A COMPARISON OF OPEN DATA OBSERVATORIES

Naeima Hamed
School of Computer Science and Informatics
Cardiff University, UK
hamednh@cardiff.ac.uk

Omer Rana
School of Computer Science and Informatics
Cardiff University, UK
ranaof@cardiff.ac.uk

Pablo Orozco-terWengel
School of Biosciences
Cardiff University, UK
orozco-terwengelpa@cardiff.ac.uk

Benoît Goossens
School of Biosciences
Cardiff University, UK
goossensbr@cardiff.ac.uk

Charith Perera
School of Computer Science and Informatics
Cardiff University, UK
pererac@cardiff.ac.uk

April 15, 2024

ABSTRACT

Open Data Observatories refer to online platforms that provide real-time and historical data for a particular application context, e.g., urban/rural environments or a specific application domain. They are generally developed to facilitate collaboration within one or more communities through reusable datasets, analysis tools, and interactive visualizations. Open Data Observatories collect and integrate various data from multiple disparate data sources—some providing mechanisms to support real-time data capture and ingest mechanisms. Data types can include sensor data (soil, weather, traffic, pollution levels) and satellite imagery. Data sources can include Open Data providers, interconnected devices, and services offered through the Internet of Things. The continually increasing volume and variety of such data require timely integration, management, and analysis, yet presented in a way that end-users can easily understand. Data released for open access preserve their value and enable a more in-depth understanding of real-world choices. This survey compares thirteen Open Data Observatories and their data management approaches. We investigated their aims, design, and types of data. We conclude with research challenges that influence the implementation of these observatories, outlining some advantages and limitations for each one and recommending areas for improvement. Our goal is to identify best practices learned from the selected observatories to aid the development of new Open Data Observatories.

***K*eywords** Urban observatories, non-urban observatories, dashboards, data portal, smart city data, open data principles, linked open data, 5-star models

1 Introduction

Structured, semi-structured, and unstructured data can be generated from diverse sources, including government authorities, academic institutions, and citizens. These data categories apply to every sort of data, with structured data including inventories and catalogs organized in tables, semi-structured data such as operational manuals in JSON (JavaScript Object Notation) and XML (eXtensible Markup Language) formats, and unstructured data including text and media. These data are collected through various methods, such as questionnaires, web scraping and Internet of Things (IoT) devices. While many governments have embraced the "Open Data" concept and made some of their data public, some commercial organizations collect large volumes of data, but only a fraction is accessible. Open Data refer to data that are made available to the public by governments, organizations, and individuals [1]. They promote transparency, collaboration, and innovation, which can improve the quality of scientific research and contribute to the development of a sustainable ecosystem [2, 3]. Open Data portals serve as gateways to a wide range of datasets and resources from various sources, including governments, non-profit organizations, and private companies. They provide search and discovery tools, data visualization capabilities, and options for downloading data [4].

Open Data Observatories curate and integrate real-time and historical data from different sources, presenting them in a unified manner. Previous research initiatives in [5] developed methods to survey Open Data Observatories, providing insights into their availability and helping data publishers select the most suitable platforms for their data.

Stall et al. [6] introduced the Generalist Repository Comparison Chart (GRCC) to assist researchers in identifying a generalist repository when a domain-specific repository [7] is unavailable for storing their research data. They provide a broad platform for sharing diverse research outputs such as articles, datasets, codes, and digital research products. These repositories (e.g., Zenodo, Figshare, and Dryad) require users to deposit their research outputs under open licenses, ensuring accessibility for further use.

Open Data portals, Open Data Observatories, and generalist repositories represent distinct system within the data sharing ecosystem, each serving unique functions and targeting specific audiences. Open Data portals are centralized platforms where governments, organizations, and institutions release datasets to the public, aimed at enhancing transparency, enabling societal and economic benefits, and fostering innovation through open access to information on a variety of topics such as demographics, economics, and government operations [8]. Open Data Observatories focus on monitoring and analyzing specific datasets for trends and insights, typically in public or research domains, while generalist repositories archive diverse types of scholarly work, including datasets, articles, and preprints, thus supporting interdisciplinary research and increasing the visibility and impact of academic work beyond traditional publication venues. The reliance on Open Data Observatories has become increasingly crucial in tackling the complex challenges faced by contemporary society and the environment. A series of studies by Miller et al. [9], Moustaka et al. [10], Ma et al. [11], and Liu et al. [12] provided an understanding of the role of Open Data Observatories in areas such as urban sustainability, smart city analytics, and ocean science. Our study aims to compare different Open Data Observatories to highlight their distinct characteristics, methodologies, and challenges they encounter. By identifying and extrapolating best practices from these observatories, the goal is to facilitate the development of new Open Data Observatories and to better understand their impact on decision-making and policy formulation in urban and non-urban settings.

This study's research questions are formulated as follows:

- What are the key features and functionalities of different Open Data Observatories?
- How do different Open Data Observatories compare regarding data coverage, accessibility, and usability?
- What are the strengths and limitations of different Open Data Observatories?
- What are the challenges organizations face when building Open Data Observatories, and how can these challenges be addressed?

To achieve the research questions, we:

1. We selected and compared thirteen Open Data Observatories based on various criteria, such as data types, data coverage, accessibility, and usability.
2. We investigated the data management approaches in the context of Open Data Observatories.
3. We outlined their strengths and limitations and suggested areas for improvement.
4. We identified the critical challenges faced by organizations when building Open Data Observatories, such as technical and intellectual challenges.

This research is structured as follows: Section 2 investigates the use of the term Open Data, its principles, and main sources. Section 3discusses the study's research methodology. Section 4 introduces the thirteen selected Open Data Observatories, individually describing their aim, data management approaches, and the (smart) services they support. Section 5 recapitulates the types of data they support, examining their themes, sources and the methods employed in their processing. Section 6 describes four key research challenges, namely data integration, quality, provenance, and privacy. Section 7 interpret the study's findings, compare them with existing knowledge, address research questions, evaluate implications, and guide future research directions. Finally, Section 8 concludes the study.

Table 1: Description and comparison of Open Data principles proposed by Sebastopol, the Sunlight Foundation and how they map to the FAIR (Findable, Accessible, Interoperable and Reusable) data principles.

Principle	Description	Sebastopol	Sunlight Foundation	FAIR Data Principles
1. Complete	Data must be a complete and accurate representation of the original observations including all computational details.	✓	✓	Findable
2. Primary	Data collected at the source with detailed metadata.	✓	✓	Findable
3. Timely	Data published promptly after collection.	✓	✓	Accessible
4. Accessible	Data must be easily accessible both physically and electronically.	✓	✓	Accessible
5. Machine-processable	Data in a format that can be easily processed by computers.	✓	✓	Interoperable
6. Non-discriminatory	Data is accessible to anyone without restrictions.	✓	✓	Accessible
7. Non-proprietary	Data in a format that does not require proprietary software.	✓	✓	Interoperable
8. License Free	Data freely available without restrictions.	✓	✓	Reusable
9. Permanence	Data remains accessible online, including all versions.		✓	Accessible
10. Usage costs	Accessing and obtaining data incur no fees.		✓	Accessible and reusable

2 Open Data

Open Data are free data, released under open licenses [13] and organized in structured formats that follow established standards and conventions. This allows the data to be easily understood and processed by both humans and machines. They are accompanied by metadata, which provides additional information about the data, such as their source, creation date, data dictionary, and other relevant details. This metadata helps users better understand and contextualize the data. Open Data are also presented in formats that are designed to be easily read and processed by computer programs and algorithms [3]. This enables automated analysis, integration of the data, making it more accessible and useful for a wide range of applications [14]. This section investigates Open Data principles and sources.

2.1 Open Data Principles

The expansion of Open Data is influenced by fundamental frameworks such as the Berners-Lee Five-Star Model [1] principles established by organizations like the Sunlight Foundation [15]. This Five-Star Model evaluates Open Data on a scale from one to five stars, with higher ratings indicating data that are open, machine-readable, and compliant with open standards. Kucera et al. [16] investigated the challenges related to publishing and reusing Open Government Data, emphasizing methodologies and best practices in this domain. This includes the establishment of a publication methodology within the COMSODE project, which highlights the role of Open Government Data in fostering transparency and citizen engagement. Open Data principles, further expanded upon by groups such as the Sebastopol [17] attendees and the Sunlight Foundation, establish a comprehensive framework to ensure government data are openly accessible. The FAIR data principles [18, 19, 20] provide a set of guidelines aimed at enhancing data reusability for both humans and machines, stressing the importance of data being *Findable, Accessible, Interoperable, and Reusable.* Table 1 integrates Open Data principles, as discussed by both the Sebastopol group and the Sunlight Foundation, with the broader framework of the FAIR data principles, providing a comparative overview of their alignment. It shows ten critical principles identified for the openness and availability of government data, ranging from ensuring data completeness and primacy to guaranteeing accessibility, machine processability, and non-discrimination. Moreover, it introduces considerations for non-proprietary formats, license freedom, permanence, and the elimination of usage costs to foster a more inclusive and accessible digital ecosystem. This alignment is further enhanced by indicating which of these Open Data principles correspond to which element FAIR data principles.

2.2 Open Data Sources

Scientific research heavily relies on Open Data sources for replication, validation, and growth. Open Data can be obtained from various entities, including government bodies, academic institutions, and citizens. For example, Open Government Data encompass a wide range of information such as demographics (age, gender, race), economic indicators (GDP, unemployment rates), weather data, and public health indices. These data types enable researchers to examine social trends, economic patterns, public health outcomes, and their interrelationships. Academic research data from universities and institutions also contribute to Open Data sources. Researchers are increasingly required by publishers to make the data contributing to a paper available. This includes making their data available for others to use and build upon, including surveys and observational data that provide empirical evidence. By sharing these data openly, researchers foster collaboration, facilitate replication, and allow for the expansion of scientific knowledge. In recent years, citizen-generated data through smartphones and mobile devices have gained increasing value, particularly in social science and humanities studies [21]. These data include information collected through social media platforms, GPS tracking, and other mobile applications. Researchers can use citizen-generated data to study many topics, including online communities, human behaviour, social interactions, urban dynamics, and cultural trends. Sensor networks significantly contribute data on environmental conditions, vehicle movement, and electricity usage. These networks provide valuable information for research related to urban planning, environment sustainability, transportation patterns, and energy consumption. While Open Data sources offer numerous benefits, they also present challenges. Data quality assurance, privacy protection, and managing diverse data types are some hurdles researchers must address. However, the potential of Open Data sources is evident, and they are expected to play an increasingly significant role in scientific research.

3 Research Method

We employed a methodology known as SPIDER (Sample, Phenomenon of Interest, Design, Evaluation, Research type) [22] to guide our review. SPIDER is a framework specifically designed for conducting rigorous, transparent, and reproducible reviews. To ensure comprehensive coverage, we extracted keywords for each SPIDER element based on synonyms and related terms derived from our research questions. We conducted searches using the Google search engine, Google Scholar, ACM digital library, and Cardiff University library, focusing on the following terms:

1. **S**ample: Open Data observatory.
2. **P**henomenon of Interest: domain-specific data observatory, multi-domain data observatory.
3. **D**esign: Open Data platforms.
4. **E**valuation: relevance, transparency, accessible.
5. **R**esearch type: descriptive, survey, research article.

3.1 Search Plan

Our search plan used the Boolean operators AND and OR to connect the search items corresponding to each SPIDER element. This approach allowed us to construct comprehensive search queries that incorporated relevant terms. For instance, the search query for the SPIDER elements would look like this: Sample AND Phenomenon of Interest AND Design AND Evaluation AND Research type ("Open Data platform" OR "Open Data observatory") AND ("domain-specific data observatories" OR "domain-specific observatory" OR "multi-domain observatory" OR "data integration") AND ("accessible online platforms" OR "data platform") AND ("relevance" OR "transparency" OR "rigour") AND ("descriptive" OR "survey"). Using the OR operator within parentheses, we expanded the search to include variations and synonyms for terms such as "Open Data platform" and "Open Data observatory." We incorporated terms related to the phenomenon of interest, such as "domain-specific data observatories," "domain-specific observatory," "multi-domain observatory," and "data integration." To capture different aspects of the design and evaluation, we included phrases like "accessible online platforms" and "data platform." We also encompassed terms related to the desired research attributes, such as "relevance," "transparency," and "rigour," and the research types, such as "descriptive" and "survey." This search strategy ensured we covered a wide range of relevant literature and maximised the chances of identifying relevant studies for our review.

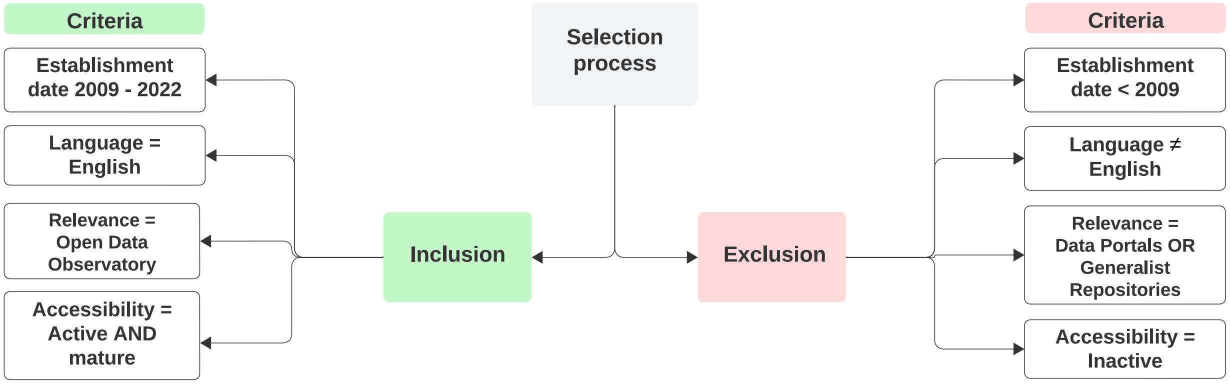

Criteria
Establishment date 2009 - 2022
Language = English
Relevance = Open Data Observatory
Accessibility = Active AND mature
Selection process
Inclusion
Exclusion
Criteria
Establishment date < 2009
Language ≠ English
Relevance = Data Portals OR Generalist Repositories
Accessibility = Inactive

Figure 1: Inclusion and exclusion criteria for selecting the reviewed Open Data Observatories .

3.2 Observatories Selection Process

The results obtained from the previous step yielded a substantial number of platforms, some of which were not directly relevant to our research questions. We established specific inclusion and exclusion criteria to refine the selection process and ensure that only the most relevant platforms were included in our study. These criteria, outlined in Figure 1, were based on several factors, including the domain experts' suggestions, platforms' establishment date, and relevance to our research questions. By setting these criteria, we aimed to focus our analysis on the most recent platforms available in English. We prioritised platforms that demonstrated clear relevance to our research questions.

3.3 Observatories Selection Result

The initial search process yielded 40 Open Data environments. We manually checked each one to ensure that we focused specifically on Open Data Observatories. Through this evaluation, we were able to filter out and identify 34 environments that met the criteria of being Open Data Observatories. After completing a thorough manual evaluation, we arrived at a final selection of 13 Open Data Observatories that satisfied all the necessary criteria. These 13 observatories (Figure 2) will be introduced and discussed in the subsequent section of the study. By employing this rigorous manual verification process, we ensured that the selected Open Data Observatories were reliable, accessible, and relevant to our research questions.

4 Open Data Observatories

This section provides a chronological overview of the selected Open Data Observatories, starting from the older and progressing to the newer ones (Figure 2). Each observatory is concisely outlined and characterized by its attributes, kinds of data, and significant accomplishments or obstacles. This presentation aims to offer readers a thorough understanding of the selected observatories and their contributions to Open Data research and implementation.

4.1 Terrestrial Ecosystem Research Network (TERN)

Terrestrial Ecosystem Research Network (TERN)[1] is a national research infrastructure program in Australia that supports ecosystem science, observations, and data management. TERN was established in 2009 by the Australian Government in response to a growing need for a coordinated approach to terrestrial ecosystem research and management. The network comprises a range of field sites and data infrastructure that supports long-term environmental monitoring and research, including measurements of ecosystem processes, biodiversity, and land surface properties. TERN's infrastructure includes over 600 environmental monitoring sites across Australia and advanced data management systems that allow researchers to access and analyse data from multiple sources. TERN aims to support evidence-based decision-making for ecosystem management and conservation in Australia and to promote a greater understanding of terrestrial ecosystems and their role in maintaining global environmental health.

TERN hosts a substantial and growing collection of diverse ecosystem datasets from across Australia, covering topics such as mangroves, vegetation, soil, and phenology. TERN provides a variety of data tools and services, including

[1] tern.org.au/

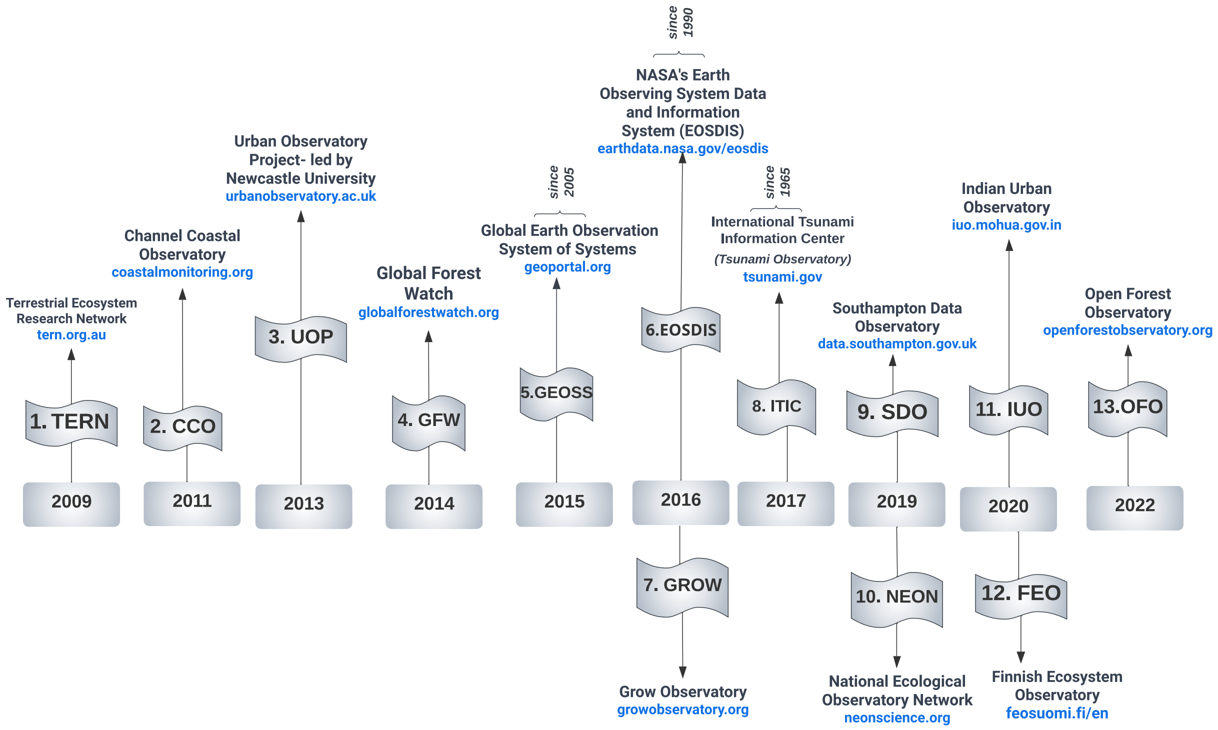


Figure 2: Timeline displays the selected Open Data Observatories. 1. The Terrestrial Ecosystem Research Network (TERN) [23], 2. Channel Coastal Observatory (CCO), 3. The Urban Observatory Project (UOP), 4. Global Forest Watch (GFW) [24], 5. Global Earth Observation System of Systems (GEOSS) [25, 26], 6. NASA's Earth Observing System Data and Information System (EOSDIS) [27], 7. Grow Observatory (GROW), 8. International Tsunami Information Center (ITIC)- Tsunami Observatory, 9. Southampton Data Observatory (SDO), 10. National Ecological Observatory Network (NEON) [28], 11. Indian Urban Observatory (IUO) 12. Finnish Ecosystem Observatory (FEO) [29], 13. Open Forest Observatory (OFO).

SHaRED for data submission and harmonization, aligning with the FAIR principles, a Data Discovery Portal for accessing diverse ecosystem datasets, tools for data analysis and visualization such as MCAS-S and the Data Visualiser, cloud-based research platforms like CoESRA, and resources for field data collection, including a network of monitoring sites. In addition, the Threatened Species Index- TSX(tsx.org.au) is a dynamic tool that helps understand how Australia's threatened species are faring over time. It provides visualizations and detailed data on temporal trends for 286 species of threatened and near-threatened mammals, birds, and plants in Australia.

4.2 Channel Coastal Observatory (CCO)

Since 2011, the National Network of Regional Coastal Monitoring Programmes has supported six projects along the English coastline. The overarching objective of these projects is to gather in-situ coastal monitoring data [30]. However, Contarinis et al. [31] highlighted some inconsistencies in the quality of the data collected and the methodologies employed by traditional management approaches. The Channel Coastal Observatory (CCO)[2] was established in response to these challenges. In England, 520,000 properties face the risk of coastal flooding, while 8,900 are threatened by coastal erosion. CCO aims to provide consistent and reliable data to aid decision-makers in understanding coastal behaviour and identifying potential risks associated with coastal flooding and erosion [32]. CCO covers various coastal regions, including the Northeast, East Riding of Yorkshire, Anglian, Southeast region (low-lying land), and Northwest. The primary data types collected and displayed on its platform include topographic and hydrographic surveys. Topographic surveys focus on features such as beaches, cliffs, dunes, and coastal defence structures, while hydrographic surveys extend from the Mean Low Water (MLW) contour to 1 kilometre offshore. CCO offers a collection of real-time data on waves, tides, meteorology, and GPS measurements, which are crucial for understanding

[2] coastalmonitoring.org/

and managing coastal environments. CCO offers a public API that allows developers to access and integrate the real-time coastal data (waves, tides, meteorology) collected by the monitoring programs. It also provides information on how to access the coastal data through Web Map Services (WMS) in GIS software such ArcMap and QGIS.

4.3 Urban Observatory Project (UOP)

Urban Observatory Project (UOP)[3] was launched in 2013 and sponsored by the UK Collaboratorium for Research on Infrastructure and Cities (UKCRIC) - led by Newcastle University in collaboration with five other British universities; Sheffield, Bristol, Cranfield, Birmingham, and Manchester. UOP aims to monitor and analyse urban areas through the deployment of various sensors across these cities. It collects vast amounts of real-time data from sensors and other sources to gain insights into urban dynamics. Each participating university focuses on specific aspects of urban life. For instance, Sheffield Urban Flows Observatory examines the impact of energy and resource flows on economic performance and social well-being. At the same time, Bristol Urban Flows Observatory transforms Bristol into a living laboratory for community engagement. Cranfield Urban Observatory provides data-centric and remote-sensing solutions for addressing environmental, social, and economic issues. Birmingham Urban Observatory monitors critical infrastructure and its interplay with the environment, economy, and society. Lastly, Manchester Urban Observatory collects, analyses, and shares urban data to support decision-making processes. The collaborative efforts of these observatories contribute to a better understanding of urban dynamics and offer insights for sustainable and efficient urban development [33]. UOP's data types include traffic flow, parking spaces, cycling docking, pedestrian count, weather data, air quality, water quality, seismic activity, noise-level, water-level (rainfall), beehives, energy usage data, thermal imaging, visual and hyper-spectral mapping, social media feeds, employee feedback, and quantifying the impacts of COVID-19 measures. More details about UKRIC observatories are available as supplement materials in Appendix A 8.

4.4 Global Forest Watch (GFW)

Global Forest Watch (GFW) initiative[4] is a non-profit organization that is part of the World Resources Institute (wri.org). GFW collaborates with over 100 organizations to provide a transparent and actionable platform that is supported by satellite technology and cloud computing. This initiative empowers various stakeholders, including law enforcement, companies, and governments, in forest management and combating deforestation. The GFW's web-based platform (observatory), which was launched in 2014, provides data and tools for monitoring forests and land use. The platform has amassed over four million users worldwide, benefiting diverse groups such as local law enforcement, park managers, international corporations, and civil society organizations in their endeavors to safeguard forests. GFW's key applications include the Forest Watcher mobile app for real-time threat detection, GFW Pro for managing deforestation risks in supply chains, and the Global Forest Review for monitoring global forest objectives. Moreover, national governments employ GFW's technology for forest resource management, while small grants and fellowships support additional advocacy and research. Collectively, GFW assists in forest surveillance and management, combats illegal deforestation, promotes sustainable commodity sourcing, and supports conservation research on a global scale. GFW data types include satellite imagery for observing changes in forest cover, forest change data for tracking deforestation and regrowth, and land cover data for understanding land usage. In addition to data about biodiversity, climate dynamics, and commodity supply chains, as well as legal and administrative boundaries, fire alerts, and water resources. GFW provides both developer-focused tools (APIs and open-source code) and a user-friendly MapBuilder platform to enable the creation of customized interactive mapping applications that leverage GFW's robust spatial data and analysis capabilities.

4.5 Global Earth Observation System of Systems (GEOSS)

Global Earth Observation System of Systems (GEOSS)[5] was created following directives from the 2002 United Nations World Summit on Sustainable Development and the G8's 2005 commitment. Its purpose was to improve the development and application of earth observation technologies for environmental monitoring and management. Initiated in 2005 with a 10-year implementation plan, GEOSS aimed to provide comprehensive, coordinated, and sustained observations of the Earth, focusing on nine key societal benefits such as sustainable agriculture, biodiversity conservation, and climate change adaptation. The success of GEOSS's first decade led to the implementation of a renewed 10-year plan (2016-2025), which aligned with global initiatives such as the UN Committee of Experts on Global Geospatial Information Management (UN-GGIM) and the G8 Open Data Charter to enhance data sharing and

[3] urbanobservatory.ac.uk

[4] wri.org/initiatives/global-forest-watch

[5] geoportal.org/

management. GEOSS evolved into more than just a technological project; it became a global partnership that advocated for the significance of Earth observations and engaged with stakeholders to tackle global challenges. One of GEOSS's notable achievements was the establishment of the GEOSS's data sharing principles, which advocated for Open Data access, minimal use restrictions, and prompt availability of data, metadata, and products. These principles significantly influenced global data policies, including the European Union's Copernicus program [26]. GEOSS encompasses a wide array of data types, aiming to facilitate comprehensive, coordinated, and continuous observations of the Earth system. Data types include but are not limited to, satellite imagery, atmospheric data, oceanographic data, geological data, biodiversity information, and climate metrics.

4.6 NASA's Earth Observing System Data and Information System (EOSDIS)

The Earth Observing System Data and Information System (EOSDIS)[6] is a vital part of NASA's Earth Science Data Systems Program, providing extensive capabilities for managing data from various sources, including satellites, aircraft, field measurements, and other programs. EOSDIS supports the Earth Observing System (EOS) satellite missions by handling tasks such as command and control, scheduling, data capture, and initial processing. These mission operations are overseen by NASA's Earth Science Mission Operations Project. EOSDIS's Science Operations, managed by NASA's Earth Science Data and Information System Project, involve generating higher-level science data products (levels 1-4), archiving, and distributing data products from EOS missions, as well as other satellite missions, aircraft, and field measurement campaigns. This function is carried out within a distributed system that consists of interconnected nodes of Science Investigator-led Processing Systems and Distributed Active Archive Centers (DAACs), which are discipline-specific. EOSDIS offers a variety of curated data types that are crucial for evaluating ecosystem conditions, predicting species' geographical distributions, identifying materials based on spectral properties, and monitoring human-induced environmental changes. These data types include vegetation health, spectroscopy, species distribution, and environmental change tracking data.

4.7 Grow Observatory (GROW)

Grow Observatory (GROW)[7] serves as a citizens' observatory that has enabled individuals and communities to take proactive measures about soil and climate across Europe. GROW collected soil moisture, temperature, and light level data from low-cost "Flower Power" sensors deployed across 24 locations in 13 European countries, resulting in a network of 6,502 ground-based soil sensors and a dataset of 516 million rows of soil data. The sensors were installed and maintained by a network of citizen scientists, community groups, land managers, and researchers. The sensors' data were collected every 15 minutes and uploaded to the GROW servers using mobile phones. GROW integrated the sensors' data through an online platform, allowing members to register their sensors and visualise the data through time-series graphs and maps. GROW also used GEOSS (observatory 6) data to provide public access to archived earth observation data. This information was then used to more accurately predict extreme events, such as floods, droughts, and wildfires. In addition, GROW data played a significant role in validating and calibrating satellite observations, such as those from the European Space Agency's (ESA), SMOS (Soil Moisture and Ocean Salinity) mission and the future SMAP (Soil Moisture Active Passive) satellite. Artists and designers have played a significant role in GROW, with the former creating artworks reflecting the importance of soil ecosystems and remote sensing satellites and designing dynamic visualizations for agriculture and climate forecasting. It has also helped farmers in the Canary Islands reduce their water usage for irrigation by 30%. GROW received awards, including the Land and Soil Management Award 2019, the Stephen Fry Award for Excellence in Public Engagement 2020, and recognition as the first in the European Commission's annual GEO Plenary Statement on significant Earth Observation developments in 2019.

4.8 International Tsunami Information Center (ITIC)- Tsunami Observatory

In March 2017, NOAA's National Tsunami Warning Center and Pacific Tsunami Warning Center, in partnership with the Tsunami Service Program, centralized their information on a Tsunami Observatory[8]. Serving as a hub for information on tsunamis, it provides warnings, advisories, watches, and threat evaluations for Alaska, British Columbia, Washington, Oregon, and California regions. The observatory offers real-time updates on seismic events that could cause tsunamis. These updates include specific information such as event magnitude, depth, coordinates, and the time the seismic event occurred. It also shares bulletins and statements about the current tsunami status, clearly indicating if there are any warnings, advisories, watches, or threats in effect for the mentioned areas. Tsunami Observatory aims to inform the public about tsunami risks following seismic activities, promoting safety and preparedness among residents of

[6] earthdata.nasa.gov/eosdis

[7] growobservatory.org/

[8] tsunami.gov

potentially affected regions. It also provides connections to various initiatives, such as the Deep-ocean Assessment and Reporting of Tsunamis (DART) project, which is a component of the U.S. National Tsunami Hazard Mitigation Program. DART employs seafloor bottom pressure recorders (BPR) and surface buoys to identify and report tsunamis in real-time. DART system has two generations, with the second-generation DART II enabling bidirectional communication since 2008. This system can detect tsunamis as small as 1 cm and transmits information to ground stations through a GOES satellite link for early detection and data collection. Moreover, the NOAA Tsunami Stations offer information on tide stations equipped to detect tsunamis along various coastlines, while the IOC Sea Level Monitoring Facility provides real-time monitoring of sea level stations worldwide.

4.9 Southampton Data Observatory (SDO)

Southampton Data Observatory[9] collects data from various stakeholders in Southampton and Hampshire and combines them with nationally published data, providing access to professionals, businesses, the voluntary sector, citizens, and communities. The observatory has been developed in partnership with statutory partners, including the National Health Service (NHS) Hampshire, Southampton, and Isle of Wight (CCG), and Southampton Voluntary Services, with data contributions from other partners such as the National Office of Statistics (ONS), Hampshire Constabulary, Hampshire Fire and Rescue Service, and South Central Ambulance Service. SDO is accountable to the Southampton Health and Well-being Board and the Southampton Safe City Partnership for delivering the Joint Strategic Needs Assessment (JSNA) and the Safe City Strategic Assessment. It considers data protection issues and ensures sufficient safeguards and disclosure controls are in place to protect the identity of individuals. SDO's data types include links to demographics, economy, education, health, housing, road safety and environment specific to Southampton and its immediate surroundings within the United Kingdom.

4.10 National Ecological Observatory Network (NEON)

The National Ecological Observatory Network (NEON)[10] is an Open Data observatory funded by the National Science Foundation. Initiating its operational phase in the summer of 2019, NEON allows access to data on various topics, including climate, land use, and biodiversity. NEON adopts a specialized method for selecting its study locations spanning across the United States, including Hawaii and Puerto Rico, to capture a diverse range of environmental conditions. These areas were divided into 20 distinct zones, each comprising its own set of ecosystems, landscapes, and natural processes. This approach allowed NEON to gather extensive data on various aspects, such as the well-being of plants and animals, soil and water quality, and more, using state-of-the-art sensor technology and direct field observations. As a result, NEON provides standardized data on a continental scale collected from 81 field sites equipped with automated sensor systems and field instruments that continuously collect data on environmental factors. NEON's focus on long-term, standardized data collection allows researchers to track and analyse changes in ecological systems over time, providing insights into the impacts of climate change and other environmental factors. The program also encourages engagement with the scientific community, allowing researchers to use NEON data for their research projects.

4.11 India Urban Observatory (IUO)

The India Urban Observatory (IOU)[11] is an Open Data Observatory established by the Ministry of Housing and Urban Affairs (MoHUA) in India. IOU serves as a centralized hub for data and insights related to urban areas in the country. Its primary objective is to provide policymakers, researchers, and citizens access to reliable urban planning and development information. IUO aims to facilitate evidence-based decision-making and improve the efficiency of urban planning processes. It offers a wide range of data, including city-level indicators encompassing population statistics, infrastructure development, and economic growth. The observatory also provides data on various urban services such as water supply, sanitation, and waste management. IUO offers visualization and analysis tools to enhance data re-use and understanding. These tools enable users to explore and interpret the data in a user-friendly manner, promoting more significant insights and informed decision-making.

[9] data.southampton.gov.uk/

[10] data.neonscience.org/

[11] iuo.mohua.gov.in/portal/apps/sites

Table 2: Lists the Open Data Observatories and their data types.

Open Data Observatory	Data types
1. Terrestrial Ecosystem Research Network (TERN)	Mangroves, vegetation, soil, and phenology.
2. Channel Coastal Observatory (CCO)	Topographic and hydrographic surveys. Real-time data about waves, tides, weather and GPS data .
3. Urban Observatory Project (UOP)	Urban data include traffic flow, parking spaces, cycling docking, pedestrian count, weather data, air quality, water quality, seismic activity, noise-level, water-level (rainfall, river and tides), beehives, energy usage data, thermal imaging, visual and hyper-spectral mapping, social media feeds, employee feedback.
4. Global Forest Watch (GFW)	Satellite imagery, biodiversity, soil, climate dynamics, commodity supply chains, legal and administrative boundaries, fire alerts, and water resources.
5. Global Earth Observation System of Systems (GEOSS)	Satellite imagery, soil, atmospheric data, oceanographic data, geological data, biodiversity information, and climate metrics.
6. Earth Observing System Data and Information System (EOSDIS)	Soil, vegetation, spectroscopy, species distribution, and environmental change.
7. Grow Observatory (GROW)	Soil, temperature, and light level.
8. International Tsunami Information Center (ITIC)	Water-level data, historical tsunami, recent tsunamis.
9. Southampton Data Observatory (SDO)	Urban data include links to demographics, economy, education, health, housing, road safety and environmental data.
10. National Ecological Observatory Network (NEON)	Soil, atmospheric data for climate change, biogeochemistry, ecohydrology, land cover processes, organisms, populations, and communities.
11. Indian Urban Observatory (IUO)	Urban data include population statistics, infrastructure development, and economic growth, water supply, sanitation, and waste management..
12. Finnish Ecosystem Observatory (FEO)	Climate, soil, hydrology, biogeochemistry, and biodiversity.
13. Open Forest Observatory (OFO)	Forest drone imagery, forest structure metrics, tree sizes and species

4.12 The Finnish Ecosystem Observatory (FEO)

Finnish Ecosystem Observatory (FEO)[12] is a research and monitoring infrastructure that serves as a resource for obtaining high-quality ecosystem data across diverse terrestrial and aquatic ecosystems in Finland. FEO aims to facilitate access to data and observations for researchers, policymakers, and the general public. The data available through FEO encompass a wide range of parameters, including climate, hydrology, biogeochemistry, and biodiversity. To gather such data, FEO employs various monitoring techniques such as eddy covariance flux towers, radiometers, anemometers, and infrared gas analysers. FEO provides standardized field monitoring methods, calibration guidelines, and field data collection apps to ensure consistent and reliable data collection. One of the research at FEO, Mäyrä et al. [34] combined deep learning and remote sensing to improve forest monitoring, specifically by classifying tree species using airborne hyperspectral imagery and LIDAR data. Conducted in Finland's Boreal forests, the study demonstrated the effectiveness of high-resolution hyperspectral data and ground reference measurements in efficiently distinguishing between different tree species for improved biodiversity monitoring.

4.13 Open Forest Observatory (OFO)

The Open Forest Observatory (OFO)[13] employs drones and Artificial Intelligence (AI) to map and identify trees without needing traditional ground surveys. It establishes more than 100 forest plots, each roughly 25 hectares in size, to gather data vital for forest management in the face of issues such as droughts and wildfires. This initiative aims to improve research in forest ecology and disturbance ecology by creating three innovative cyberinfrastructure tools. The first tool is an AI-driven software workflow that efficiently transforms drone-captured imagery into detailed forest inventory information. This includes creating maps that accurately pinpoint individual trees, along with their size and species. The second tool is a searchable and open database that contains tree maps from over 100 plots, each covering 25 hectares. These plots are coordinated with existing forest inventory networks, such as the NSF's NEON, and cover a range of environmental and disturbance gradients. Lastly, the initiative includes comprehensive documentation and training programs, both online and in-person, to empower researchers to generate and share their data and tools. The software used in this observatory employs advanced photogrammetry to create 3D models of forest structures. It also uses multi-view computer vision, supported by neural networks, for accurate species classification and to filter out incorrect tree identifications. OFO is primarily funded by the National Science Foundation with additional support from The Nature Conservancy. OFO is housed in three academic institutions, the Department of Plant Sciences at the University of California, Davis, the CIRES Earth Lab at the University of Colorado, Boulder, and the Bio5 Institute at the University of Arizona. It relies on ground-reference forest inventory data from two sources, the USDA Forest Service Pacific Southwest Region and the National Ecological Observatory Network (NEON) 4.10. OFO also uses CyVerse and Jetstream2 computing infrastructure to support its operations.

[12] feosuomi.fi/en/

[13] openforestobservatory.org/

Table 3: Lists Open Data Observatories, including their geographic scope and the data themes they provide.

Open Data Observatory	<abbr>	Geographic Scope	Data API	Urban Data	Non-urban Data
1. Terrestrial Ecosystem Research Network	TERN	Australia	Yes		*
2. Channel Coastal Observatory	CCO	UK	Yes		*
3. Urban Observatory Project	UOP	UK	Yes	*	
4. Global Forest Watch	GFW	USA	Yes		*
5. Global Earth Observation System of Systems	GEOSS	Worldwide	Yes		*
6. Earth Observing System Data and Information System	EOSDIS	USA	Yes		*
7. Grow Observatory	GROW	Europe	No		*
8. International Tsunami Information Center	ITIC	Worldwide	Yes		*
9. Southampton Data Observatory	SDO	UK	No	*	*
10. National Ecological Observatory Network	NEON	North America	Yes		*
11. Indian Urban Observatory	IUO	India	No	*	
12. Finnish Ecosystem Observatory	FEO	Finland	No		*
13. Open Forest Observatory	OFO	USA	No		*

5 Data Themes and Management

This section delves into the data from the selected Open Data Observatories, examining their themes, sources and the methods employed in their processing. Our thematic analysis, referencing [35], revealed two main themes, urban data and no-urban data. We started the thematic analysis by reading through the data types collected for the selected observatories and taking notes. Table 2 shows data types managed by the selected observatories. Then, using NVIVO 12 software, we generated codes that helped us with the data themes. Words coded under "Transport" are indicative of `urban data`, while the words coded under "Soil Data" and "Seismic Events" fell under the `non-urban` data theme.

5.1 Urban Data

Urban data refer to information generated within the context of cities, including data on smart transportation, human behavior, demographics, and social systems. Smart transportation data involve metrics such as traffic flow, vehicle counts, public transit usage, parking availability, congestion levels, average speeds, journey times, and pedestrian counts. Several observatories, such as Urban Observatory Project (UOP), Southampton Data Observatory (SDO), and Indian Urban Observatory (IUO), collect and analyze various types of urban data. UOP focuses on providing real-time data on city transportation, including traffic congestion, parking availability, and public transit usage. SDO gathers links to data on transportation usage and behavior, including walking, cycling, and driving patterns, as well as transportation infrastructure like roads and public transit systems. Similarly, IUO collects data on transportation infrastructure (roads, highways, railways), transportation usage and behavior (vehicle ownership, mode choice, travel patterns). These observatories aim to provide insights into how urban transportation systems function and how they can be improved to better meet the needs of city residents. The data collected by these observatories cover a range of urban data metrics, as analyzed in Figure 3. Environmental data are collected in cities by one of the UOP observatories, to illustrate the concept, Figure 4 shows the environmental data types and parameter counts at Newcastle's Urban Observatory Project. Table 4 lists examples of the data types' parameters and their measuring units. Here, weather data include temperature, humidity, wind speed, and precipitation through a network of sensors deployed across Newcastle and the surrounding region, and the water level data entail river and tide Level. Raw data were obtained from (newcastle.urbanobservatory.ac.uk/api-docs/doc/sensors-dash-types-csv/).

5.2 Non-urban Data

Non-urban data refer to information and metrics collected from areas outside of city boundaries, including rural, wilderness, and natural environments. The non-urban data collected by our selected Open Data Observatories as listed in Table 2 span a wide array of environmental variables crucial for understanding ecosystem dynamics, climate change, and biodiversity. Terrestrial Ecosystem Research Network (TERN) focuses on mangroves, vegetation, soil, and phenological data. Channel Coastal Observatory (CCO) delivers topographic, hydrographic, meteorological, and GPS data relevant to coastal dynamics. Global Forest Watch (GFW) and Global Earth Observation System of Systems (GEOSS) both utilize satellite imagery to monitor biodiversity, climate dynamics, and environmental changes. Earth Observing System Data and Information System (EOSDIS) emphasizes soil, vegetation, and environmental change data. Grow Observatory (GROW) contribute data on soil conditions, temperature, light levels. National Ecological Observatory Network (NEON) offers comprehensive data on soil, atmosphere, biogeochemistry, and biodiversity to track climate change impacts. Finnish Ecosystem Observatory (FEO) and Open Forest Observatory (OFO) provide insights into forest structure. This diverse range of data supports a holistic understanding of Earth's non-urban environments, facilitating research and conservation efforts across multiple disciplines.

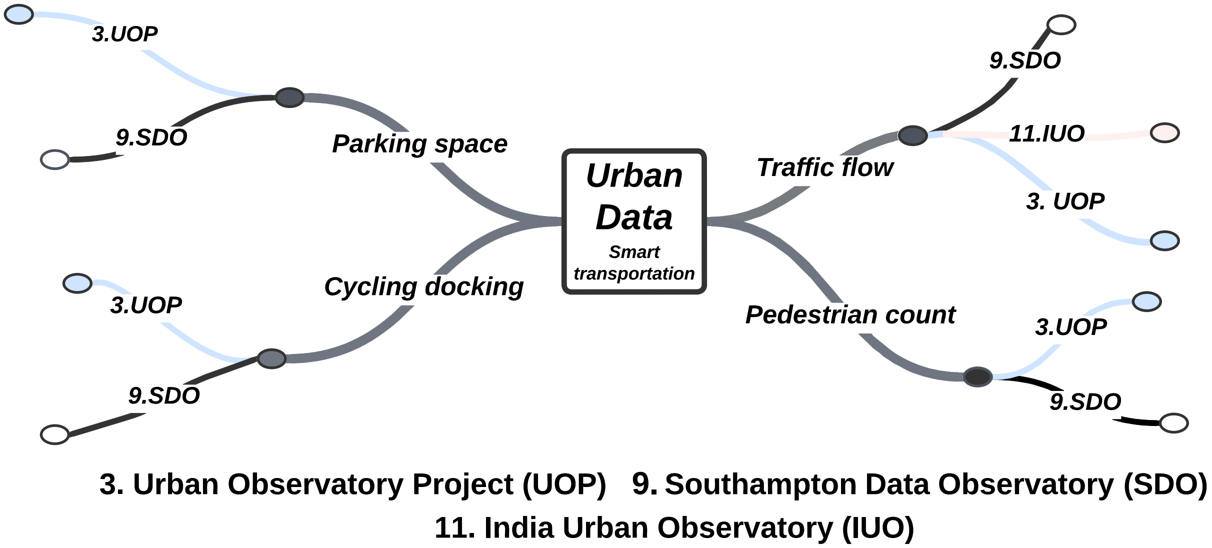


Figure 3: Transport data metrics collected by Open Data Observatories.

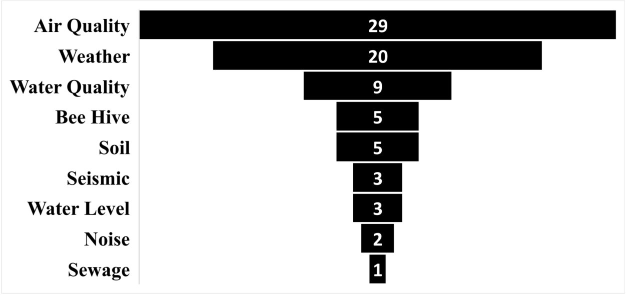


Figure 4: Newcastle Urban Observatory parameters count by data type.

Theme	Parameter	Unit
Air Quality	CO	ugm -3
Weather	Rain	mm
Water Quality	Dissolved Oxygen	mg/l
Bee hive	Brood nest temperature	Celsius
Soil	Soil Moisture	%VWC
Seismic	Vertical Displacement	m
Water Level	River Level	m
Noise	Sound	db
Sewage	Sewage Level	mm

Table 4: Newcastle Urban Observatory parameters examples and their measuring unit.

5.3 Data Sources

Open Data Observatories obtain data from Open Data portals, wireless sensor networks, and smart devices. Wireless Sensor Networks (WSNs) play a significant role in urban and non-urban data collection [36]. A notable example is the Urban Observatory Project (UOP), which uses a network of over 3600 sensors to capture diverse data streams from different physical environments. Grow Observatory (GROW) employs Flower Power sensors to monitor in-situ soil moisture, fertiliser levels, and air temperature at 15-minute intervals [37, 38]. Other technologies contributing data to these observatories include Lidar, ARGUS cameras, and satellites. ITIC- tsunami observatory provides data on water-levels, historical and recent tsunamis. The water-levels data sourced from the DART (Deep-ocean Assessment and Reporting of Tsunamis) system and the National Oceanic and Atmospheric Administration (NOAA) coastal water-level stations. The DART system obtains water-levels data from bottom pressure recorders on the seafloor, which measure water pressure with a resolution of approximately 1 mm of sea water and take 15-second averaged samples. The data are then transmitted to a ground station via satellite telecommunications, enabling real-time reporting. The DART II systems transmit standard mode data containing 24 estimated sea-level height observations at 15-minute intervals, once every six hours. Open Forest Observatory (OFO) uses drone imagery in a multi-step process to source data. First, numerous overlapping drone photos are taken from various angles to estimate each tree's three-dimensional structure. Next, the Canopy Height Model (CHM) is generated by processing the data to create a high-resolution Digital Surface Model (DSM) that displays the vegetation's height in each pixel above the ground. Then, an algorithm identifies individual

Table 5: Lists and compares the Open Data Observatories data sources.

Open Data Observatory	Wireless Sensors	Smart Devices	Citizen Data	Weather Stations	Digital Cameras	Satellite/Lidar	Field Surveys	Sensing Vehicles	Drones	Crowd-sourcing
1. Terrestrial Ecosystem Research Network (TERN)	*	*		*	*	*	*		*	
2. Channel Coastal Observatory (CCO)	*	*		*	*	*	*	*	*	
3. Urban Observatory Project (UOP)	*	*	*	*	*	*		*	*	*
4.Global Forest Watch (GFW)			*			*	*			*
5. Global Earth Observation System of Systems (GEOSS)	*	*		*		*	*		*	
6. Earth Observing System Data and Information System (EOSDIS)	*	*		*		*				
7. Grow Observatory (GROW)	*	*	*	*			*			*
8. International Tsunami Information Center (ITIC)	*	*		*		*				*
9. Southampton Data Observatory (SDO)		*	*							*
10. National Ecological Observatory Network (NEON)	*	*	*	*	*	*	*	*	*	*
11. Indian Urban Observatory (IUO)	*	*	*							*
12. Finnish Ecosystem Observatory (FEO)	*	*				*	*			
13. Open Forest Observatory (OFO)			*						*	*

trees in the forest area using drone imagery and CHM data, resulting in tree-level maps of forest stands. National Ecological Observatory Network (NEON) sources data and samples using a combination of automated instruments, field technicians, and airborne remote sensing. Terrestrial Ecosystem Research Network (TERN) gathers data using a variety of sensors, including Eddy covariance flux towers, heat flux plates, radiometers, anemometers, infrared gas analyzers, spectrometers, CosmOz soil moisture meters, groundwater bores, ecoacoustic sensors, phenocams, terrestrial laser scanners, UAV/drones, camera traps, and photopoints [39]. Table 5 groups and compares some of the observatories' primary data sources.

5.4 Data Processing

Most of the selected Open Data Observatories develop open-source software to harmonize and integrate diverse open data sources. Such data processing techniques are set to realize the potential value of Open Data by making them FAIR (findable, accessible, interoperable, and reusable) for researchers, decision-makers, and the broader community. Terrestrial Ecosystem Research Network (TERN) includes several tools and applications for data processing and analysis. To mention a few, SHaRED Data Submission (shared.tern.org.au) allows ecologists to upload their research data to the Australian Ecological Knowledge and Observation System (ÆKOS) and assists in creating structured metadata and assigns Digital Object Identifiers (DOIs). CoESRA Virtual Desktop (coesra.tern.org.au) enables access to a web-based virtual desktop from any device and equipped with scientific software such as RStudio, Jupyter Notebook, and QGIS. EcoImages (ecoimages.tern.org.au) serves as a repository that organizes images of vegetation, soil, and landscapes. To process live streams of diverse data, Urban Observatory Project (UOP) deploys real-time machine learning models on CCTV feeds and uses data queues, data sharding, and many edge processors along with hourly replication to reduce the occurrence of problems during live data streaming. Global Forest Watch (GFW) uses machine learning for detecting and mapping tree cover and loss, involving image segmentation, classification, and change detection to produce forest datasets. At ITIC tsunami observatory, raw data from the tide gauges and DART buoys are processed by the PMEL (Pacific Marine Environmental Laboratory) and NGDC (National Geophysical Data Center) to remove errors and archive. National Ecological Observatory Network (NEON) developed proprietary software to process raw data from sensors and field apps into standardized data products. NEON employs a unique "NEON Ingest Conversion Language" to establish and update data processing protocols as necessary. Open Forest Observatory (OFO) presents three cyber-infrastructure innovations to enhance data processing capabilities. These include a scalable, reproducible, AI-enabled software workflow for converting drone imagery into forest inventory data, a searchable database of tree maps that are aligned with forest inventory plot networks and accessible to the public, and documentation and training resources to encourage researchers to contribute their own data and analytical tools. Moreover, research [40], which offers resources for individuals who want to create efficient and detailed tree maps of conifer forests without requiring extensive customization of image acquisition and processing parameters.

5.5 Data Visualization

Data visualizations transform information into meaningful graphical representations that intended audiences can interpret [41]. The selected observatories employ various visualizations techniques to present and communicate their collected data effectively. Visualizations include static and interactive maps [42], charts such as time series, scatter plots, histograms [43], bar, and pie graphs. TERN-ANU Landscape Data Visualizer (maps.tern.org.au) is a user-friendly atlas that offers comprehensive spatial data on Australian landscapes, soil, ecosystems, and water resources. The data can be visualized on a map and explored through time-series data for specific locations. Urban Observatory Project (UOP) employs interactive maps, digital comparison tools, thematic cartography, real-time data visualizations to explore and understand urban dynamics. National Ecological Observatory Network (NEON) collaborates with Google to enhance the visualization and accessibility of its environmental data via the Google Cloud Platform, incorporating tools such Google Earth Engine and BigQuery. This integration enables users to engage with and visualize extensive NEON datasets directly in the cloud. Global Forest Watch (GFW) visualizes data through its Open Data portal, interactive map features, downloadable datasets, geospatial monitoring frameworks, and software like the Forest Trends Analysis Tool. NASA's Earth Observing System Data and Information System (EOSDIS) visualizes data through the Earthdata Cloud, which provides users with free access to NASA Earth science data for research purposes. ITIC - tsunami observatory provides real-time and historical tsunami data through 1-minute water level readings, event search tools, and interactive maps. These resources offer numerical and graphical representations of water-levels, crucial for early tsunami detection. India Urban Observatory (IUO) employs diverse visualization methods such as data stories, interactive maps using ArcGIS, thematic dashboards, and Open Data portal to share urban insights with stakeholders such as government institutions, researchers, and the public. Grow Observatory (GROW) uses interactive maps, visualization tools to effectively visualize the soil moisture data it collects and share them with its stakeholders [44].

6 Research Challenges

Establishing Open Data Observatories involves addressing various challenges related to integrating diverse data sources and systems. These challenges include ensuring data interoperability, scalability, and replicability since each data source has its own design and computing specifications. Combining and merging disparate data, without careful consideration, can lead to service conflicts, resulting in degraded data quality, loss of data provenance, and potential privacy breaches. This section explores these challenges, as depicted in Figure 5 and how each observatory addresses each challenge.

6.1 Data Integration

Data integration is the process of combining data from disparate sources into a unified view [45]. Integrating heterogeneous data can positively impact decision-making, however, achieving valid integration faces many challenges, as stated by many researchers such as [46, 47, 48]. Open Data Observatories may suffer primarily from the Interoperability challenge, which refers to the difficulty of integrating and harmonising disparate data sources and systems. It ensures that different datasets with varying formats, structures, and standards can effectively work together and exchange information. Interoperability is one of the Open Data FAIR principles as explained in section 2.1 and a significant obstacle for Open Data Observatories [49, 47]. To overcome this challenge, several observatories implemented various strategies. For instance, Terrestrial Ecosystem Research Network (TERN) harmonized plot-based ecology using EcoPlots (ecoplots.tern.org.au), a semantic data integration system that maps each data source to the TERN Plot ontology. Urban Observatory Project (UOP) deployed a platform called the "Urban Data Exchange (UDX)" (urbandatacollective.com/urban-observatories-case-study) that acts as a central hub for onboarding, harmonizing, and serving the real-time data streams from the different urban observatory systems. NASA's Earth Observing System Data and Information System (EOSDIS) enhanced data interoperability through standardization of data formats and metadata, a distributed and interoperable architecture across nodes like the Science Investigator-led Processing Systems (SIPS) and Distributed Active Archive Centers (DAACs), which enabled efficient data retrieval [50].

6.2 Data Quality

Applied research defined the term data quality differently [51], a commonly used definition by Strong et al. [52] describing data quality as data fit for the intended purpose. Byabazaire et al. [53] and Taleb et al. [54] testified that data quality is a mature research topic in big data and database management. However, Perez-Castillo et al. [51] claimed its youth in Smart Connected Products (SCP) [55] and the Internet of Things. Data quality plays a significant role in Open Data Observatories, as a sufficient quality level can build trust between the cyber and physical world [53, 51]. Each observatory addresses data quality using different strategies, Urban Observatory Project (UOP) manages data quality by using automated checks for data anomalies, calibrating sensors against precision stations, and incorporating user feedback. They also recognize the limitations of low-cost sensors and design their data use accordingly. Global Forest Watch (GFW) ensures data are up-to-date by automating updates or requesting providers to notify them of changes. NASA's Earth Observing System Data and Information System (EOSDIS) methodology ensures metadata quality of Earth observation data hinges on a framework prioritizing correctness, completeness, and consistency. NASA uses automated and manual reviews to identify and rectify issues, demanding active collaboration with data providers to implement enhancements [56]. Channel Coastal Observatory (CCO) and National Ecological Observatory Network (NEON) implement quality assurance and control practices. CCO ensures the reliability of marine observations, flagging poor data but not eliminating them, while NEON applies rigorous quality measures to ensure data quality. For example, observation system data use mobile apps with constraints and validation rules. Instrument System data benefit from sensor placement, maintenance, and calibration. Airborne Remote Sensing data are calibrated and tested pre- and post-flight. Automated checks and expert reviews ensure reliability, while flags and metrics provide transparency. The India Urban Observatory (IUO) handles quality through trusted data sources, accuracy, transparency, and interactive visualizations but has limitations in completeness and update frequency. Open Forest Observatory (OFO) prioritizes data quality through standardized, open-source workflows for drone-based forest mapping, accessible via GitHub. It also employs cloud-based tools to process drone imagery into detailed forest maps, facilitating ease of use as well as a central database to support data sharing and quality enhancement through community feedback.

6.3 Data Provenance

Data provenance, which traces the origins and lineage of data, is crucial in Open Data Observatories. Maintaining rigorous data provenance allows observatories to ensure data transparency, reliability, and reproducibility [57, 58, 59]. Terrestrial Ecosystem Research Network (TERN) releases weather data accompanied by their lineage, including the type and model of the automatic weather station used for collection. The specific location and characteristics of the site. The instruments used for measuring different weather parameters, along with their accuracy and resolution. The methodology for data recording and the intervals at which data were stored. The procedures followed in case of sensor failure, including using alternative data sources for gap filling and indicating this within the dataset. The availability of data and contact information for access to more granular data (hourly data). Similarly, Southampton Data Observatory (SDO) commits to full metadata inclusion for all its published data compendiums and resources, encompassing data sources and time frames. National Ecological Observatory Network's (NEON) dedication to rich metadata and thorough documentation strengthens the provenance and traceability of its data offerings. This commitment includes the provision of Digital Object Identifiers (DOIs) for NEON data packages, enhancing their findability and citability. NEON's approach to data provenance involves metadata management, adherence to FAIR principles, data citation tracking, and handling data from diverse sources, focusing on transparency and accessibility. In a different vein, research [60] recommends applying blockchain technology for data provenance. Blockchain can revolutionize how data are managed, enhancing transparency, security, and trust. By leveraging its immutable ledger, data integrity and authenticity can be guaranteed, ensuring that once data are recorded, it cannot be altered. Moreover, the decentralization offered by blockchain reduces risks associated with centralized data storage by distributing data across a network, thus enhancing data resilience and accessibility through peer-to-peer sharing. Furthermore, blockchain's encryption and smart contracts safeguard sensitive data and automate data access permissions, ensuring only authorized access. It also offers a transparent audit trail for all data modifications and transactions, facilitating traceable data lineage and enforcing open data licenses automatically.

6.4 Data Privacy

Data privacy is critical in protecting personal and sensitive information from unauthorised access and disclosure. Open Data Observatories implemented various measures to address data privacy challenges, including data anonymization, access controls, and encryption [61, 62, 63, 64, 59]. These observatories handle massive amounts of data from various data sources through orderly collection, aggregation, and analytics. However, these data may contain sensitive details such as personally identifiable information and endangered species locations [63, 65, 66, 67, 68, 69]. Terrestrial Ecosystem Research Network (TERN), Channel Coastal Observatory (CCO), and Urban Observatory Project (UOP) all have dedicated privacy statements that outline their data privacy practices. These include compliance with regulations like GDPR, providing privacy notices, defining lawful data processing, implementing security measures, and respecting user rights. Similarly, the Global Forest Watch (GFW) and Global Earth Observation System of Systems (GEOSS) approach data privacy through transparency, consent-based processing, security, and clear points of contact for users. NASA's EOSDIS also has a privacy policy that emphasizes protection and proper use of information in line with relevant laws and regulations. Grow Observatory (GROW) addresses privacy by using an open data license, collecting only anonymized sensor data without personal identifiers, and operating under institutional oversight. ITIC-tsunami observatory's privacy policy covers aspects like cookies, email handling, and user rights under the Privacy Act. Southampton Data Observatory adheres to the overall privacy policy of Southampton City Council, while National Ecological Observatory Network (NEON) securely manages user accounts, anonymizes data reporting, and applies Creative Commons licensing. In contrast, India Urban Observatory (IUO) has a privacy-focused approach, avoiding automatic capture of personal information and only collecting such data if explicitly provided by users, with appropriate security measures. Finally, Open Forest Observatory focuses on openly sharing its forest mapping data and tools, rather than collecting or managing personal user information, implying a commitment to data transparency and accessibility.

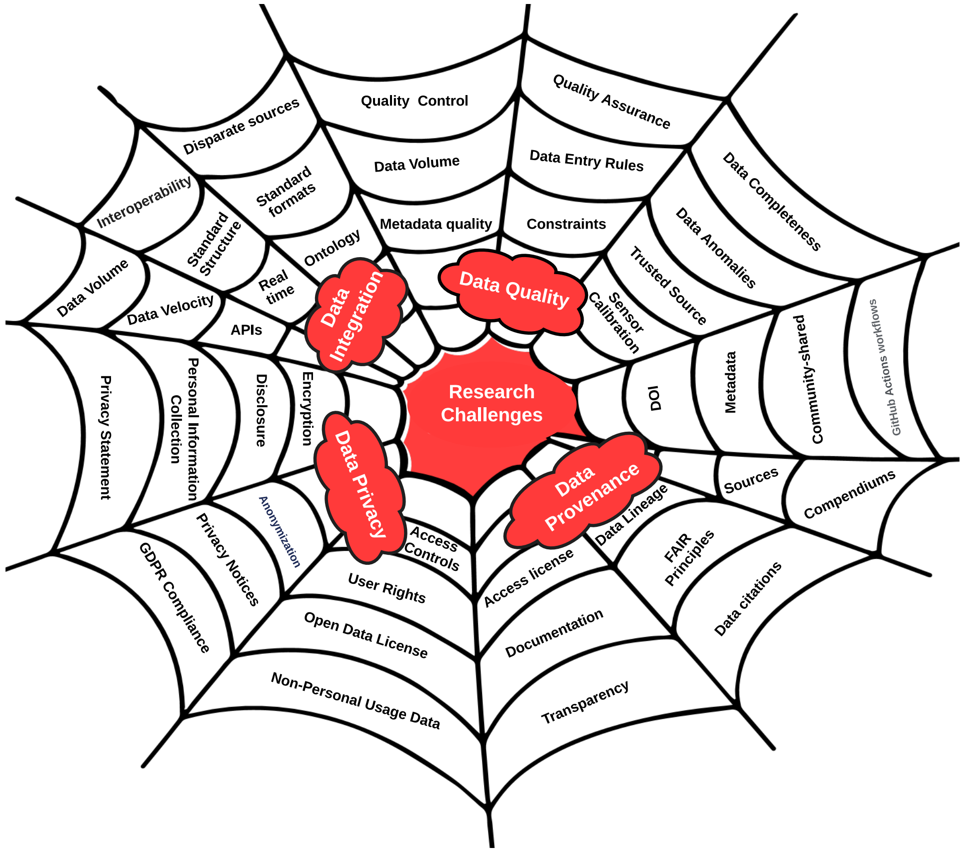


Figure 5: This diagram captures the intricate web of research challenges in data management, segmented into four primary categories, Data Integration, Data Quality, Data Privacy, and Data Provenance. Each challenge extends into related subtopics and approaches to overcome them that touch the periphery of the web, symbolizing the complex and interconnected nature of these issues. The visual metaphor of a spider web aptly conveys the idea that each aspect is a critical thread in the overall structure of data management.

7 Discussion

The selected Open Data Observatories are pushing the boundaries of the FAIR principles through the creation of open-source software and the application of advanced data processing methods. Terrestrial Ecosystem Research Network (TERN), for example, not only simplifies the process of data submission and organization through the SHaRED Data Submission tool but also promotes data discoverability and citability with structured metadata and Digital Object Identifiers (DOIs). On another front, Urban Observatory Project (UOP)'s deployment of machine learning models for the real-time analysis of CCTV data showcases innovative data handling techniques. The application of machine learning by Global Forest Watch (GFW) for analyzing forest coverage highlights the pivotal role of advanced technology in the efforts to preserve natural habitats. Moreover, proprietary software developed by National Ecological Observatory Network (NEON) and the drone imagery processing innovations introduced by Open Forest Observatory (OFO) mark progress in data standardization and quality improvement. Through these diverse data processing efforts, these observatories are not just elevating the value of Open Data but are also providing deeper insights into environmental and urban challenges, thereby equipping researchers and stakeholders with the necessary resources for informed decision-making. Urban data observatories such as UOP, SDO and IUO provide essential insights into the fabric of city life, tracking urban expansion and infrastructure development to support urban sustainability, smart city analytics [9, 10, 11, 12]. Observatories like CCO and ITIC tsunami observatory contribute to our preparedness and response strategies for coastal hazards, safeguarding communities and ecosystems, and relying on real-time and historical non-urban data. The observatories offered a variety of data types, with soil, vegetation, and climate data being among the most common. Our study embarked on facilitating the development of new Open Data Observatories. This effort led us through a complex maze of challenges, from making different data sources work together to ensuring the data were reliable and protected. Interoperability, a cornerstone of the FAIR principles for Open Data, presents a notable challenge in data integration for Open Data Observatories. Efforts, including the implementation of semantic data systems for real-time data integration, demonstrate advancements in overcoming this obstacle. Similarly, adopting standardized formats and metadata improved the ease of access and usefulness of integrated data. Different observatories adopt tailored strategies to maintain and enhance the quality of their data. For instance, some focus on rigorous quality control measures and real-time data verification, while others prioritize the accuracy, transparency, and up-to-dateness of their data through both automated systems and manual oversight. These methods reflect a shared commitment across observatories to uphold the integrity and reliability of their data. Tracing data back to their origins, a practice known as data provenance is essential for establishing trust and ensuring transparency within data-centric environments. Observatories that rigorously document their data sources set a benchmark for data management, enhancing both the reliability and reproducibility of their data. Using detailed metadata documentation and Digital Object Identifiers (DOIs) improves the traceability and accessibility of data. Furthermore, adherence to the FAIR principles and metadata handling amplifies the integrity of the collected data. Implementing standardized workflows and open-source software also contributes to transparency, making it easier for the wider scientific community to verify data. Protecting Data Privacy: the methods used by different observatories to tackle data privacy issues demonstrate their commitment to meeting regulatory standards, yet they vary in their approaches to data collection, use, and management. For example, while some observatories comply with the General Data Protection Regulation (GDPR), others emphasize data anonymization and the use of open data licenses to reduce the collection of personal data. The depth and breadth of these privacy policies also differ significantly. Some observatories have developed comprehensive policy frameworks that address a broad range of legal and operational concerns, whereas others adopt more focused privacy strategies that rely on obtaining explicit user consent before gathering personal data. Few observatories protected threatened species by reducing their taxonomic identification precision to a safer classification level, and in certain areas, such data were completely withheld from publication. This careful processing respects both data integrity and ecological sensitivities, supporting robust scientific analysis while safeguarding vulnerable taxa.

Study limitations: Determining the precise size and quality of data was difficult due to variations among the chosen observatories; ideally, a summary of the data inventory should have been provided. A model like that of 4TU.ResearchData (data.4tu.nl/) would have simplified the inventory process. Consequently, this information was not readily available in each observatory examined. In addition, our study lacked detailed information on the funding and sponsorships of the observatories, which can be useful for understanding their sustainability and longevity. Building Open Data Observatories is challenging but also filled with potential for significant impact. The collaboration between technology, policy, and practice is key to navigating these challenges, ensuring that observatories can thrive long-term. As we move forward, the lessons learned from our work will undoubtedly influence the growth and development of open data ecosystems. Table 6 lists some advantages and limitations of the selected observatories and takeaways that can assist the establishment of new Open Data Observatories.

8 Conclusion

This study analyzed thirteen Open Data Observatories, offering data that spans both urban and non-urban settings on a global and regional scale. These observatories, including global initiatives such as GEOSS and ITIC, and region-specific ones such as GFW, EOSDIS, and OFO in the USA, GROW, FEO, CCO, UOP, SDO in Europe, IUO in Asia, and TERN in Australia, were evaluated for their core features, data accessibility, and usability. Despite the inherent difficulty in comparing the observatories due to their varied sizes and development phases, we noted significant collaborations and connections, for example, between NEON and OFO, and between GROW and GEOSS. The data were organized into urban and non-urban themes, highlighting commonalities in data types and processing approaches across the observatories. Challenges related to integrating diverse data sources while maintaining their reliability and integrity were explored, revealing that solutions varied widely depending on the source of the data. We pinpointed specific strengths and weaknesses for each observatory, forming the basis for our recommendations for future developments. These findings mark the importance of collaboration, the standardization of data, and adaptable strategies for overcoming integration challenges, essential for developing new Open Data Observatories. These results highlight the critical role of working together, standardizing data, and developing flexible methods to navigate the complexities of data integration.

Table 6: Strengths and limitations of the selected Open Data Observatories, future recommendations and some takeaways.

Data Observatory	Strengths	Limitations	Future Recommendation	Takeaways
1. TERN[14]	High-quality data on environmental monitoring, along with tools and expertise, provided to researchers.	Limited coherent national capability for monitoring freshwater ecosystems.	Integrating blockchain for data provenance and artificial intelligence for Linked Data.	Semantic data integration and the Threatened Species Index (TSX)[15]
2. CCO[16]	Access to tools and models to analyze coastal data and predict morphological changes.	Outsourcing data storage may impose security concerns.	Incorporate extreme events alert system.	Extreme events analysis.
3. UOP[17]	Ability to provide a wide variety of real-time and historical data on different aspects of the urban environment.	Urban observatories do not extend their coverage to all cities across the UK, resulting in a limited geographical reach.	Lack of evident research documenting the positive impact of the project (e.g., reduce crime rates).	Real-time data integration.
4. GFW[18]	Forest Watcher mobile app for real-time threat detection, GFW Pro for managing deforestation risks in supply chains, grants and fellowships.	Limited data lineage.	Provide details how data are collected and evolved over time to enhance data provenance.	Real-time forest monitoring via satellite imagery and remote sensing.
5. GEOSS[19]	Data platform flexibility enabling users to adapt it to their needs.	GEOSS does not guarantee its Earth Observations' accuracy or take responsibility for their use.	Invest in quality assurance and control.	Platform flexibility.
6. EOSDIS[20]	Global, long-term and reliable Open Data.	Limited validation for satellite-based data with ground-based measurements.	Consider real-time update and alert system for extreme events.	Data long-term archiving useful for analysis and training AI applications.
7. GROW[21]	Empowers citizens and communities to have a say on soil and climate matters across Europe.	Limited data types.	Integrate more data sources as air quality and noise level.	Citizen science.
8. ITIC[22]	Centralized and authoritative source for providing real-time information, and warnings about tsunami events and risks.	Data quality and provenance challenges causing errors in tsunami database.	Addressing data quality for improving the reliability and usability of the tsunami data.	Alert system
9. SDO[23]	Crowd-sourcing, allowing citizens to understand local issues and contribute to problem-solving in urban development and sustainability matters.	Lack of real-time data and APIs.	Extend geographic scope.	Civic engagement and transparency.
10. NEON[24]	Open Data with good quality and sufficient documentation.	A location of some Instrument System (IS) Data sensors is seasonally adjusted or removed due to unsuitable conditions for measurement.	Implement robust power solutions (solar panels or wind turbines) for OKSR site where operations cease during winter.	Educational resources such as the learning and code hub.
11. IOU[25]	Wide range of urban data.	Inconsistent data frequency.	Consider using applications for data quality assurance.	Data diversity.
12. FEO[26]	Ongoing monitoring and research initiatives related to Finland ecosystems.	Limited data coverage, lack of data privacy statement.	Expand geographic scope.	Platform presentation in multiple languages.
13. OFO[27]	Educational resources to understand forests.	Limited data diversity, privacy policy not shared in the website.	Integrate more remote sensing wildlife data, supplemented with contextual information	Drones and Artificial Intelligence (AI).

References

[1] Open Knowledge Foundation. Open data handbook. `https://opendatahandbook.org/`, 2021. Accessed: 2024-03-07.

[2] Adrienne Colborne and Michael Smit. Characterizing disinformation risk to open data in the post-truth era. *J. Data and Information Quality*, 12(3), jun 2020.

[3] Philipp Lämmel, Benjamin Dittwald, Lina Bruns, Nikolay Tcholtchev, Yuri Glikman, Silke Cuno, Mathias Flügge, and Ina Schieferdecker. Metadata harvesting and quality assurance within open urban platforms. *J. Data and Information Quality*, 12(4), oct 2020.

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

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

[6] Shelley Stall, Maryann E. Martone, Ishwar Chandramouliswaran, Mercè Crosas, Lisa Federer, Julian Gautier, Mark Hahnel, Jennie Larkin, Daniella Lowenberg, Nicole Pfeiffer, Ida Sim, Tim Smith, Ana E. Van Gulick, Erin Walker, Julie Wood, Maryam Zaringhalam, and Alberto Zigoni. Generalist repository comparison chart, July 2020. Thank you the American Geophysical Union for designing the document.

[7] Corinna Gries, Paul C Hanson, Margaret O'Brien, Mark Servilla, Kristin Vanderbilt, and Robert Waide. The environmental data initiative: Connecting the past to the future through data reuse. *Ecology and Evolution*, 13(1):e9592, 2023.

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

[9] Harvey Miller, Kelly Clifton, Gulsah Akar, Kristin Tufte, Sathya Gopalakrishnan, John MacArthur, Elena Irwin, Rajiv Ramnath, and Jonathan Stiles. Urban Sustainability Observatories: Leveraging Urban Experimentation for Sustainability Science and Policy. *Harvard Data Science Review*, 3(2), may 14 2021. https://hdsr.mitpress.mit.edu/pub/zunejoo2.

[10] Vaia Moustaka, Athena Vakali, and Leonidas G. Anthopoulos. A systematic review for smart city data analytics. *ACM Comput. Surv.*, 51(5), dec 2018.

[11] Meiyi Ma, Sarah M. Preum, Mohsin Y. Ahmed, William Tärneberg, Abdeltawab Hendawi, and John A. Stankovic. Data sets, modeling, and decision making in smart cities: A survey. *ACM Transactions on Cyber-Physical Systems*, 4(2), 2019.

[12] Yingjian Liu, Meng Qiu, Chao Liu, and Zhongwen Guo. Big data challenges in ocean observation: a survey. *Personal and Ubiquitous Computing*, 21:55–65, 2017.

[13] John Doe. My creative commons work. Creative Commons Attribution-ShareAlike 4.0 International License, 2023.

[14] Vishanth Weerakkody, Zahir Irani, Kawal Kapoor, Uthayasankar Sivarajah, and Yogesh K. Dwivedi. Open data and its usability: an empirical view from the Citizen's perspective. *Information Systems Frontiers*, 19(2):285–300, 2017.

[15] Sunlight Foundation. Ten Principles for Opening Up Government Information. *Sunlight Foundation*, (October 2007):3, 2010.

[16] Jan Kucera, Dusan Chlapek, Jakub Klímek, and Martin Necaskỳ. Methodologies and best practices for open data publication. In *DATESO*, pages 52–64, 2015.

[17] Victoria Wang and David Shepherd. Exploring the extent of openness of open government data - A critique of open government datasets in the UK. *Government Information Quarterly*, 37(1):101405, 2020.

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

[19] Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T Evelo, et al. Fair principles: interpretations and implementation considerations, 2020.

[20] Louise Bezuidenhout. Being fair about the design of fair data standards. *Digital Government: Research and Practice*, 1(3):1–7, 2020.

[21] Rob Kitchin and Tracey P Lauriault. Small data in the era of big data. *GeoJournal*, 80:463–475, 2015.

[22] Alison Cooke, Debbie Smith, and Andrew Booth. Beyond pico: the spider tool for qualitative evidence synthesis. *Qualitative health research*, 22(10):1435–1443, 2012.

[23] James Cleverly, Derek Eamus, Will Edwards, Mark Grant, Michael J Grundy, Alex Held, Mirko Karan, Andrew J Lowe, Suzanne M Prober, Ben Sparrow, et al. Tern, australia's land observatory: addressing the global challenge of forecasting ecosystem responses to climate variability and change. *Environmental Research Letters*, 14(9):095004, 2019.

[24] Global Forest Watch. Global forest watch. *World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002)*, 2002.

[25] Eliot Christian. Planning for the global earth observation system of systems (geoss). *Space Policy*, 21(2):105–109, 2005.

[26] Max Craglia, Jiri Hradec, Stefano Nativi, and Mattia Santoro. Exploring the depths of the global earth observation system of systems. *Big Earth Data*, 1(1-2):21–46, 2017.

[27] Jeanne Behnke. Nasa's earth observing system data and information system (eosdis). Technical report, 2017.

[28] David T Barnett, Peter B Adler, Benjamin R Chemel, Paul A Duffy, Brian J Enquist, James B Grace, Susan Harrison, Robert K Peet, David S Schimel, Thomas J Stohlgren, et al. The plant diversity sampling design for the national ecological observatory network. *Ecosphere*, 10(2):e02603, 2019.

[29] Petteri Vihervaara, Saku Anttila, Peter Kullberg, Pekka Härmä, Markus Törmä, Tytti Jussila, Kaisu Aapala, Risto Heikkinen, Janne Mäyrä, Mikko Kervinen, et al. Finnish ecosystem observatory (feo)-operationalizing remote sensing analyses for threatened habitats and biodiversity monitoring. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 735–738. IEEE, 2021.

[30] EM Lee. Reflections on the decadal-scale response of coastal cliffs to sea-level rise. *Quarterly Journal of Engineering Geology and Hydrogeology*, 44(4):481–489, 2011.

[31] Stilianos Contarinis, Athanasios Pallikaris, and Byron Nakos. The Value of Marine Spatial Open Data Infrastructures-Potentials of IHO S-100 Standard t Become the Universal Marine Data Model. *Journal of Marine Science and Engineering*, 8(8):564, 2020.

[32] Travis Mason and Thomas Dhoop. Cover photograph: Datawell Directional Waverider Mk III in Weymouth Bay Photo courtesy of Fugro GB Marine Limited National Network of Regional Coastal Monitoring Programmes of England Quality Assurance & Quality Control of Wave Data. 2017.

[33] Luke Smith and Mark Turner. Building the Urban Observatory : Engineering the largest set of publicly available real-time environmental urban data in the UK. 21:10456, 2019.

[34] Janne Mäyrä, Sarita Keski-Saari, Sonja Kivinen, Topi Tanhuanpää, Pekka Hurskainen, Peter Kullberg, Laura Poikolainen, Arto Viinikka, Sakari Tuominen, Timo Kumpula, et al. Tree species classification from airborne hyperspectral and lidar data using 3d convolutional neural networks. *Remote Sensing of Environment*, 256:112322, 2021.

[35] David Byrne. A worked example of braun and clarke's approach to reflexive thematic analysis. *Quality & quantity*, 56(3):1391–1412, 2022.

[36] Ammar Gharaibeh, Mohammad A. Salahuddin, Sayed Jahed Hussini, Abdallah Khreishah, Issa Khalil, Mohsen Guizani, and Ala Al-Fuqaha. Smart cities: A survey on data management, security, and enabling technologies. *IEEE Communications Surveys Tutorials*, 19(4):2456–2501, 2017.

[37] Karoly Zoltan Kovács, Drew Hemment, Mel Woods, Naomi K. van der VELDEN, Angelika Xaver, Rianne H. Gi Esen, Victoria J. Burton, Natalie L. Garrett, Luca Zappa, Deborah Long, Endre Dobos, and Rastislav Skalsky. Citizen observatory based soil moisture monitoring - The GROW example. *Hungarian Geographical Bulletin*, 68(2):119–139, 2019.

[38] M. Woods, D. Hemment, R. Ajates, A. Cobley, A. Xaver, and G. Konstantakopoulos. GROW Citizens' Observatory: Leveraging the power of citizens, open data and technology to generate engagement, and action on soil policy and soil moisture monitoring. *IOP Conference Series: Earth and Environmental Science*, 509(1):10–12, 2020.

[39] Victoria M. Scholl, Megan E. Cattau, Maxwell B. Joseph, and Jennifer K. Balch. Integrating national ecological observatory network (neon) airborne remote sensing and in-situ data for optimal tree species classification. *Remote Sensing*, 12(9), 2020.

[40] Derek JN Young, Michael J Koontz, and JonahMaria Weeks. Optimizing aerial imagery collection and processing parameters for drone-based individual tree mapping in structurally complex conifer forests. *Methods in Ecology and Evolution*, 13(7):1447–1463, 2022.

[41] Hsiao-Fang Yang, Chia-Hou Kay Chen, and Kuei-Ling Belinda Chen. Using Big Data Analytics and Visualization to Create IoT-enabled Science Park Smart Governance Platform. In Fiona Fui-Hoon Nah and Keng Siau, editors, *HCI in Business, Government and Organizations. Information Systems and Analytics*, pages 459–472, Cham, 2019. Springer International Publishing.

[42] Michael Evans, Dragomir Yankov, Pavel Berkhin, Pavel Yudin, Florin Teodorescu, and Wei Wu. LiveMaps: Converting Map Images into Interactive Maps. SIGIR '17, pages 897–900. ACM, 2017.

[43] Anurag Srivastava. *Mastering Kibana 6. x: Visualize Your Elastic Stack Data with Histograms, Maps, Charts, and Graphs*. Packt Publishing, Limited, Birmingham, 2018.

[44] Samuel Stehle and Rob Kitchin. Real-time and archival data visualisation techniques in city dashboards. *International Journal of Geographical Information Science*, 34(2):344–366, 2020.

[45] Yassine Chahid, Mohamed Benabdellah, and Abdelmalek Azizi. Internet of things protocols comparison, architecture, vulnerabilities and security: State of the art. *ACM International Conference Proceeding Series*, pages 0–5, 2017.

[46] Ana Maria de Carvalho Moura, Fabio Porto, Vania Vidal, Regis Pires Magalhães, Macedo Maia, Maira Poltosi, and Daniele Palazzi. A semantic integration approach to publish and retrieve ecological data. *International Journal of Web Information Systems*, 11(1):87–119, jan 2015.

[47] Maggi Bansal, Inderveer Chana, and Siobhán Clarke. A survey on iot big data: Current status, 13 v's challenges, and future directions. *ACM Comput. Surv.*, 53(6), dec 2020.

[48] Xin Luna Dong and Divesh Srivastava. Big data integration. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, pages 1245–1248, 2013.

[49] Mahda Noura, Mohammed Atiquzzaman, and Martin Gaedke. Interoperability in internet of things: Taxonomies and open challenges. *Mobile Networks and Applications*, 24:796–809, 2019.

[50] Hampapuram K Ramapriyan and John Moses. Nasa's earth science data systems: Lessons learned and future directions. In *Proceedings of the 2010 Roadmap for Digital Preservation Interoperability Framework Workshop*, pages 1–9, 2010.

[51] R Perez-Castillo, A G Carretero, M Rodriguez, I Caballero, M Piattini, A Mate, S Kim, and D Lee. Data Quality Best Practices in IoT Environments. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*, pages 272–275, 2018.

[52] Diane M Strong, Yang W Lee, and Richard Y Wang. Data Quality in Context. *Commun. ACM*, 40(5):103–110, 5 1997.

[53] J Byabazaire, G O'Hare, and D Delaney. Data Quality and Trust : A Perception from Shared Data in IoT. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pages 1–6, 6 2020.

[54] I Taleb, M A Serhani, and R Dssouli. Big Data Quality: A Survey. In *2018 IEEE International Congress on Big Data (BigData Congress)*, pages 166–173, 7 2018.

[55] Pai Zheng, Xun Xu, and Chun Hsien Chen. A data-driven cyber-physical approach for personalised smart, connected product co-development in a cloud-based environment. *Journal of Intelligent Manufacturing*, 31(1):3–18, 2020.

[56] Kaylin Bugbee, Jeanné le Roux, Adam Sisco, Aaron Kaulfus, Patrick Staton, Camille Woods, Valerie Dixon, Christopher Lynnes, and Rahul Ramachandran. Improving discovery and use of nasa's earth observation data through metadata quality assessments. *Data Science Journal*, 20:17–17, 2021.

[57] Adel Alkhalil and Rabie A. Ramadan. IoT Data Provenance Implementation Challenges. *Procedia Computer Science*, 109(2014):1134–1139, 2017.

[58] Henry Pearce. The (UK) Freedom of Information Act's disclosure process is broken: where do we go from here? *Information and Communications Technology Law*, 29(3):354–390, 2020.

[59] Rui Hu, Zheng Yan, Wenxiu Ding, and Laurence T. Yang. A survey on data provenance in IoT. *World Wide Web*, 23(2):1441–1463, 2020.

[60] Karl Werder, Balasubramaniam Ramesh, and Rongen Zhang. Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2):1–23, 2022.

[61] Tong Li, Chongzhi Gao, Liaoliang Jiang, Witold Pedrycz, and Jian Shen. Publicly verifiable privacy-preserving aggregation and its application in IoT. *Journal of Network and Computer Applications*, 126(October 2018):39–44, 2019.

[62] Mohsen Marjani, Fariza Nasaruddin, Abdullah Gani, Ahmad Karim, Ibrahim Abaker Targio Hashem, Aisha Siddiqa, and Ibrar Yaqoob. Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges. *IEEE access*, 5:5247–5261, 2017.

[63] Eugene Siow, Thanassis Tiropanis, and Wendy Hall. Analytics for the internet of things: A survey. *ACM Computing Surveys*, 51(4), 2018.

[64] Thanga S. Revathi, N. Ramaraj, and S. Chithra. Tracy-Singh Product and Genetic Whale Optimization Algorithm for Retrievable Data Perturbation for Privacy Preserved Data Publishing in Cloud Computing. *Computer Journal*, 63(2):239–253, 2020.

[65] Charith Perera, Yongrui Qin, Julio C Estrella, Stephan Reiff-Marganiec, and Athanasios V Vasilakos. Fog computing for sustainable smart cities: A survey. *ACM Computing Surveys*, 50(3), 2017.

[66] Hossein Ahmadi, Goli Arji, Leila Shahmoradi, Reza Safdari, Mehrbakhsh Nilashi, and Mojtaba Alizadeh. *The application of internet of things in healthcare: a systematic literature review and classification*, volume 18. Springer Berlin Heidelberg, 2019.

[67] P Ravi Kumar, Au Thien Wan, and Wida Susanty Haji Suhaili. Exploring Data Security and Privacy Issues in Internet of Things Based on Five-Layer Architecture. *International journal of communication networks and information security*, 12(1):108–121, 2020.

[68] George Dunea. Privacy concerns. *BMJ*, 329(7464):519, 2004.

[69] Yi Ning Liu, Yan Ping Wang, Xiao Fen Wang, Zhe Xia, and Jing Fang Xu. Privacy-preserving raw data collection without a trusted authority for IoT. *Computer Networks*, 148:340–348, 2019.

[70] Philip James, Ronnie Das, Agata Jalosinska, and Luke Smith. Smart cities and a data-driven response to COVID-19, 2020.

[71] Markus Lanthaler and Christian Gütl. On using JSON-LD to create evolvable RESTful services. WS-REST '12, pages 25–32. ACM, 2012.

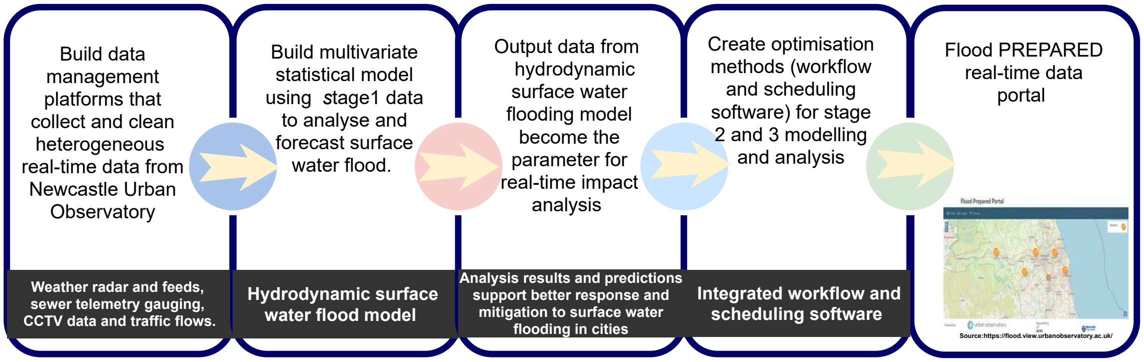
A Supplementary Materials

A.1 Urban Observatory Project (UOP)

The overall framework uniquely applies scientific methods to support decision-making through a multi-scale urban system that observes, analyses, and models both real-time and historical data. For example, air quality monitoring sensors deployed across Newcastle and Gateshead measure key air quality parameters such as Nitrogen Dioxide, Ozone, Carbon Monoxide, and Particulates, generating accurate readings for both authorities and citizens to act upon, thus reducing exposure to air pollution. There are over 50 data types, including many real-time datasets, freely available at the *urbanobservatory.ac.uk* website. These data encompass earth observations, traffic flow, air pollution readings, water quality parameters, and more [33].

1. Newcastle Urban Observatory[28] collects and analyses a vast amount of real-time data from sensors and other sources in urban areas. It uses a wide array of smart devices capturing more than a hundred different metrics per second, in addition to static images, videos, radar, and laser-scan matrices acquired separately. The data generated by these sensors are precise and actionable by both authorities and citizens to mitigate issues such as air pollution and traffic congestion. Nevertheless, managing such massive data volumes presents a significant challenge, necessitating an efficient data management approach. Among the Newcastle Urban Observatory many projects, we examined the Predicting Rainfall Events by Physical Analytics of Real-time Data (Flood-PREPARED) project. This initiative represents a pioneering resource for assessing real-time water surface flood risks and their impacts on cities, equipping them with innovative physical, analytical methods to predict surface water flooding and providing decision-makers with actionable real-time predictions. The project's implementation progressed through five correlated stages, as shown in Figure 6. Another work by James et al. [70] quantifies the impact of COVID-19 measures in the UK. Leveraging existing Internet of

[28] newcastle.urbanobservatory.ac.uk/


Build data management platforms that collect and clean heterogeneous real-time data from Newcastle Urban Observatory
Weather radar and feeds, sewer telemetry gauging, CCTV data and traffic flows.
Build multivariate statistical model using stage1 data to analyse and forecast surface water flood.
Hydrodynamic surface water flood model
Output data from hydrodynamic surface water flooding model become the parameter for real-time impact analysis
Analysis results and predictions support better response and mitigation to surface water flooding in cities
Create optimisation methods (workflow and scheduling software) for stage 2 and 3 modelling and analysis
Integrated workflow and scheduling software
Flood PREPARED real-time data portal
Flood Prepared Portal
urban observatory

Source:https://flood.view.urbanobservatory.ac.uk/

Figure 6: Predicting Rainfall Events by Physical Analytics of REaltime Data (Flood-**PREPARED)**

Things data and a comprehensive analytics infrastructure, the authors developed an interactive COVID-19 dashboard. It visualizes various indicators that update in real-time, comparing data changes against baselines and offering frequent automated comparative descriptive statistics (e.g., daily, weekly updates) to facilitate decision-making. For instance, data from air quality stations, car parks, and traffic sensors analyzed showed a significant decline in pedestrian footfall and traffic volume across Tyne and Wear city during the UK COVID-19 national lockdown in March 2020. Moreover, the Newcastle Urban Observatory archives a collection of historical data for various metrics, serving as a reference for validating the new predictions generated by James et al.'s dashboard. Overall, this dashboard aims to repurpose part of the observatory's real-time data for crisis and disaster management, with analyses replicated in other cities like Sheffield, yielding similar results. Newcastle Observatory may offer insights that could be adapted by observatories in rural locations, including an interactive map of various data and sensors, the ability to download data in multiple formats, and the integration of live Twitter feeds.

2. Sheffield Urban Flows Observatory[29]: Sponsored by the Engineering and Physical Sciences Research Council (EPSRC) and in partnership with UKCRIC Universities, the Sheffield Urban Flows Observatory actively aims to foster a carbon-free, healthy environment. It has developed a dynamic understanding of how the flows of energy and resources impact economic performance and social well-being. The observatory collects, stores, and analyzes city data to monitor the city's environmental performance interactively, engaging citizens and social systems. Its technical platform captures real-time data, including air quality, weather, energy consumption, and both thermal and visual imaging. It consists of various types of sensors (fixed, mobile, and atmospheric), middleware (to gather, integrate, and transform data into meaningful information), data storage, and a data analytics unit.

3. Bristol Urban Flows Observatory [30]: The UKCRIC Bristol Infrastructure Collaboratory aims to transform Bristol into a living laboratory, engaging diverse communities from academia, business, and the citizenry. It uses Open Data, Wireless Sensor Network (WSN), and smart technology solutions to address environmental and social sustainability concerns.

4. Cranfield Urban Observatory[31]: The Cranfield Urban Observatory provides data-centric and remote sensing solutions for environmental, social, and economic issues. It boasts a well-established information technology unit that connects a network of spatially distributed sensors. Its Internet of Things (IoT) network consists of various types of sensors to monitor noise and air pollution, water consumption, and citizens' observations. The observatory extracts data from these sensors and publishes them in real-time, alongside dedicated analytics tools and visualizations, enabling domain experts to monitor the city's environmental performance and make informed decisions to improve life quality, health, and well-being.

5. Birmingham Urban Observatory[32]: With the UK's second-largest population after London, Birmingham's high population density may strain infrastructure, public services, and the environment. Consequently, city administrators invest resources in managing housing, transportation, health, and energy conditions to sustain

[29] urbanflows.ac.uk

[30] bristol.ac.uk/engineering/research/ukcricbristol/collaboratory/

[31] cranfield.ac.uk/facilities/urban-observatory

[32] cityobservatory.birmingham.gov.uk/

adequate living standards, particularly monitoring the environmental, economic, and social factors impacting these critical infrastructures.

6. Manchester Urban Observatory[33]: An interdisciplinary research hub that collects, analyzes, and shares urban data for decision support. The observatory collaborates on various themes with other universities, operating under the dedicated platform "Manchester-I". It offers free and real-time air quality, flood monitoring, and traffic flow information. Linked to Triangulum, a European Union-funded smart city data ecosystem, the Manchester Urban Observatory team has comprehensively rebuilt the platform, integrating data from numerous city-wide sensors. They have also developed a web API that leverages the capabilities of semantic web technology, using JSON-LD [71].

[33] manchester-i.com/home