conventional approach to address this problem is to increase the visibility of projects requiring validation to potential validators. However, a strategy more in line with the principle of collective intelligence would be to broaden the pool of collaborators authorized to validate, potentially including less experienced mappers. While novice mappers alone may lack the expertise to accurately evaluate a task, the collective input of a large number of novice users can yield reliable results. Voting or averaging mechanisms that have proven effective in other contexts could be explored in this case [167]. To this end, validation tasks could be further fragmented, making them more accessible to novice users. MapSwipe,[11] another micro-task platform within the OSM humanitarian ecosystem, has already implemented simple, fragmented validation. However, this functionality is still limited to pilot projects and should be extended to more HOT-TM projects to maximize its potential.

To conclude our reflection, we turn to the implications that our findings may have for the presumably increasing incorporation of artificial intelligence (AI) tools in humanitarian mapping tasks. By introducing the concept of collective intelligence, it is clear that computer agents are an integral part of this equation [152, 153]. Studies such as those by Tipnis et al. [193] on the introduction of Rapid show that while AI tools can improve productivity, they can also exacerbate differences in

[11] https://mapswipe.org/

participation between novice and experienced users, an outcome that does not necessarily align with fostering a more sustainable community.

In this context, our findings highlight several areas where AI can make a sustainable contribution. On the one hand, further productivity gains can be achieved by anticipating and mitigating events that slow down mapping workflows, especially in high-priority or more complex projects where such interruptions are more frequent. One such approach could be to predict split-prone areas and perform a preliminary split during project setup. In addition, novice users could be guided to complete tasks more efficiently by avoiding repeated self-unlocking and marking tasks as completed in a single cycle. In some cases, an extra minute of guidance for a mapper could speed up the completion of a task by a whole day. However, the greatest potential lies in the strategic use of AI to support equity-based goals. Initiatives such as gamified progression through task difficulty levels, personalized notifications, layered orchestration of task assignments with enrichment goals, and collective validation through aggregated contributions would be overwhelming for human managers to coordinate on their own. In the framework of computational participation in collective intelligence systems [152], the use of AI to implement these strategies would redefine the role of computational agents from mere tools to active assistants and managers.

4.7 Summary

This chapter illustrates the value of analyzing collaborative production environments through the lens of collective intelligence. Adopting this perspective fosters a more systemic understanding, revealing that the components of the framework are not independent but interdependent, continually influencing each other. For instance, the composition of a given group can influence initial decisions about task design and performance evaluation in crowdsourcing activities, which in turn can create incentives that reinforce the original composition of the group.

The collective intelligence framework has not only proven effective in coherently organizing the numerous quantitative results generated in this study but has also served as a unifying lens to synthesize evidence from previous notable research efforts in the field of VGI. The dispersed nature of these studies often makes it difficult for systemic perspectives to emerge. However, the main advantage of this framework lies in its ability to support the formulation of structural recommendations for improving system intelligence. As Malone et al. [152, 153] argue, understanding the factors that influence collective intelligence allows system managers to intervene more effectively. By adjusting the fundamental components of a system, managers can significantly improve their collective intelligence, an advantage not usually achieved with individual intelligence, which tends to be much less adaptive.

This case serves as an invitation to future researchers, designers and managers of crowdsourcing platforms to consider the advantages offered by this framework. The availability of such diagnostics is a valuable asset in this era of rapid and widespread adoption of artificial intelligence, a phenomenon that undoubtedly also affects computer-supported collaborative work initiatives. As discussed above, careless adoption of these technologies can undermine rather than foster productive collaboration.

With proper diagnosis it is possible to better guide interventions. In this sense, the basal literature on collective intelligence is rich in conceptual ideas [152, 153, 161]. However, this abundance of concepts is not always accompanied by detailed use cases that can guide practical application. In this sense, our study, while not aiming to establish a formal methodology, can provide a source of application ideas that can be adapted and replicated in other contexts. The key prerequisite is access to an equally rich source of data.

Data dependency serves as a call to crowdsourcing platform managers to follow the example of HOT in collecting comprehensive data on contributors, workflows and outcomes, as well as to share these data with relevant communities, ideally in open formats. While specific data fields and metrics may vary across crowdsourcing scenarios, this case study highlights the value of quantifying the components of a collective intelligence system. The general action lines of (1) describing relevant attributes of contributor profiles, (2) analyzing the mechanics of collective action through techniques such as process mining, and (3) tracking factors associated with successful or unsuccessful outcomes, can be broadly applied to a variety of contexts. Finally, the specific recommendations proposed in this study to improve the sustainability of the HOT-TM mapping community may also be valuable for other peer production environments, given that the uneven distribution of human activity in collaborative efforts seems to be the norm rather than the exception [192].

Adopting such an ambitious framework as collective intelligence offers significant advantages but it comes with inherent challenges. Efforts to address these challenges, however comprehensive, can often seem insufficient. Some limitations become more apparent and, rather than solving issues, may raise more questions. However, we see this as a constructive outcome. Below we set out several key limitations and threats to validity of our study and identify corresponding opportunities for future research.

Starting with group composition, our analysis, based on attribute profiling, was sufficient to provide valuable information. In addition, we consider that the mapping level classification system used by HOT-TM, although simple, is effective in identifying distinct user behaviors. However, this approach may be somewhat simplistic and overlook broader or more nuanced patterns of behavior. Future research could explore characterizations based on emergent group properties [198] or adopt mapper categorization frameworks [183].

When analyzing collective actions, it is important to keep in mind that HOT-TM operates on two interconnected levels: the OSM editor and the tasking manager. Our analysis focuses on what the states of the tasking manager can reveal about the behavior of the mapper. This focus is a defining characteristic of this type of mapping compared to other mapping approaches. However, we do not ignore the importance of object-level edit histories, which provide a more granular view of behaviors and actions. We see the exploration of these detailed operations and their relationship to tasking manager states as a valuable direction for future research. For example, it would be valuable to measure the impact of micro-task structure (particularly task boundaries) on spatial interactions, such as the alignment and connection of adjacent or continuous elements [186]. This consideration highlights two more general limitations of our study. The first is our reliance primarily on HOT-TM API data, which

were chosen for reasons of manageability. The incorporation of external data, such as OSM-level edit histories, could significantly enrich our analysis and provide a more complete understanding of mapping dynamics. The second limitation relates to the dominance of quantitative analysis. Qualitative research activities played a secondary role, serving to complement the interpretation and discussion of the results. Future research could deepen the understanding of HOT-TM mappers by employing qualitative approaches, such as those applied by Kogan et al. [186] or through usability testing to further explore the mapping experience within the task manager [73]. This approach could help clarify the nature of behaviors such as hesitation of beginners, shedding light on whether such behaviors stem more from intrinsic mapper factors or from the design of the mapping interface.

In assessing the intelligence of the system through its outputs, we have used validation as a credibility metric. However, this approach may not represent the most objective measure of system quality. In this regard, the extensive OSM literature on data quality offers opportunities to incorporate more objective indicators, such as positional accuracy of geographic data, attribute accuracy, data completeness and other dimensions of quality [190, 191]. Leveraging these indicators requires further exploration of the OSM database to collect additional data. Another important limitation of the study is that it focuses on the impact of collaboration on task outcomes. A broader and more insightful approach would involve examining how individual tasks and their integration contribute to the overall quality of the total project area.

Other considerations include the volatility of platforms like HOT-TM and OSM, which are constantly evolving in terms of policies and user interfaces. This means that certain behaviors we observed may change suddenly with new updates, making them no longer directly comparable to what we have documented here. Additionally, we prioritized the analysis of the mapping level of contributors because it was accessible to all users, while other demographic segmentation criteria, such as the location of the mappers, received secondary treatment due to their limited availability. Information about mapping levels on the HOT-TM API was restricted to the time of data retrieval. Consequently, a contributor identified as advanced or intermediate may have been a beginner during their participation in certain projects. To reduce this impact, we selected archived projects launched within a two-year period.

The case of HOT-TM must also be understood within the broader context of open data ecosystems. The platform not only coordinates volunteer contributions but also channels them into OpenStreetMap, where the data are openly available for reuse across humanitarian, governmental, and civic applications. This circular flow (from volunteers producing geospatial information, to its integration in a global open data infrastructure, to its uptake in disaster response and development initiatives) illustrates how collective intelligence contributes directly to the sustainability of open data. By analyzing the mechanisms that enable or constrain this flow, the chapter highlights the role of user interfaces not only as technical tools for mapping but also as mediating elements in the production, integration, and reuse of open geographic information.

This chapter has shown that collective intelligence in humanitarian mapping projects can be systematically evaluated by analyzing contributor profiles, task workflows, and patterns of collaboration

in the HOT Tasking Manager. This analysis allowed us to characterize how groups of mappers are composed, how they coordinate their efforts, and where evidence of intelligent action emerges. This directly answers the research question (RQ3) linked to this chapter by demonstrating that the proposed framework provides a rigorous approach for identifying and analyzing collective intelligence in open data ecosystems. While the findings are grounded in the HOT Tasking Manager, the methodological principles developed here are transferable to other platforms that rely on large-scale user contributions, thereby reinforcing the broader relevance of the approach.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Summary of contributions

This dissertation has investigated how the evaluation of user interfaces can contribute to the development of more sustainable and value-creating open data ecosystems. Anchored in the domain of geographic information, it advances an integrated and multi-level approach to interface evaluation that spans functional requirements, usability alignment, and collective user behavior. The work is motivated by the recognition that open data platforms are not merely technical infrastructures but socio-technical systems whose success depends on meaningful human interaction, collaborative dynamics, and long-term user engagement.

In response to the overarching research question (How can we systematically assess the design of user interfaces in open data portals to anticipate deployment failures and guide improvements that align with sustainability-oriented principles?) this thesis demonstrates that such assessment is feasible through a multi-layered methodological framework that integrates acceptance testing, usability evaluation, and collective intelligence analysis. Together, these approaches generate technical and human-centered evidence, allowing risks to be identified prior to deployment and guiding design improvements in line with principles of user-driven design, inclusivity, and circularity.

Chapter 2 laid the methodological foundation by proposing a structured framework for the acceptance testing of geospatial search engines, drawing from TMAP and aligning it with the specific demands of geospatial search interfaces. This directly addresses RQ1 by showing that existing software testing methodologies can be adapted to evaluate user interfaces effectively. Functionality, effectiveness, and user-friendliness were defined as central quality attributes, ensuring that evaluations extend beyond technical performance to address user expectations. The findings presented in this chapter have been accepted for publication in Environmental Modelling & Software. [199].

Building upon this foundation, Chapter 3 deepened the investigation by focusing on the cognitive dimension of usability. It addressed a key challenge left implicit in traditional testing frameworks: whether the conceptual model of the system, as envisioned by designers, aligns with the mental models formed by users during actual use. By extending usability testing with process mining techniques

and statistical analysis, this chapter directly answers RQ2, demonstrating that such methods can reveal mismatches between conceptual and mental models and thereby highlight inclusivity gaps across different levels of expertise. The findings of this chapter are published in the 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023) [73]. A journal article has also been prepared and submitted for review [200].

Chapter 4 expanded the evaluative perspective further to consider collective dynamics within open data ecosystems, with a focus on the Humanitarian OpenStreetMap Team Tasking Manager. Here, user interactions were treated not only as isolated episodes but as part of a broader pattern of collaborative engagement and knowledge production. By analyzing contributor profiles, task workflows, and collaboration patterns, the chapter provides a direct answer to RQ3, showing that the collective intelligence of users can be systematically evaluated through their circular interactions with open data ecosystems. The results highlight uneven participation, adaptive division of labor, and emergent coordination, all of which illustrate both the opportunities and challenges of sustaining collective intelligence. The findings of this chapter are published in the ACM Transactions on Computer-Human Interaction [201] and the Association of Geographic Information Laboratories in Europe (AGILE) 2024 conference [194].

Taken together, these chapters trace a trajectory from testing systems against predefined specifications, to understanding how users internalize and navigate those systems, and finally to examining how interfaces support or hinder group-level problem solving. Each step not only builds upon the last but also reveals new layers of interaction and complexity. Methodologically, the thesis combines structured software testing, cognitive modeling, and behavioral analytics to offer a comprehensive evaluation toolkit. Conceptually, it makes the case for treating open data portals not just as repositories or tools, but as interactive environments that require sustained evaluation, adaptation, and stewardship.

The contributions of this dissertation thus lie not only in the development of specific methods or case studies, but also in the articulation of a broader vision: that interface evaluation must be multi-dimensional and sustainability-oriented. In this view, functionality, usability, and collective intelligence are not isolated attributes but interdependent components of a successful open data ecosystem. Systems that fail to meet functional requirements may never reach users; those that neglect usability may alienate them; and those that ignore group dynamics may fail to deliver on the promises of participation, openness, and innovation.

The findings of this dissertation should be considered in light of both their internal and external validity. Internal validity has been supported through systematic research design, including the use of established testing frameworks, process mining techniques, and structured usability experiments that ensured methodological rigor and reproducibility. External validity relates to the transferability of the proposed methods beyond the specific case studies. While the empirical work has been carried out in the domain of geographic information systems (particularly the IGN geospatial search engine and the HOT Tasking Manager) the methodological contributions are broadly applicable to other open data ecosystems facing similar challenges. At the same time, the contextual characteristics of the case studies, such as domain-specific data types and community practices, should be acknowledged as

potential boundary conditions. Future research is therefore encouraged to extend these approaches to additional domains in order to further strengthen the generalizability of the results. Additionally, as noted in the limitations of each chapter, the evaluation of user interfaces must be particularly attentive to the potential influence of bias, which can arise from multiple sources, including the choice of interfaces, the selection of users, and, most critically, the perspectives of the evaluators themselves.

This PhD project was developed within the framework of the "ODECO: Towards a Sustainable Open Data ECOsystem" project. ODECO was a four-year Horizon 2020 Marie Skłodowska-Curie Innovative Training Network initiative (H2020-MSCA-ITN-2020, grant agreement 955569) whose main aim was to address current and future challenges in the creation of user-driven, circular and inclusive open data ecosystems. In addition to the original contributions presented in this dissertation, the project offered opportunities to collaborate on various academic and outreach deliverables related to open data ecosystems, user interfaces, and user experience. Among the relevant outcomes is a study on user participation in open government data initiatives, whose findings were published in IEEE Access [202] and presented at the 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022) [203]. Additionally, a journal article addressing the thematic annotation of open data has been prepared and submitted for review [204]. Other project outputs include several technical reports addressing key aspects of open data ecosystems, such as user needs, contributions from data users and non-governmental data providers, and strategies for long-term sustainability [7, 10, 13]. Dissemination efforts have included contributions to forums such as the XI Jornada de Jóvenes Investigadores del I3A in Spain (2022) [205], the XV Jornadas Ibéricas de Infraestructuras de Datos Espaciales in Spain (2024) [206], and the State of the Map Europe conference in Poland (2024) [207].

5.2 Work in progress

We are expanding our analysis of collective intelligence in humanitarian mapping by incorporating an objective assessment of data quality, an aspect that remained unaddressed in our previous study. Much of the research on crowdsourced geographic information has focused on participation, motivation, and community dynamics, while questions about the integrity of the final spatial products have received less systematic attention. In particular, there is still limited understanding of how emerging microtask-based production models, such as those widely used in humanitarian mapping, affect the coherence and reliability of outputs [151]. Conversely, the studies that do engage with quality issues often treat them in isolation from the sociotechnical systems and workflows that produce the data, which makes it difficult to link quality outcomes to underlying design choices and organizational practices [187].

Our ongoing work seeks to bridge this gap by explicitly examining the relationship between microtask-based workflows and the quality of the resulting geographic data. We hypothesize that the division of mapping into narrowly defined microtasks, while effective for mobilizing large numbers of volunteers, may unintentionally hinder the emergence of spatial collective intelligence. When contributions are fragmented into many small units, inconsistencies and integration problems are

more likely to occur once these units are recombined. Such problems, such as discontinuities along task boundaries, duplications of mapped features, or uneven levels of detail, can compromise both the internal coherence of the data and their value for decision-making in disaster preparedness and response.

To explore this hypothesis, we return to the case of HOT-TM, a platform that represents the microtasking model in humanitarian contexts. Building on our previous findings about the behavioral and organizational dimensions of mapping, we now extend the analysis to examine how task structure and contributor interactions translate into measurable outcomes in the data. By analyzing a large number of completed projects, we aim to move beyond localized observations and provide a systematic account of integration issues across diverse mapping scenarios.

Methodologically, our approach combines spatial statistical techniques with expert-based assessments to achieve a balanced view of quality. Spatial statistics make it possible to detect recurring patterns of misalignment or fragmentation at scale, while expert visual inspection provides contextual interpretation and validation. Through this mixed strategy, we intend to (i) identify the most common forms of integration problems that arise in humanitarian mapping, (ii) develop robust indicators that can measure these problems consistently, (iii) assess the prevalence of such problems across projects with varying geographies and objectives, and (iv) reflect on practical strategies for mitigating them, both at the level of platform design and project management.

The expected contribution of this work in progress is twofold. First, it enriches our understanding of spatial collective intelligence by demonstrating that data quality is not only a technical issue but also a property of the sociotechnical system through which contributions are organized. Second, it offers actionable insights for the humanitarian mapping community by pointing to ways in which task design, contributor guidance, and integration mechanisms can be adapted to produce more coherent and reliable data. In doing so, this research aims to inform both theory and practice: advancing collective intelligence studies with an empirical account of data integration challenges, while also equipping practitioners with knowledge to improve the sustainability and utility of crowdsourced geographic information.

5.3 Future work

Future research on the evaluation of user interfaces in open data ecosystems could build upon the directions outlined throughout the preceding chapters. In Chapter 2, which focused on a testing framework for geospatial search engines, it was suggested that future work should extend the proposed methodology with empirical studies conducted under real-world conditions. This includes further analysis of how user interface decisions affect long-term user engagement, trust, and the sustainability of open data ecosystems. There is also a need to refine testing strategies to better account for the dynamic nature of open data platforms, which are continuously updated and expanded.

Chapter 3, which applied usability testing and process mining to evaluate the geospatial search engine developed by the Spanish National Geographic Institute, highlighted the potential of expanding

evaluations to include a broader range of user types and search tasks. Future studies should explore the interaction patterns of users from different domains and levels of expertise to assess how mental models vary. Integrating behavioral analytics—such as eye tracking or clickstream data—may also enhance our understanding of user behavior and inform more adaptive interface designs.

In Chapter 4, which addressed user interfaces in crowdsourcing platforms such as the HOT Tasking Manager, it was recommended that the analytical framework be extended to other platforms that rely on microtask-based models. Incorporating longitudinal data could offer insights into user learning curves and retention patterns. Furthermore, future work should continue developing objective quality metrics that account for spatial complexity, contributor experience, and platform-level task design. These metrics can serve as feedback mechanisms not only for platform managers but also for contributors, ultimately improving mapping quality and engagement.

Building on these specific suggestions, several general directions can guide future work on evaluating user interfaces in open data ecosystems. First, evaluation frameworks should be multidimensional, integrating usability metrics, behavioral data, and data quality indicators to provide a holistic view of interface performance. Second, future studies should prioritize longitudinal and contextual evaluations that capture the evolution of user experience over time. Third, greater emphasis should be placed on inclusiveness by involving a wide spectrum of users, including those with limited technical backgrounds and individuals from underrepresented communities. Fourth, evaluations should explicitly connect interface design decisions to broader ecosystem goals, such as transparency, openness, and participatory governance. Finally, fostering reproducibility and open benchmarking through the publication of test datasets, protocols, and tools will enhance the comparability and impact of future research.

In sum, this dissertation underscores that the sustainability of open data ecosystems depends not only on the availability of data, but on the quality of interaction between users, systems, and—now more than ever—artificial agents. Thoughtfully designed and continuously evaluated user interfaces are key to unlocking the full societal potential of open and intelligent data infrastructures.

CHAPTER 6

CONCLUSIONES Y TRABAJO FUTURO

6.1 Resumen de contribuciones

Esta tesis ha investigado cómo la evaluación de interfaces de usuario puede contribuir al desarrollo de ecosistemas de datos abiertos más sostenibles y generadores de valor. Anclada en el dominio de la información geográfica, propone un enfoque integrado y multinivel para la evaluación de interfaces que abarca los requisitos funcionales, la alineación con la usabilidad y el comportamiento colectivo de los usuarios. El trabajo parte del reconocimiento de que las plataformas de datos abiertos no son meras infraestructuras técnicas, sino sistemas sociotécnicos cuyo éxito depende de una interacción humana significativa, dinámicas colaborativas y una participación sostenida de los usuarios a largo plazo.

En respuesta a la pregunta de investigación general —¿Cómo podemos evaluar sistemáticamente el diseño de las interfaces de usuario en los portales de datos abiertos para anticipar fallos de implementación y guiar mejoras que se alineen con principios orientados a la sostenibilidad?— esta tesis demuestra que dicha evaluación es factible mediante un marco metodológico multinivel que integra pruebas de aceptación, evaluación de usabilidad y análisis de inteligencia colectiva. En conjunto, estos enfoques generan evidencias técnicas y centradas en las personas, permitiendo identificar riesgos antes de la implementación y guiar mejoras de diseño en consonancia con principios de orientación al usuario, inclusividad y circularidad.

El Capítulo 2 sentó las bases metodológicas al proponer un marco estructurado para la prueba de aceptación de motores de búsqueda geoespaciales, tomando como referencia TMAP y alineándolo con las demandas específicas de las interfaces de búsqueda geoespacial. Esto responde directamente a la RQ1 al mostrar que las metodologías existentes de pruebas de software pueden adaptarse para evaluar eficazmente las interfaces de usuario. La funcionalidad, la efectividad y la facilidad de uso se definieron como atributos de calidad centrales, garantizando que las evaluaciones vayan más allá del rendimiento técnico para abordar las expectativas de los usuarios. Los resultados presentados en este capítulo han sido aceptados para su publicación en Environmental Modelling & Software. [199].

Sobre esta base, el Capítulo 3 profundizó la investigación centrándose en la dimensión cognitiva de la usabilidad. Abordó un desafío clave que quedaba implícito en los marcos de prueba tradicionales: si el modelo conceptual del sistema, tal como fue concebido por los diseñadores, se alinea con los modelos mentales formados por los usuarios durante el uso real. Al ampliar las pruebas de usabilidad con técnicas de minería de procesos y análisis estadístico, este capítulo responde directamente a la RQ2, demostrando que dichos métodos pueden revelar desajustes entre modelos conceptuales y modelos mentales y, por tanto, poner de manifiesto brechas de inclusividad en distintos niveles de experiencia. Los resultados de este capítulo se publicaron en la 22ª International Conference on Perspectives in Business Informatics Research (BIR 2023) [73]. Asimismo, se ha preparado y enviado un artículo a una revista para su revisión [200].

El Capítulo 4 amplió aún más la perspectiva evaluativa para considerar las dinámicas colectivas dentro de los ecosistemas de datos abiertos, con un enfoque en el Humanitarian OpenStreetMap Team Tasking Manager. Aquí, las interacciones de los usuarios se trataron no solo como episodios aislados, sino como parte de un patrón más amplio de compromiso colaborativo y producción de conocimiento. Al analizar los perfiles de los colaboradores, los flujos de trabajo de las tareas y los patrones de colaboración, el capítulo da una respuesta directa a la RQ3, mostrando que la inteligencia colectiva de los usuarios puede evaluarse sistemáticamente a través de sus interacciones circulares con los ecosistemas de datos abiertos. Los resultados destacan la participación desigual, la división adaptativa del trabajo y la coordinación emergente, todo lo cual ilustra tanto las oportunidades como los retos de sostener la inteligencia colectiva. Los hallazgos de este capítulo se publicaron en la revista ACM Transactions on Computer-Human Interaction [201] y en la conferencia de la Association of Geographic Information Laboratories in Europe (AGILE) 2024 [194].

En conjunto, estos capítulos trazan una trayectoria que va desde la evaluación de sistemas frente a especificaciones predefinidas, hasta la comprensión de cómo los usuarios interiorizan y navegan dichos sistemas, y finalmente a la evaluación de cómo las interfaces apoyan o dificultan la resolución colectiva de problemas. Cada paso no solo se construye sobre el anterior, sino que también revela nuevas capas de interacción y complejidad. Metodológicamente, la tesis combina pruebas estructuradas de software, modelado cognitivo y análisis de comportamiento para ofrecer un conjunto de herramientas de evaluación integral. Conceptualmente, argumenta que los portales de datos abiertos deben considerarse no solo como repositorios o herramientas, sino como entornos interactivos que requieren evaluación continua, adaptación y gestión.

Los resultados de esta tesis deben considerarse a la luz de su validez interna y externa. La validez interna se ha respaldado mediante un diseño de investigación sistemático, que incluyó el uso de marcos de prueba consolidados, técnicas de minería de procesos y experimentos de usabilidad estructurados que garantizaron el rigor metodológico y la reproducibilidad. La validez externa se relaciona con la transferibilidad de los métodos propuestos más allá de los estudios de caso específicos. Si bien el trabajo empírico se ha llevado a cabo en el ámbito de los sistemas de información geográfica, particularmente el buscador geoespacial del IGN y el HOT Tasking Manager, las contribuciones metodológicas son aplicables a otros ecosistemas de datos abiertos que enfrentan desafíos similares.

Al mismo tiempo, es importante reconocer las características contextuales de los casos de estudio, como los tipos de datos específicos del dominio y las prácticas comunitarias, que pueden constituir condiciones de frontera. Por ello, se recomienda que futuras investigaciones extiendan estos enfoques a otros dominios con el fin de reforzar aún más la generalización de los resultados.

Las contribuciones de esta tesis no se limitan al desarrollo de métodos específicos o estudios de caso, sino que también articulan una visión más amplia: la evaluación de interfaces debe ser multidimensional y orientada a la sostenibilidad. Desde esta perspectiva, la funcionalidad, la usabilidad y la inteligencia colectiva no son atributos aislados, sino componentes interdependientes de un ecosistema de datos abiertos exitoso. Sistemas que no cumplen con los requisitos funcionales pueden no llegar nunca a los usuarios; aquellos que descuidan la usabilidad pueden alejarlos; y aquellos que ignoran las dinámicas grupales pueden no cumplir con las promesas de participación, apertura e innovación.

Este proyecto de doctorado se desarrolló en el marco del proyecto "ODECO: Towards a Sustainable Open Data ECOsystem". ODECO fue una iniciativa de la red de formación innovadora Marie Skłodowska-Curie del programa Horizonte 2020 (H2020-MSCA-ITN-2020, acuerdo de subvención 955569), cuyo objetivo principal fue abordar los desafíos actuales y futuros en la creación de ecosistemas de datos abiertos impulsados por los usuarios, circulares e inclusivos. Además de las contribuciones originales presentadas en esta tesis, el proyecto brindó oportunidades de colaboración en diversos productos académicos y de divulgación relacionados con ecosistemas de datos abiertos, interfaces de usuario y experiencia de usuario. Entre los resultados relevantes se encuentra un estudio sobre la participación de usuarios en iniciativas de datos abiertos gubernamentales, cuyos hallazgos fueron publicados en IEEE Access [202] y presentados en la 26th International Conference on Theory and Practice of Digital Libraries (TPDL 2022) [203]. Asimismo, un artículo sobre anotación temática de datos abiertos ha sido preparado y enviado para su revisión [204]. Otros productos del proyecto incluyen varios informes técnicos que abordan aspectos clave de los ecosistemas de datos abiertos, como las necesidades de los usuarios, las contribuciones de usuarios de datos y proveedores no gubernamentales, y estrategias para la sostenibilidad a largo plazo [7, 10, 13]. Las actividades de difusión incluyeron contribuciones a foros como la XI Jornada de Jóvenes Investigadores del I3A en España (2022) [205], las XV Jornadas Ibéricas de Infraestructuras de Datos Espaciales en España (2024) [206] y la conferencia State of the Map Europe en Polonia (2024) [207].

6.2 Trabajo en curso

Estamos ampliando nuestro análisis de la inteligencia colectiva en el mapeo humanitario mediante la incorporación de una evaluación objetiva de la calidad de los datos, un aspecto que permaneció sin abordar en nuestro estudio anterior. Gran parte de la investigación sobre información geográfica generada por la multitud se ha centrado en la participación, la motivación y las dinámicas comunitarias, mientras que las preguntas sobre la integridad de los productos espaciales finales han recibido menos atención sistemática. En particular, aún existe un conocimiento limitado sobre cómo los modelos

emergentes de producción basados en microtareas, como los utilizados de manera generalizada en el mapeo humanitario, afectan la coherencia y la fiabilidad de los resultados [151]. Por el contrario, los estudios que sí abordan cuestiones de calidad suelen tratarlas de manera aislada respecto a los sistemas sociotécnicos y a los flujos de trabajo que producen los datos, lo que dificulta vincular los resultados de calidad con las decisiones de diseño y las prácticas organizativas subyacentes [187].

Nuestro trabajo actual busca reducir esta brecha examinando explícitamente la relación entre los flujos de trabajo basados en microtareas y la calidad de los datos geográficos resultantes. Partimos de la hipótesis de que la división del mapeo en microtareas estrechamente definidas, aunque eficaz para movilizar a un gran número de voluntarios, puede afectar negativamente la emergencia de inteligencia colectiva espacial. Cuando las contribuciones se fragmentan en múltiples unidades pequeñas, es más probable que aparezcan inconsistencias y problemas de integración al recombinarse dichas unidades. Problemas como discontinuidades en los límites de las tareas, duplicaciones de objetos mapeados o niveles de detalle desiguales pueden comprometer tanto la coherencia interna de los datos como su valor para la toma de decisiones en la preparación y respuesta ante desastres.

Para explorar esta hipótesis, retomamos el caso de HOT-TM, una plataforma que ejemplifica el modelo de microtareas en contextos humanitarios. Basándonos en hallazgos previos sobre las dimensiones conductuales y organizativas del mapeo, extendemos ahora el análisis para examinar cómo la estructura de las tareas y las interacciones de los colaboradores se traducen en resultados medibles en los datos. Al analizar un conjunto amplio de proyectos completados, buscamos superar observaciones localizadas y ofrecer un panorama sistemático de los problemas de integración en distintos escenarios de mapeo.

Metodológicamente, nuestro enfoque combina técnicas estadísticas espaciales con evaluaciones basadas en expertos para lograr una visión equilibrada de la calidad. Las estadísticas espaciales permiten detectar a gran escala patrones recurrentes de desalineación o fragmentación, mientras que la inspección visual de expertos aporta interpretación contextual y validación. Con esta estrategia mixta, pretendemos (i) identificar las formas más comunes de problemas de integración que surgen en el mapeo humanitario, (ii) desarrollar indicadores robustos que puedan medir dichos problemas de manera consistente, (iii) evaluar la prevalencia de estos problemas en proyectos con diferentes geografías y objetivos, y (iv) reflexionar sobre estrategias prácticas para mitigarlos, tanto a nivel de diseño de la plataforma como de gestión de proyectos.

La contribución esperada de este trabajo en progreso es doble. En primer lugar, enriquece nuestra comprensión de la inteligencia colectiva espacial al demostrar que la calidad de los datos no es únicamente un asunto técnico, sino también una propiedad del sistema sociotécnico a través del cual se organizan las contribuciones. En segundo lugar, ofrece información práctica para la comunidad de mapeo humanitario al señalar maneras en que el diseño de tareas, la orientación a los colaboradores y los mecanismos de integración pueden adaptarse para producir datos más coherentes y confiables. De este modo, esta investigación busca informar tanto a la teoría como a la práctica: por un lado, avanzando en los estudios de inteligencia colectiva con un análisis empírico de los retos de integración

de datos, y por otro, proporcionando a los profesionales conocimiento para mejorar la sostenibilidad y la utilidad de la información geográfica generada de manera colaborativa.

6.3 Trabajo futuro

La investigación futura sobre la evaluación de interfaces de usuario en ecosistemas de datos abiertos podría desarrollarse a partir de las líneas esbozadas a lo largo de los capítulos anteriores. En el Capítulo 2, centrado en un marco de pruebas para motores de búsqueda geoespacial, se sugirió que futuros trabajos extiendan la metodología propuesta mediante estudios empíricos en condiciones reales. Esto incluye un análisis más profundo de cómo las decisiones de diseño de interfaces afectan la participación sostenida de los usuarios, la confianza y la sostenibilidad del ecosistema. También se requiere refinar las estrategias de prueba para adaptarse mejor a la naturaleza dinámica de las plataformas de datos abiertos, que están en constante evolución.

El Capítulo 3, que aplicó pruebas de usabilidad y minería de procesos al motor de búsqueda desarrollado por el Instituto Geográfico Nacional de España, destacó el potencial de ampliar las evaluaciones para incluir una mayor diversidad de tipos de usuarios y tareas de búsqueda. Futuros estudios deberían explorar los patrones de interacción de usuarios de distintos dominios y niveles de experiencia para evaluar cómo varían los modelos mentales. La integración de técnicas de análisis del comportamiento —como el seguimiento ocular o los registros de clics— también podría enriquecer la comprensión del comportamiento del usuario e informar diseños de interfaz más adaptativos.

En el Capítulo 4, que abordó las interfaces de usuario en plataformas de crowdsourcing como HOT Tasking Manager, se recomendó ampliar el marco analítico a otras plataformas que dependen de modelos basados en microtareas. Incorporar datos longitudinales podría ofrecer información sobre las curvas de aprendizaje y los patrones de retención de usuarios. Además, el trabajo futuro debería continuar desarrollando métricas objetivas de calidad que consideren la complejidad espacial, la experiencia de los colaboradores y el diseño de tareas a nivel de plataforma. Estas métricas pueden servir como mecanismos de retroalimentación tanto para los gestores de plataformas como para los colaboradores, mejorando así la calidad del mapeo y la participación.

A partir de estas recomendaciones específicas, se pueden proponer varias líneas generales para orientar futuras investigaciones sobre la evaluación de interfaces en ecosistemas de datos abiertos. Primero, los marcos de evaluación deben ser multidimensionales, integrando métricas de usabilidad, datos de comportamiento e indicadores de calidad de los datos para ofrecer una visión integral del rendimiento de la interfaz. Segundo, los estudios futuros deberían priorizar evaluaciones longitudinales y contextuales que capten la evolución de la experiencia de usuario a lo largo del tiempo. Tercero, se debe poner mayor énfasis en la inclusión, involucrando a una amplia gama de usuarios, incluidos aquellos con conocimientos técnicos limitados y personas de comunidades subrepresentadas. Cuarto, las evaluaciones deben vincular explícitamente las decisiones de diseño de interfaces con los objetivos más amplios del ecosistema, como la transparencia, la apertura y la gobernanza participativa. Finalmente, fomentar la reproducibilidad y la comparación abierta mediante la publicación de

conjuntos de datos, protocolos y herramientas de prueba mejorará la comparabilidad e impacto de futuras investigaciones.

En resumen, esta tesis subraya que la sostenibilidad de los ecosistemas de datos abiertos depende no solo de la disponibilidad de datos, sino de la calidad de la interacción entre usuarios, sistemas y —cada vez más— agentes artificiales. Interfaces cuidadosamente diseñadas y evaluadas de manera continua son clave para liberar todo el potencial social de las infraestructuras de datos abiertas e inteligentes.

REFERENCES

[1] European Commission. "Digital Agenda: Commission's Open Data Strategy, Questions Answers". Technical report, European Commission, 2011.

[2] Esther Huyer and Laura van Knippenberg. The Economic Impact of Open Data Opportunities for Value Creation in Europe. Technical report, European Commission, 2020.

[3] Aditya Agrawal. Data Roadmaps for Sustainable Development. Assessment and Lessons Learned. Technical report, Global Partnership for Sustainable Development Data Secretariat, 2017.

[4] 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, Andrea Polini, Barbara Re, and Cesar Casiano Flores. 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.

[5] European Commission. A European Strategy for Data. Technical report, European Commission, 2020.

[6] Ingrid Mulder, Davide Di Staso, María Elena López Reyes, Georgios Papageorgiou, Alejandra Celis Vargas, Liubov Pilshchikova, Caterina Santoro, Héctor Ochoa Ortiz, Umair Ahmed, and Ashraf Shaharudin. Open data user needs: seven flavours. Deliverable D2.1, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, May 2023.

[7] Francisco J. Lopez-Pellicer, Abdul Aziz, Dagoberto José Herrera-Murillo, Mohsan Ali, and Maria Ioanna Maratsi. User Needs from a Technical Perspective. Deliverable D2.2, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, September 2023.

[8] Silvia Cazacu, Ramya Chandrasekhar, Mélanie Dulong de Rosnay, and Glenn Vancauwenberghe. User needs from a governance perspective. Deliverable D2.3, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, February 2024.

[9] Manolis Ktistakis, Davide Di Staso, Maria Elena Lopez Reyes, Giorgos Papageorgiou, Alejandra Celis Vargas, Liubov Pilshchikova, Caterina Santoro, Héctor Ochoa Ortiz, and Ashraf Shaharudin. Closing the cycle: Understanding potential contributions of open government data users to the open data ecosystem. Deliverable D3.1, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, November 2023.

[10] Andrea Polini, Umair Ahmed, Abdul Aziz, Dagoberto José Herrera-Murillo Herrera, and Mohsan Ali. Closing the Cycle: Promoting Open Data Users' Contribution from a Technical Perspective. Deliverable D3.2, Towards a Sustainable Open Data ECOsystem (ODECO) Project, Grant Agreement No. 955569, April 2024.

[11] Rikke Magnusse, Birger Larsen, Davide Di Staso, Silvia Cazacu-Bucica, Ramya Chandrasekhar, María Elena López Reyes, Giorgos Papageorgiou, Alejandra Celis Vargas, Liubov Pilshchikova, Caterina Santoro, Héctor Ochoa Ortiz, and Ashraf Shaharudin. Closing the cycle: Promoting open data users contributions from a governance perspective. Deliverable D3.3, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, May 2024.

[12] Barbara Re, Héctor Ochoa Ortiz, Ahmad Ashraf Ahmad Shaharudin, Davide Di Staso, Giorgos Papageorgiou, Alejandra Celis Vargas, and Liubov Pilshchikova. Motivations of non-government actors to become active contributors to the Open Data ecosystem. Deliverable D4.1, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, June 2024.

[13] Charalampos Alexopoulos, Maria Ioanna Maratsi, Mohsan Ali, Georgios Papageorgiou, Abdul Aziz, and Dagoberto José Herrera-Murillo. An Approach to Steer the Behaviour of Non-government Data Holders Towards Open Data Through a Technical Strategy. Deliverable D4.2, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, October 2024.

[14] Joep Crompvoets, Caterina Santoro, Ashraf Shaharudin, Davide Di Staso, Giorgos Papageorgiou, Alejandra Celis Vargas, Liubov Pilshchikova, and Héctor Ochoa Ortiz. An approach to steer the behaviour of non-government data holders towards open data through a governance strategy. Deliverable D4.3, Towards a Sustainable Open Data ECOsystem (ODECO), Grant Agreement No. 955569, September 2024.

[15] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated Quality Assessment of Metadata Across Open Data Portals. Journal of Data and Information Quality, 8(1), 2016.

[16] Wendy Carrara, Wae San Chan, Sander Fischer, and Eva van Steenbergen. Creating Value Through Open Data: Study on the Impact of Re-use of Public Data Resources. Technical report, European Commission, 2015.

[17] Donald A. Norman. The Design of Everyday Things. Basic Books, New York, revised and expanded edition, 2013.

[18] Ben Shneiderman and Catherine Plaisant. Designing the User Interface: Strategies for Effective Human-Computer Interaction. Addison-Wesley, Boston, 5th edition, 2010.

[19] Michael McTear. The Conversational Interface: Talking to Smart Devices. Springer, Cham, 2016.

[20] Susanne Bødker. Third-wave hci, 10 years later—participation and sharing. Interactions, 22(5):24–31, 2015.

[21] Douglas D. Nebert, editor. Developing Spatial Data Infrastructures: the SDI Cookbook. Global Spatial Data Infrastructure (GSDI), 2004.

[22] Paolo Corti, Athanasios Tom Kralidis, and Benjamin Lewis. Enhancing Discovery in Spatial Data Infrastructures Using a Search Engine. PeerJ. Computer science, 4, 2018.

[23] Javier Lacasta, Francisco Javier Lopez-Pellicer, Javier Zarazaga-Soria, Rubén Béjar, and Javier Nogueras-Iso. Approaches for the Clustering of Geographic Metadata and the Automatic Detection of Quasi-spatial Dataset Series. ISPRS International Journal of Geo-Information, 11(2):87, 2022.

[24] Sergio Martin-Segura, Francisco Javier Lopez-Pellicer, Javier Nogueras-Iso, Javier Lacasta, and Francisco Javier Zarazaga-Soria. The Problem of Reference Rot in Spatial Metadata Catalogues. ISPRS International Journal of Geo-Information, 11(1):27, 2022.

[25] OpenStreetMap Wiki. Stats - OpenStreetMap Wiki, 2025.

[26] Tim Koomen, Leo van der Aalst, Bart Broekman, and Michiel Vroon. TMap® Next for Result-driven Testing. UTN Publishers, Willem van Oranjelaan 5 5211 CN 's-Hertogenbosch The Netherlands, 2 edition, 2007.

[27] IEEE. IEEE 610.12-1990—IEEE Standard Glossary of Software Engineering Terminology. https://standards.ieee.org/standard/610_12-1990.html, 1990.

[28] Kai Xu, Min Chen, Songshan Yue, Fengyuan Zhang, Jin Wang, Yongning Wen, and Guonian Lü. The Portal of OpenGMS: Bridging the Contributors and Users of Geographic Simulation Resources. Environmental Modelling Software, 180, 2024.

[29] Miguel Ángel Latre, Francisco J. Lopez-Pellicer, Javier Nogueras-Iso, Rubén Béjar, F. Javier Zarazaga-Soria, and Pedro R. Muro-Medrano. Spatial Data Infrastructures for Environmental E-government Services: the Case of Water Abstractions Authorisations. Environmental Modelling Software, 48:81–92, 2013.

[30] Stefan Wiemann, Johannes Brauner, Pierre Karrasch, Daniel Henzen, and Lars Bernard. Design and Prototype of an Interoperable Online Air Quality Information System. Environmental Modelling Software, 79:354–366, 2016.

[31] Ziheng Sun, Liping Di, and Juozas Gaigalas. SUIS: Simplify the Use of Geospatial Web Services in Environmental Modelling. Environmental Modelling Software, 119:228–241, 2019.

[32] Ekaterina Galimova. Features of Software Testing in the Development of Geographic Information Systems. volume 177, page 02008. E3S Web of Conferences, EDP Sciences, 2020.

[33] Trismayanti Dwi Puspitasari, Arvita Agus Kurniasari, and Pramuditha Shinta Dewi Puspitasari. Analysis and Testing Using Boundary Value Analysis Methods for Geographic Information System. IOP Conference Series: Earth and Environmental Science, 1168(1), 2023.

[34] Mohsen Kalantari, Syahrudin Syahrudin, Abbas Rajabifard, and Hannah Hubbard. Synchronising Spatial Metadata Records and Interfaces to Improve the Usability of Metadata Systems. ISPRS International Journal of Geo-Information, 10(6):393, 2021.

[35] Tymoteusz Horbiński, Paweł Cybulski, and Beata Medyńska-Gulij. Web Map Effectiveness in the Responsive Context of the Graphical User Interface. ISPRS International Journal of Geo-Information, 10(3):134, 2021.

[36] Bénédicte Bucher, Erwin Folmer, Rob Brennan, Wouter Beek, Elio Hbeich, Falk Würriehausen, Lexi Rowland, Ricardo Alonso Maturana, Elena Alvarado, Raf Buyle, and Pasquale Di Donato, editors. Spatial Linked Data in Europe: Report from Spatial Linked Data Sessions at Knowledge Graph in Action, chapter Linked Cartography and Maps: the National Geographic Institute of Spain knowledge graph and its semantic web for the public (Spain), pages 17–18. Official Publication - EuroSDR, 2021. http://www.eurosdr.net/sites/default/files/uploaded_files/eurosdr_publication_ndeg_73.pdf.

[37] Javier Jesús Gutiérrez, María José Escalona, Manuel Mejías, and Jesús Torres. Generation of Test Cases from Functional Requirements. a Survey. pages 1–10. Proceedings 4rd Workshop on System Testing and Validation, 2006.

[38] María José Escalona, Javier Jesús Gutierrez, Manuel Mejías, Gustavo Aragón, Isabel Ramos, Jesús Torres, and Francisco José Domínguez. An Overview on Test Generation from Functional Requirements. Journal of Systems and Software, 84(8):1379–1393, August 2011.

[39] Gianni Pucciani. Boozang from the Trenches: Learn Test Automation with Boozang in an Enterprise Environment, chapter Gherkin and Behavior Driven Development, pages 217–224. Springer, 2022.

[40] Tanja Vos, Pekka Aho, Fernando Pastor Ricós, Olivia Rodriguez-Valdes, and Ad Mulders. TESTAR – Scriptless Testing Through Graphical User Interface. Software Testing, Verification and Reliability, 31, 05 2021.

[41] Julián Alberto García-García, Manuel Alba Ortega, Laura García-Borgoñon, and Maria Jose Escalona. NDT-Suite: a Model-Based Suite for the Application of NDT. In 12th International Conference on Web Engineering. Lecture Notes in Computer Science, pages 469–472. Springer Berlin Heidelberg, 2012.

[42] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schtze. Relevance Feedback and Query Expansion. Introduction to Information Retrieval. Cambridge University Press, New York, 2008.

[43] Zhi Quan Zhou, Shujia Zhang, Markus Hagenbuchner, T.H. Tse, Fei-Ching Kuo, and Tsong Chen. Automated Functional Testing of Online Search Services. Software Testing, Verification and Reliability, 22:221 – 243, 06 2012.

[44] Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. Clinical Information Retrieval: a Literature Review. Journal of Healthcare Informatics Research, 2024.

[45] Shrestha Ghosh, Simon Razniewski, and Gerhard Weikum. Answering Count Queries with Explanatory Evidence. page 2415–2419. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022.

[46] Eduardo Gabriel Cortes, Vinicius Woloszyn, Dante Barone, Sebastian Möller, and Renata Vieira. A Systematic Review of Question Answering Systems for Non-factoid Questions. Journal of Intelligent Information Systems, 58(3):453–480, 2022.

[47] Marti Hearst. Search User Interfaces. Cambridge University Press, 2009.

[48] Ben Shneiderman and Catherine Plaisant. Strategies for Evaluating Information Visualization Tools: Multi-dimensional In-depth Long-term Case Studies. Proceedings of the 2006 conference Advanced Visual Interfaces (AVI'04), 2006.

[49] Rebecca C. Vandewalle, William C. Barley, Anand Padmanabhan, Daniel S. Katz, and Shaowen Wang. Understanding the Multifaceted Geospatial Software Ecosystem: a Survey Approach. International Journal of Geographical Information Science, 35(11):2168–2186, 2021.

[50] Stanislav Popelka, Lukáš Herman, Tomas Řezník, Michaela Pařilová, Karel Jedlička, Jiří Bouchal, Michal Kepka, and Karel Charvát. User evaluation of map-based visual analytic tools. ISPRS International Journal of Geo-Information, 8(8), 2019.

[51] Erik P.W.M. van Veenendaal. Building on Success – Beyond the Obvious: a Closer Look at Good Enough Testing. page 91–92. Proceedings of the Federated Africa and Middle East Conference on Software Engineering, Association for Computing Machinery, 2022.

[52] Vuk Vukovic, Jovica Djurkovic, and Jelica Trninic. A Business Software Testing Process-Based Model Design. International Journal of Software Engineering and Knowledge Engineering, 28(05):701–749, 2018.

[53] Maarten van Banerveld, Mohand-Tahar Kechadi, and Nhien-An Le-Khac. A Natural Language Processing Tool for White Collar Crime Investigation. In Abdelkader Hameurlain, Josef Küng, Roland Wagner, Tran Khanh Dang, and Nam Thoai, editors, Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIII: Selected Papers from FDSE 2014, pages 1–22. Springer, 2016.

[54] Stuart Wrigley, Dorothee Reinhard, Khadija Elbedweihy, Abraham Bernstein, and Fabio Ciravegna. Methodology and Campaign Design for the Evaluation of Semantic Search Tools. Proceedings of the Semantic Search 2010 Workshop, 2010.

[55] Zhi Quan Zhou, Shaowen Xiang, and Tsong Chen. Metamorphic Testing for Software Quality Assessment: a Study of Search Engines. IEEE Transactions on Software Engineering, 42(3):264–284, 2015.

[56] ISO/IEC/IEEE. ISO/IEC/IEEE 29119-2:2021 Software and Systems Engineering — Software Testing — Part 2: Test Processes. https://www.iso.org/standard/79430.html, 2021.

[57] TMMi Foundation. Test Maturity Model Integration(TMMi). Guidelines for Test Process Improvement. Release 1.3. https://www.tmmi.org/tmmi-documents/, 2022.

[58] ISO/IEC. ISO/IEC 25010:2011 Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — System and Software Quality Models. https://www.iso.org/standard/35733.html, 2011.

[59] ISO/IEC/IEEE. ISO/IEC/IEEE 29119-4:2021 Software and Systems Engineering — Software Testing — Part 4: Test Techniques. https://www.iso.org/standard/79430.html, 2021.

[60] Srinivasan Desikan and Gopalaswamy Ramesh. Software Testing. Pearson Education India, 2006.

[61] Anne Mette Jonassen. Guide to Advanced Software Testing. Artech House, 2008.

[62] Jens Engel, Benno Rice, and Richard Jones. Welcome to Behave! https://behave.readthedocs.io/en/latest/, 2023.

[63] Boni García, Micael Gallego, Francisco Gortázar, and Mario Munoz-Organero. A Survey of the Selenium Ecosystem. Electronics, 9(7), 2020.

[64] ISO/IEC/IEEE. ISO/IEC/IEEE 29119-1:2022 Software and Systems Engineering - Software Testing - Part 1: General Concepts. https://www.iso.org/standard/83636.html, 2022.

[65] Jakob Nielsen. Usability Engineering. Academic Press Professional, 1993.

[66] Jakob Nielsen. Finding Usability Problems Through Heuristic Evaluation. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, page 373–380, 1992.

[67] Thomas Mahatody, Mouldi Sagar, and Christophe Kolski. State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. International Journal of Human–Computer Interaction, 26(8):741–785, 2010.

[68] Kate Moran. Usability Testing 101. https://www.nngroup.com/articles/usability-testing-101/, 2019.

[69] Khadija Elbedweihy, Stuart Wrigley, Paul Clough, and Fabio Ciravegna. An Overview of Semantic Search Evaluation Initiatives. Journal of Web Semantics, 30:82–105, 2015.

[70] John Brooke. Usability Evaluation in Industry, chapter SUS: a quick and dirty usability scale, pages 189–194. Taylor and Francis, 1996.

[71] K. Latha. Experiment and Evaluation in Information Retrieval Models. Chapman and Hall/CRC, 2017.

[72] Aaron Bangor, Philip Kortum, and James Miller. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies, 4(3):114–123, 2009.

[73] Dagoberto José Herrera-Murillo, Javier Nogueras-Iso, Paloma Abad-Power, and Francisco J. Lopez-Pellicer. User Interaction Mining: Discovering the Gap Between the Conceptual Model of a Geospatial Search Engine and Its Corresponding User Mental Model. pages 3–15. Perspectives in Business Informatics Research, Springer, 2023.

[74] Yavuz Köroğlu and Alper Sen. Functional Test Generation from UI Test Scenarios Using Reinforcement Learning for Android Applications. Software Testing, Verification and Reliability, 31, 10 2020.

[75] Kathleen Gregory and Laura Koesten. Discovering Data, pages 33–48. Springer, CH, 2022.

[76] Toshihisa Doi. Mental Model Formation in Users with High and Low Comprehension of a Graphical User Interface. Journal of Human Ergology, 48(1):9–24, 2019.

[77] Courtney Titus, Mary Gordon, Krisanne Graves, and Curt Braun. Getting the Complete Picture: Using Surveys as Complementary Method for Assessing Usability. In Tareq Z. Ahram and Christianne Falcão, editors, Advances in Usability, User Experience and Assistive Technology, pages 197–203, Cham, 2019. Springer International Publishing.

[78] Gang Wang, Xinyi Zhang, Shiliang Tang, Christo Wilson, Haitao Zheng, and Ben Y. Zhao. Clickstream User Behavior Models. ACM Transactions on the Web, 11(4), 2017.

[79] Ennio Visconti, Christos Tsigkanos, and Laura Nenzi. Automated Monitoring of Web User Interfaces. ACM Transactions on the Web, 19(2), 2025.

[80] Raphael Menges, Steffen Staab, Christoph Schaefer, Tina Walber, and Chandan Kumar. What Did My Users Experience? Discovering Visual Stimuli on Graphical User Interfaces of the Web. ACM Transactions on the Web, 19(2), 2025.

[81] Renat Faizrakhmanov, Mohammad Reza Bahrami, and A.E. Platunov. Prototype, Method, and Experiment for Evaluating Usability of Smart Home User Interfaces. Computer Standards and Interfaces, 92:103903, 2025.

[82] Imen Benzarti, Hafedh Mili, Renata Medeiros de Carvalho, and Abderrahmane Leshob. Domain Engineering for Customer Experience Management. Innovations in Systems and Software Engineering, 18:171 – 191, 2022.

[83] Ergonomics of Human-system Interaction — Part 11: Usability: Definitions and Concepts. Standard, International Organization for Standardization, Geneva, CH, 2018.

[84] Sebastian Möller. Usability Engineering, pages 55–72. Springer, DE, 2023.

[85] Joseph S. Dumas and Marilyn C. Salzman. Usability Assessment Methods. Reviews of Human Factors and Ergonomics, 2:109–140, 2006.

[86] Lauren E. Snyder, Rebecca Hazen, and Amanda K. Hall. What Is the New Future of Work for UX Research Practice? Assessing UX Research Practice During the COVID-19 Pandemic and Beyond. In Aaron Marcus, Elizabeth Rosenzweig, and Marcelo M. Soares, editors, Design, User Experience, and Usability, pages 324–342, CH, 2023. Springer.

[87] Tasha Hollingsed and David G. Novick. Usability Inspection Methods After 15 Years of Research and Practice. In ACM International Conference on Design of Communication, 2007.

[88] Jakob Nielsen. Usability Engineering. AP Professional, USA, 1993.

[89] Ted Boren and Judith Ramey. Thinking Aloud: Reconciling Theory and Practice. IEEE Transactions on Professional Communication, 43(3):261–278, 2000.

[90] Inge De Bleecker and Rebecca Okoroji. Remote Usability Testing: Actionable Insights in User Behavior Across Geographies and Time Zones. Packt Publishing, UK, 2018.

[91] Emily Geisen and Jennifer Romano Bergstrom. Usability Testing for Survey Research. Morgan Kaufmann, USA, 2017.

[92] Carol M. Barnum. Usability Testing Essentials: Ready, Set...Test! Morgan Kaufmann Publishers Inc., USA, 2010.

[93] Jeffrey Rubin and Dana Chisnell. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. John Wiley & Sons, USA, 2011.

[94] Morten Hertzum. Usability Testing: a Practitioner's Guide to Evaluating the User Experience. Synthesis Lectures on Human-Centered Informatics, 1:i–105, 2020.

[95] Usability of Consumer Products and Products for Public Use — Part 2: Summative Test Method. Standard, International Organization for Standardization, Geneva, CH, 2013.

[96] Mathew Stange, Amanda Barry, Jolene Smyth, and Kristen Olson. Effects of Smiley Face Scales on Visual Processing of Satisfaction Questions in Web Surveys. Social Science Computer Review, 36(6):756–766.

[97] John Brooke. Usability Evaluation in Industry, chapter SUS: a quick and dirty usability scale. Taylor and Francis, UK, 1996.

[98] René Unrau and Christian Kray. Mining map interaction semantics in web-based geographic information systems (WebGIS) for usability analysis. AGILE: GIScience Series, 2:16, 2021.

[99] Dedy Kurniawan, Dwi Indah, Purwita Sari, and Rahmat Alif. Understanding the Landscape of Usability Evaluation in Geographic Information Systems: a Systematic Literature Review. Journal of Applied Science, Engineering, Technology, and Education, 5:35–45, 2023.

[100] Joshua Sterling Wells, Robert Grant, John Chang, and Reem Kayyali. Evaluating the Usability and Acceptability of a Geographical Information System (GIS) Prototype to Visualise Socioeconomic and Public Health Data. BMC Public Health, 21, 2021.

[101] Valéria Oliveira Henrique de Araújo, Moema José de Carvalho Augusto, Hesley da Silva Py, and Raquel A. Abrahão Costa e Oliveira. The Usability of the National Spatial Data Infrastructure (INDE) Geoportal. In Proceedings of the 27th of International Cartographic Conference Brazil, 2015.

[102] Lei Kristoffer R. Lactuan, Danilo J. Mercado, and Jaime M. Samaniego. LabGIS a real estate property geographical information system (gis) for local government units. pages 14–26, 2019.

[103] Wil Van der Aalst. Process Mining: Data Science in Action. Springer, NL, 2016.

[104] Wil Van der Aalst. Using Process Mining to Bridge the Gap Between BI and BPM. Computer, 44:77–80, 2011.

[105] Guangming Li, Renata Medeiros de Carvalho, and Wil M. P. van der Aalst. Object-Centric Behavioral Constraint Models: a Hybrid Model for Behavioral and Data Perspectives. SAC '19, page 48–56, New York, NY, USA, 2019. Association for Computing Machinery.

[106] Wil Van der Aalst, M. de Leoni, and A.H.M. ter Hofstede. Process Mining and Visual Analytics: Breathing Life into Business Process Models, pages 107–138. Computer Science, Technology and Applications. Nova Publishers, USA, 2012.

[107] Tom Thaler. Towards Usability Mining. In GI-Jahrestagung, 2014.

[108] Sharam Dadashnia, Constantin Houy, and Peter Loos. Usability Mining. In Jan vom Brocke, Alan Hevner, and Alexander Maedche, editors, Design Science Research. Cases, pages 155–176. Springer, CH, 2020.

[109] William Rouse and Nancy Morris. On Looking into the Black Box. Prospects and Limits in the Search for Mental Models. Psychological Bulletin, 100(3):349–363, 1984.

[110] Nancy Staggers and Anthony F. Norcio. Mental Models: Concepts for Human-Computer Interaction Research. International Journal of Man-Machine Studies, 38(4):587–605, 1993.

[111] Robert Andrews, J. Lilly, Divya Srivastava, and Karen Feigh. The Role of Shared Mental Models in human-AI Teams: a Theoretical Review. Theoretical Issues in Ergonomics Science, 24:1–47, 2022.

[112] John M. Carroll and Judith Reitman Olson. Mental Models in Human-computer Interaction: Research Issues About What the User of Software Knows. In Handbook of Human-Computer Interaction, pages 45–65. North-Holland, USA, 1988.

[113] Xiaofan Qian, Ying Yang, and Yong Gong. The Art of Metaphor: a Method for Interface Design Based on Mental Models. In Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry, page 171–178, USA, 2011. Association for Computing Machinery.

[114] Johan De Kleer and John Seely Brown. Assumptions and Ambiguities in Mechanistic Mental Models. In Mental Models, pages 155–190. Psychology Press, USA, 1983.

[115] Lilian Crum. Laws of UX: Using Psychology to Design Better Products & Services. Design and Culture, 12(3):357–359, 2020.

[116] Yan Zhang. Undergraduate Students' Mental Models of the Web as an Information Retrieval System. Journal of the American Society for Information Science and Technology, 59(13):2087–2098, 2008.

[117] Marina Papastergiou. Students' Mental Models of the Internet and Their Didactical Exploitation in Informatics Education. Education and Information Technologies, 10:341–360, 2005.

[118] Efthimis Efthimiadis and David Hendry. Search Engines and How Students Think They Work. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 595–596, 2005.

[119] Sandra P. Roth, Peter Schmutz, Stefan L. Pauwels, Javier A. Bargas-Avila, and Klaus Opwis. Mental Models for Web Objects: Where Do Users Expect to Find the Most Frequent Objects in Online Shops, News Portals, and Company Web Pages? Interacting with Computers, 22(2):140–152, 2010.

[120] Xinhui Hu and Michael Twidale. A Scoping Review of Mental Model Research in HCI from 2010 to 2021. In HCI International 2023 – Late Breaking Papers: 25th International Conference on Human-Computer Interaction, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I, page 101–125, 2023.

[121] Marti Hearst. Search User Interfaces. Cambridge University Press, USA, 2009.

[122] Gerard Salton. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, USA, 1989.

[123] Gary Marchionini. Information-seeking Strategies of Novices Using a Full-text Electronic Encyclopedia. Journal of the Association for Information Science and Technology, 40:54–66, 1989.

[124] Alistair G. Sutcliffe and Mark Ennis. Towards a Cognitive Theory of Information Retrieval. Interacting with Computers, 10:321–351, 1998.

[125] Ben Shneiderman, Donald Byrd, and W. Bruce Croft. Clarifying Search: a User-Interface Framework for Text Searches. D-Lib Magazine, 3, 1997.

[126] Gary Marchionini and Ryen White. Find What You Need, Understand What You Find. International Journal of Human–Computer Interaction, 23:205–237, 12 2007.

[127] Vicki L. O'Day and Robin Jeffries. Orienteering in an Information Landscape: How Information Seekers Get from Here to There. In Proceedings of the INTERCHI '93 Conference on Human Factors in Computing Systems, page 438–445, 1993.

[128] Marcia J. Bates. The Design of Browsing and Berrypicking Techniques for the Online Search Interface. Online Review, 13:407–424, 1989.

[129] Michael Khoo and Catherine Hall. What Would 'Google' Do? Users' Mental Models of a Digital Library Search Engine. In International Conference on Theory and Practice of Digital Libraries, 2012.

[130] Yan Zhang. The Development of Users' Mental Models of MedlinePlus in Information Searching. Library & Information Science Research, 35(2):159–170, 2013.

[131] Mengtian Guo, Zhilan Zhou, David Gotz, and Yue Wang. GRAFS: Graphical Faceted Search System to Support Conceptual Understanding in Exploratory Search. ACM Transactions on Interactive Intelligent Systems, 13(2), 2023.

[132] Gaganpreet Sharma. Pros and Cons of Different Sampling Techniques. International Journal of Applied Research, 3:749–752, 2017.

[133] Jiapei Ren, Haiyan Wang, and Junkai Shao. Experimental Study on Dynamic Map Information Layout Based on Eye Tracking. In Tareq Ahram, Waldemar Karwowski, Alberto Vergnano, Francesco Leali, and Redha Taiar, editors, Intelligent Human Systems Integration 2020, pages 1238–1243, CH, 2020. Springer.

[134] Christian W. Günther and Anne Rozinat. Disco: Discover Your Processes. In International Conference on Business Process Management, pages 40–44, 2012.

[135] Wil Van der Aalst. Foundations of Process Discovery, pages 37–75. Springer, CH, 2022.

[136] Hugh Jack. Chapter 9 - Universal Design Topics. In Hugh Jack, editor, Engineering Design, Planning, and Management, pages 323–380. Academic Press, USA, 2013.

[137] Anna Weigand and Maria Rauschenberger. Exploring the Definition of Small Data Collected with HCI Methods and Used for ML. In Mensch Und Computer 2023, Workshop on User-Centered Artificial Intelligence, pages 1–4, 2023.

[138] Robert A. Virzi. Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? Human Factors: The Journal of Human Factors and Ergonomics Society, 34:457–468, 1992.

[139] Laura Faulkner. Beyond the Five-user Assumption: Benefits of Increased Sample Sizes in Usability Testing. Behavior Research Methods, Instruments, & Computers, 35:379–383, 2003.

[140] Wonil Hwang and Gavriel Salvendy. Number of People Required for Usability Evaluation: the 10±2 Rule. Communications of the ACM, 53(5):130–133, 2010.

[141] Roobaea Alroobaea and Pam J. Mayhew. How Many Participants Are Really Enough for Usability Studies? In 2014 Science and Information Conference, pages 48–56, 2014.

[142] Josep Carmona, Boudewijn F. van Dongen, Andreas Solti, and Matthias Weidlich. Conformance Checking - Relating Processes and Models. Springer, CH, 2018.

[143] Thomas W. Malone and Michael S. Bernstein. Handbook of Collective Intelligence. MIT press, 2022.

[144] David Engel, Anita Williams Woolley, Ishani Aggarwal, Christopher F. Chabris, Masamichi Takahashi, Keiichi Nemoto, Carolin Kaiser, Young Ji Kim, and Thomas W. Malone. Collective Intelligence in Computer-Mediated Collaboration Emerges in Different Contexts and Cultures. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, page 3769–3778, New York, NY, USA, 2015. Association for Computing Machinery.

[145] Jeff Howe. The Rise of Crowdsourcing. Wired Magazine, 14(6), 2006.

[146] Daren C. Brabham. Crowdsourcing. The MIT Press, 2013.

[147] Michael F. Goodchild. Citizens as Sensors: the World of Volunteered Geography. GeoJournal, 69(4):211–221, 2007.

[148] Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. The Evolution of Humanitarian Mapping Within the Openstreetmap Community. Scientific Reports, 11, 2021.

[149] Lucia Saganeiti, Federico Amato, Beniamino Murgante, and Gabriele Nolè. VGI and Crisis Mapping in an Emergency Situation. Comparison of Four Case Studies: Haiti, Kibera, Kathmandu, Centre Italy. GEOmedia, 21(3), August 2017. Number: 3.

[150] Pascal Neis, Peter Singler, and Alexander Zipf. Collaborative Mapping and Emergency Routing for Disaster Logistics - Case Studies from the Haiti Earthquake and the UN Portal for Afrika. January 2010.

[151] Yaxuan Yin, Longjie Guo, and Jacob Thebault-Spieker. Productivity or Equity? Tradeoffs in Volunteer Microtasking in Humanitarian OpenStreetMap. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW1), apr 2024.

[152] Thomas W. Malone. Superminds : the Surprising Power of People and Computers Thinking Together. Little, Brown and Company, New York, first edition. edition, 2018.

[153] Thomas W. Malone, Robert Laubacher, and Chrysanthos Dellarocas. The Collective Intelligence Genome. Sloan Management Review, 51(3):21–31, 2010.

[154] Suzanne T. Bell, Shanique G. Brown, Anthony Colaneri, and Neal B. Outland. Team Composition and the ABCs of Teamwork. American Psychologist, 73:349–362, 2018.

[155] Richard Moreland, John Levine, and Melissa Wingert. Creating the Ideal Group: Composition Effects at Work, pages 11–35. Lawrence Erlbaum Associates, Inc., 2018.

[156] Yuanyuan Jiao, Yepeng Wu, and Steven Lu. The Role of Crowdsourcing in Product Design: the Moderating Effect of User Expertise and Network Connectivity. Technology in Society, 64, 2021.

[157] Yuyan Han and David Dunning. Metaknowledge of Experts Versus Nonexperts: Do Experts Know Better What They Do and Do Not Know? Journal of Behavioral Decision Making, 37(2), 2024.

[158] Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R. Jennings. Efficient Crowdsourcing of Unknown Experts Using Bounded Multi-Armed Bandits. Artificial Intelligence, 214:89–111, 2014.

[159] Amal Ben Rjab, Mouloud Kharoune, Zoltan Miklos, and Arnaud Martin. Characterization of Experts in Crowdsourcing Platforms. In Jiřina Vejnarová and Václav Kratochvíl, editors, Belief Functions: Theory and Applications, pages 97–104, Cham, 2016. Springer International Publishing.

[160] Jean-Christophe Dubois, Laetitia Gros, Mouloud Kharoune, Yolande Le Gall, Arnaud Martin, Zoltan Miklos, and Hosna Ouni. Measuring the Expertise of Workers for Crowdsourcing Applications, pages 139–157. Springer International Publishing, Cham, 2019.

[161] Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. Frameworks for Collective Intelligence: a Systematic Literature Review. 53(1), 2020.

[162] David W. McDonald. Task Dependency and the Organization of the Crowd. In ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '11, 2011.

[163] Kjeld Schmidt and Liam Bannon. Taking CSCW Seriously: Supporting Articulation Work. Computer Supported Cooperative Work, 1:7–40, 03 1992.

[164] Oliver Stiemerling and Armin B. Cremers. The Use of Cooperation Scenarios in the Design and Evaluation of a CSCW System. IEEE Transactions on Software Engineering, 24(12):1171–1181, 1998.

[165] Christoph Riedl, Young Ji Kim, Pranav Gupta, Thomas W. Malone, and Anita Williams Woolley. Quantifying Collective Intelligence in Human Groups. Proceedings of the National Academy of Sciences, 118(21):e2005737118, 2021.

[166] Anita Williams Woolley, Christopher F. Chabris, Alex 'Sandy' Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a Collective Intelligence Factor in the Performance of Human Groups. Science, 330:686 – 688, 2010.

[167] James Surowiecki. The Wisdom of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Abacus, 2004.

[168] Andrew W Lo and Ruixun Zhang. The Wisdom of Crowds Versus the Madness of Mobs: an Evolutionary Model of Bias, Polarization, and Other Challenges to Collective Intelligence. Collective Intelligence, 1(1), 2022.

[169] Jennings Anderson, Robert Soden, Kenneth M. Anderson, Marina Kogan, and Leysia Palen. EPIC-OSM: a Software Framework for OpenStreetMap Data Analytics. In 2016 49th Hawaii International Conference on System Sciences (HICSS), pages 5468–5477, 2016.

[170] Youjin Choe, Martin Tomko, and Mohsen Kalantari. Assessing Mapper Conflict in OpenStreetMap Using the Delphi Survey Method. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[171] Anran Yang, Hongchao Fan, and Ning Jing. Amateur or Professional: Assessing the Expertise of Major Contributors in OpenStreetMap Based on Contributing Behaviors. ISPRS International Journal of Geo-Information, 5, 02 2016.

[172] Daniel Bégin, Rodolphe Devillers, and Stéphane Roche. The Life Cycle of Contributors in Collaborative Online Communities-the Case of Openstreetmap. International Journal of Geographical Information Science, 32(8):1611–1630, 2018.

[173] Gloria Urrea and Eunae Yoo. The Role of Volunteer Experience on Performance on Online Volunteering Platforms. Production and Operations Management, 32(2):416–433, 2023.

[174] Martin Dittus, Giovanni Quattrone, and Licia Capra. Analysing Volunteer Engagement in Humanitarian Mapping: Building Contributor Communities at Large Scale. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16, page 108–118. Association for Computing Machinery, 2016.

[175] Melanie Eckle and João Porto de Albuquerque. Quality Assessment of Remote Mapping in OpenStreetMap for Disaster Management Purposes. In International Conference on Information Systems for Crisis Response and Management, 2015.

[176] Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, page 13–25, New York, NY, USA, 2016. Association for Computing Machinery.

[177] Shilad W. Sen, Heather Ford, David R. Musicant, Mark Graham, Os Keyes, and Brent Hecht. Barriers to the Localness of Volunteered Geographic Information. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15, page 197–206, New York, NY, USA, 2015. Association for Computing Machinery.

[178] Robert Soden and Leysia Palen. From Crowdsourced Mapping to Community Mapping: the Post-earthquake Work of OpenStreetMap Haiti. In COOP 2014-Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France), pages 311–326, Cham, 2014. Springer International Publishing.

[179] Jacob Thebault-Spieker, Aaron Halfaker, Loren G. Terveen, and Brent Hecht. Distance and Attraction: Gravity Models for Geographic Content Production. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[180] Brent J. Hecht and Darren Gergle. On the "Localness" of User-Generated Content. In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10, page 229–232, New York, NY, USA, 2010. Association for Computing Machinery.

[181] Isaac L. Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. The Geography and Importance of Localness in Geotagged Social Media. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, page 515–526, New York, NY, USA, 2016. Association for Computing Machinery.

[182] Nama R. Budhathoki and Caroline Haythornthwaite. Motivation for Open Collaboration: Crowd and Community Models and the Case of OpenStreetMap. American Behavioral Scientist, 57(5):548–575, 2013.

[183] Bowen Zhang, Jennings Anderson, Dipto Sarkar, and Robert Soden. A Quantitative Approach to Identifying Emergent Editor Roles in Open Street Map. In Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[184] Peter Mooney and Padraig Corcoran. How Social Is OpenStreetMap? In Proceedings of the AGILE'2012 International Conference on Geographic Information Science, pages 282–287, 2012.

[185] Peter Mooney and Padraig Corcoran. Analysis of Interaction and Co-editing Patterns Amongst OpenStreetMap Contributors. Transactions in GIS, 18, 2014.

[186] Marina Kogan, Jennings Anderson, Leysia Palen, Kenneth M. Anderson, and Robert Soden. Finding the Way to OSM Mapping Practices: Bounding Large Crisis Datasets for Qualitative Investigation. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, page 2783–2795, New York, NY, USA, 2016. Association for Computing Machinery.

[187] Seth Spielman. Spatial Collective Intelligence? Credibility, Accuracy, and Volunteered Geographic Information. Cartography and Geographic Information Science, 41:1–10, 03 2014.

[188] Hans W. Guesgen and Jochen Albrecht. Imprecise Reasoning in Geographic Information Systems. Fuzzy Sets and Systems, 113(1):121–131, 2000.

[189] Muki Haklay. How Good Is OpenStreetMap Information? a Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. Environment and Planning B: Planning and Design, 37:682–703, 2010.

[190] Nicholas Chrisman. The Error Component in Spatial Data, volume 1. 01 1991.

[191] Michael F. Goodchild and Linna Li. Assuring the Quality of Volunteered Geographic Information. Spatial Statistics, 1:110–120, 2012.

[192] Lev Muchnik, Sen Pei, Lucas C. Parra, Saulo D. S. Reis, José S. Andrade, Shlomo Havlin, and Hernán A. Makse. Origins of Power-Law Degree Distribution in the Heterogeneity of Human Activity in Social Networks. Scientific Reports, 3, 2013.

[193] Vinit Tipnis, Eunae Yoo, Gloria Urrea, and Fei Gao. AI-Powered Philanthropy: Effects on Volunteer Productivity. Social Science Research Network, pages 1–24, 2024.

[194] Dagoberto José Herrera-Murillo, Héctor Ochoa-Ortiz, Umair Ahmed, Francisco Lopez-Pellicer, Barbara Re, Andrea Polini, and Javier Nogueras-Iso. Process Analysis in Humanitarian Voluntary Geographic Information: the Case of the HOT Tasking Manager. AGILE: GIScience Series, 5:1–12, 2024.

[195] Pascal Neis and Alexander Zipf. Analyzing the contributor activity of a volunteered geographic information project — the case of openstreetmap. ISPRS International Journal of Geo-Information, 1(2):146–165, 2012.

[196] Kirsty Watkinson, Jonathan Huck, and Angela Harris. Using Gamification to Increase Map Data Production During Humanitarian Volunteered Geographic Information (VGI) Campaigns. Cartography and Geographic Information Science, 50:1–17, 2023.

[197] Radim Štampach, Lukáš Herman, Jakub Trojan, Kateřina Tajovská, and Tomáš Řezník. Humanitarian mapping as a contribution to achieving sustainable development goals: Research into the motivation of volunteers and the ideal setting of mapathons. Sustainability, 13(24), 2021.

[198] Martin G. Everett and Steve P. Borgatti. The Centrality of Groups and Classes. The Journal of Mathematical Sociology, 23(3):181–201, 1999.

[199] Dagoberto José Herrera-Murillo, Javier Nogueras-Iso, Miguel Ángel Latre, Paloma Abad-Power, and Francisco J. Lopez-Pellicer. A Framework for the Acceptance Testing of Geospatial Search Engines. Environmental Modelling Software, 2025. Just accepted.

[200] Dagoberto José Herrera-Murillo, Paloma Abad-Power, Francisco J. Lopez-Pellicer, Sandra Baldassarri, and Javier Nogueras-Iso. Applicability of Process Mining in Usability Tests: A Case Study for Identifying User Mental Models in Geospatial Search Engines. 2025. Paper under submission.

[201] Dagoberto José Herrera-Murillo, Héctor Ochoa-Ortiz, Umair Ahmed, Francisco Javier López-Pellicer, Barbara Re, Andrea Polini, and Javier Nogueras-Iso. Collective Intelligence in Humanitarian Voluntary Geographic Information: the Case of the HOT Tasking Manager. ACM Transactions on Computer-Human Interaction, 2025.

[202] 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.

[203] Dagoberto José Herrera-Murillo, Abdul Aziz, Javier Nogueras-Iso, and Francisco J. Lopez-Pellicer. "Analysing User Involvement in Open Government Data Initiatives. In Gianmaria Silvello, Oscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro, and Antonella Poggi, editors, Linking Theory and Practice of Digital Libraries", pages 175–186, Cham, 2022. Springer International Publishing.

[204] Abdul Aziz, Mohsan Ali, Dagoberto José Herrera-Murillo, Maria Ioanna Maratsi, Francisco J. Lopez-Pellicer, and Javier Nogueras-Iso. A Framework for the Thematic Annotation of Open Government Data. 2025. Paper under submission.

[205] Abdul Aziz, Dagoberto José Herrera-Murillo, Javier Nogueras-Iso, and Francisco J. Lopez Pellicer. Towards a Sustainable Open Data Ecosystem: First Steps for Optimizing Findability and User Feedback. Presented at the XI Jornada de Jóvenes Investigadores del I3A, Zaragoza, Spain, 2022.

[206] Dagoberto José Herrera-Murillo, Francisco J. Lopez Pellicer, and Javier Nogueras-Iso. Métodos de Investigación de Experiencia de Usuario al Servicio de las IDE. Presented at the XV Jornadas Ibéricas de las Infraestructuras de Datos Espaciales (JIIDE), Vitoria-Gasteiz, Spain, 2024.

[207] Hector Ochoa-Ortiz, Dagoberto José Herrera-Murillo, Umair Ahmed, Francisco J. Lopez-Pellicer, Barbara Re, Andrea Polini, and Javier Nogueras-Iso. How do Users Interact With the HOT Tasking Manager. Presented at State of the Map Europe 2024, Łódź, Poland, 2024.

TEST SCENARIOS FOR THE ACCEPTANCE TESTING OF USER INTERFACES WRITTEN IN GHERKIN

This appendix presents two tables with the implementation details of the Gherkin scenarios used for functionality testing through scenarios and usability testing through cognitive walkthroughs, as described in Chapter 2.

Table 1 Test scenarios for the search feature written in Gherkin

Test scenario	Status
Scenario: the user is able to search for resources typing text **Given** the user is on the home page of the search engine **When** the user performs a textual search for *"Madrid"* **Then** search results are displayed	Passed
Scenario: the user is able to search for resources selecting a point on the map **Given** the user is on the home page of the search engine **When** the user searches for a point in the centre of the map **Then** search results are displayed	Passed
Scenario: the user is able to search for resources drawing a geometry on the map **Given** the user is on the home page of the search engine **When** the user searches for a geometry in the centre of the map **Then** search results are displayed	Passed
Scenario: the user is able to search for resources uploading a geometry file **Given** the user is on the home page of the search engine **When** the user loads the file *"BTT0101_vivar_del_cid-burgos.gpx"* **Then** search results are displayed	Passed
Scenario: the user is able to search for resources typing a set of coordinates **Given** the user is on the home page of the search engine **When** the user types coordinate *"3.40" "40.30"* **Then** search results are displayed	Passed
Scenario: the user is able search for resources typing a cadastral reference **Given** the user is on the home page of the search engine **When** the user enters the cadastral reference *"9977715VK3797F"* **Then** search results are displayed	Passed

Table 2 Cognitive walkthroughs written in Gherkin.

Test case	Status	# results*	Ex. Time
Scenario: Discover cartographic resources of the autonomous community of *"Asturias"*			
Given the user is on the home page of the search engine			
When the user performs a textual search for *"Asturias"*	*Passed*	*T: 12,359*	*39.433s*
Then resources related to *"Asturias"* are displayed in the *"All"* view		*D: 112,308*	
When the user selects one of the available resources		*P: 141*	
Then a full metadata record describing the resource is displayed in a new tab			
Scenario: Download a trail file related to the search for the *"Way of El Cid"*			
Given the user is on the home page of the search engine			
When the user performs a textual search for *"Way of El Cid"*		*T: 862*	
Then resources related to *"Way of El Cid"* are displayed in the *"All"* view	*Passed*	*D: 12,165*	*50.610s*
When the user selects one of the available resources		*P: 35*	
Then a full metadata sheet describing the resource is displayed in a new tab			
When the user downloads one of the files available in the metadata record			
Then the file is downloaded locally			
Scenario: Buy the current map of the city of *"Toledo"*			
Given the user is on the home page of the search engine			
When the user performs a textual search for *"Toledo"*		*T: 21,720*	
Then resources related to *"Toledo"* are displayed in the *"All"* view	*Passed*	*D: 116,708*	*48.223s*
When the user selects one of the available resources		*P: 192*	
Then a full metadata record describing the resource is displayed in a new tab			
When the user buys one of the available resources in the metadata record			
Then the selected product is added to the shopping cart			
Scenario: Discover general cartographic resources of the region of *"Murcia"*		*T: 17,911*	
Given the user is on the home page of the search engine		*D: 109,854*	
When the user performs a textual search for *"Murcia"*		*P: 139*	
Then resources related to *"Murcia"* are displayed in the *"All"* view	*Passed*	*After filtering*	*44.140s*
When the user selects the filter of *"General Cartography"*		*T: 456*	
Then only the resources related to *"General Cartography"* are displayed		*D: 3,133*	
		P: 136	
Scenario: View the area of the *"Sierra Nevada"* National Park on the side map			
Given the user is on the home page of the search engine		*T: 1,759*	
When the user performs a textual search for *"Sierra Nevada"*	*Passed*	*D: 22,611*	*38.939s*
Then resources related to *"Sierra Nevada"* are displayed in the *"All"* view		*P: 117*	
When the user locates one of the available resources			
Then the location of the resource is shown in the side map			

*T (Total results), D (Downloads), P (Products)

A Closer Look At the Mapping Process

This appendix provides a more detailed analysis of the mapping process, examining the frequency and duration of task states based on the specific project type.

Table 3 analyses the frequency of task states. Whereas the first column presents the absolute frequency of each state across the entire dataset, the subsequent columns display the percentage of task coverage across all projects, as well as within different difficulty and priority categories. A darker tone indicates a higher coverage.

Task state	Frequency of states	TOTAL TASKS	Difficulty			Priority			
	N=1,853,074	N=312,289	Easy	Moderate	Challenging	Low	Medium	High	Urgent
			n=222,729	n=88,468	n=1,092	n=166,911	n=80,768	n=38,119	n=26,491
Locked for mapping	616,659	100	100	100	100	100	100	100	100
Locked for validation	379,266	100	100	100	100	100	100	100	100
Mapped	349,713	100	100	100	100	100	100	100	100
Validated	330,290	100	100	100	100	100	100	100	100
Split	70,612	14	10	21	54	10	9	20	43
Auto unlocked for mapping	62,632	9	8	12	15	7	5	12	33
Invalidated	29,334	6	5	9	17	4	4	10	23
Bad imagery	7,562	2	2	2	0	1	3	2	2
Auto unlocked for validation	3,863	1	1	1	0	1	1	1	2
Extended for mapping	3,143	0	0	0	1	0	0	0	0

Table 3 Task states according to frequency and case coverage

Overall, the results indicate that the most frequent state is Locked for mapping, where the creation activity takes place, typically requiring an average of two cycles per task. The Auto-unlocked for mapping state, which is also linked to map creation, appears in almost one tenth of the tasks. The frequencies of states associated with decision-making at the end of the mapping phase suggest that tasks are generally declared as Mapped only once, while the option Split is used infrequently, and the Bad imagery state is rarely selected by mappers. The average frequency of the Locked for validation and Validated states per task, combined with the low case coverage of the Invalidated states (6.3%), indicates that the validation phase is generally straightforward. Other states, such as Auto-unlocked for validation and Extended for mapping, occur only marginally, regardless of the type of project.

Table 4 expands on the duration of different states and transitions. Regarding the relative duration of mapping versus validation, the top section of the table presents the median duration (in minutes) of the Locked for mapping and Locked for validation states for the different project types.

		Difficulty			Priority			
	TOTAL	Easy	Moderate	Challenging	Low	Medium	High	Urgent
Locked for mapping (minutes)	2.1	2.0	2.9	7.5	1.6	2.5	2.9	4.8
Locked for validation (minutes)	4.1	5.2	2.8	1.8	4.0	4.3	4.9	3.8
% of total tasks where total validation time is higher than total mapping time	57.5	59.6	51.4	21.1	61.4	57.9	50.4	27.5
Locked for mapping -> Locked for mapping (days)	1.0	1.1	0.8	3.0	0.8	1.7	1.5	0.2
Mapped -> Locked for validation (days)	13.2	25.1	3.2	3.8	27.0	6.8	3.8	4.7
Invalidated -> Unlocked for mapping (days)	1.0	1.4	0.6	74.4	1.0	1.8	0.8	0.4

Table 4 Median duration of task states and transitions (85% of most frequent traces)

Our observations show that as the difficulty of the project increases, the median time required for mapping also increases, while the median time required for validation decreases. Furthermore, as the priority level increases, the median duration of the LOCKED FOR MAPPING state increases, while the LOCKED FOR VALIDATION state remains relatively stable. The middle section of the table illustrates the percentage of tasks in which the total duration of the LOCKED FOR VALIDATION states exceeds that of the LOCKED FOR MAPPING states. In just over half of the tasks (except for challenging and urgent projects) the validation time exceeds the mapping time. In the case of easier projects, it appears that validation activities go beyond mere verification, often taking on a greater burden of correcting the work done during the mapping phase. Regarding the cost of waiting, the lower section of the table shows the median waiting times between states, measured in days. The cost of waiting for another LOCKED FOR MAPPING by deciding not to declare a task as MAPPED or INVALIDATED is maximized for challenging projects and minimized for urgent projects. Meanwhile the waiting time for a MAPPED task to move to the LOCKED FOR VALIDATION state is more pronounced in easy and low priority project tasks and significantly reduced in other categories.