The high-value datasets as food for the AI-based digital services

Antonio Filograna[1,*,†], Francesca D'Agresti[1,*,†], Rosaria Daniela Scattarella[1,†], Vaiva Venckauskaitė[2,†], Eligijus Bujokas[2,†], Christian Drucks[3,†] and Anne Strunk[3,†]

[1] *Engineering Ingegneria Informatica S.p.A., Piazzale dell'Agricoltura 24, 00144 Rome, Italy*
[2] *ID Vilnius, Lvivo street 25-102, LT-09320 Vilnius, Lithuania*
[3] *Herne.Digital GmbH, Grenzweg 18, 44623 Herne, Germany*

Abstract

Artificial Intelligence is permeating our life more and more. Systems like ChatGPT, CoPilot or Gemini are typical examples of how AI is integrated into people's daily lives. In addition to these "front-end" AI services, easily recognised by humans, there is a set of services that work in the background totally transparent to us. AI needs a large amount of data to work effectively, and the data must be of high quality. The European Commission (EC) is promoting the generation and use of High-Value Datasets (HVDs), a set of data with a wide added value for society, environment and economy. These datasets are completed free of charge, fostering their reuse and promoting innovation. The BeOpen project funded by the European Commission aims at improving these HVDs to boost the development of AI-based Digital Services in the cities. The project offers a technical platform to enhance the quality of HVDs in order to be compliant with the Open Data Directive and the related implementing act issued by the EC. Porto, Herne, Vilnius, Torre Pacheco, Molida de Segura, Cartagena and Naples are the cities involved in the project and the main actors to exploit the improved HVDs to develop AI-based Digital Services, enhancing the quality of life of their citizens.

Keywords

High-value datasets, AI-based Digital Services, Artificial Intelligence, Open data

1. Introduction

Luciano Floridi, one of the most authoritative voices in contemporary philosophy, declared that we are taking advantage (or disadvantage) from the Fourth Revolution [1]. Every revolution has shattered people's certainty of being at the centre of everything. After the first revolution, devised by Copernicus, the Earth (and consequently humankind) was still not at the center of the Universe. The second revolution, devised by Darwin, stated that we are one of the animals in the animal kingdom without having any preferential place in the life chain. The third revolution, devised by Freud, destroyed the famous sentence "cogito, ergo sum" ("I think, therefore I am") putting the emphasis on the unconscious process that influences the human behavior. The fourth revolution, attributed to Alan Turing – the father of the Artificial Intelligence – shed "*new light on who we are and how we are related to the world. [2]*". Information and Communication Technologies (ICT) are redesigning human reality. The human is not the unique "person" to be intelligent. Artificial Intelligence is per-

RPE@CAiSE'25: Research Projects Exhibition at the International Conference on Advanced Information Systems Engineering, June 16-20, 2025, Vienna, Austria

* Corresponding author.
† These authors contributed equally.

antonio.filograna@eng.it (A. Filograna); francesca.dagresti@eng.it (F. D'Agresti); daniela.scattarella@eng.it (Rosaria Daniela Scattarella); vaiva.venckauskaite@idvilnius.lt (V. Venckauskaitė); eligijus.bujokas@idvilnius.lt (E. Bujokas); christian.drucks@herne.digital (C. Drucks); anne.strunk@herne.de (A. Strunk)
0000-0003-2127-293X (A. Filograna); 0000-0002-5385-5761 (F. D'Agresti); 0009-0006-0066-4386 (R.D. Scattarella); 0009-0009-0929-4504 (V. Venckauskaitė) ; 0009-0008-3246-8608 (E. Bujokas)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

vading many aspects of our life. A recent study [3] analyzed AI-based systems that reach the production environment. 87% of them remain in pre-production; this means that they never deployed and tested in the real scenario. An AI-based system must be trained and it needs a big amount of data to learn and adapt its behavior [4]. In this sense, data is one of the most important requirements, mainly the data quantity and the data quality [5]. Having a lot of data allows the system to be trained many times to improve performance. The difference among systems can be seen not only by the effectiveness of the developed algorithm but, at the same time, by the quality of data. Higher quality data means better effectiveness of the algorithm [6].

The BeOpen (*BeOpen an Open framework for boosting EU High Value Datasets from Public Sector*) project, funded under the European Commission under the Digital Europe Programme (DIGITAL-2022-CLOUD-AI-02 call), goes in that direction with its main objective to improve the quality of data. Specifically, BeOpen aims to improve the High-Value Datasets (HVDs), promoted by European Commission, to feed the AI-based Digital Services developed for enhancing the quality of life of citizens.

The project started in January 2023 and ends in June 2025, with a budget of €7.447.000, involving 17 partners (https://beopen-dep.eu/), that are: ENGINEERING INGEGNERIA INFORMATICA S.p.A. (the Project Coordinator), FIWARE Foundation, DATAPOWER SRL, HOP UBIQUITOUS SL LATITUDO 40 SRL, ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS, ETHNIKO ASTEROSKOPEIO ATHINON, Comune di Napoli, STADT HERNE, ASSOCIACAO PORTO DIGITAL, CONSORZIO MEDITECH UBIWHERE LDA, AYUNTAMIENTO DE MOLINA DE SEGURA AYUNTAMIENTO DE CARTAGENA, AYUNTAMIENTO DE TORRE PACHECO, UZDAROJI AKCINE BENDROVE VILNIAUS PLANAS, ARTHUR'S LEGAL BV.

In the last decade, open data generated by the public administrations has been increasing due to several key factors. Firstly, the European Commission promoted the transparency of the public sector, making data (previously inaccessible) available to the citizens. Secondly, the concept of smart cities is becoming a reality, and this is boosting the production of open data. Finally, the private sector also produces data, even if not completely open, which citizens are using in their daily life. Notably, HVDs identified by the EU Open Data Directive [7], represent the most relevant categories (Geospatial, Earth observation and environment, Meteorological, Statistics, Companies and company ownership, Mobility) of datasets that are currently the most requested data and play a key role in the development of digital services. The main objective of BeOpen project is to provide an integrated framework, tailored for public administrations, composed of technical tools, methodologies and guidelines to improve overall open data quality, interoperability and availability of High Value Datasets.

This paper is structured as follows: Section 1 gives an introduction on the important role of having good data for developing AI-based services. Section 2 describes what the High-Value Datasets are, how the EC wants to promote them and describes the BeOpen Framework where HVDs are produced, improved and published. Section 3 illustrates how the HVDs can be exploited to develop new AI-based digital services. Section 4 and 5 showcase respectively the experimentation where the digital services were tested and validated in two cities, Herne and Vilnius, involved in the project. Finally, Section 6 depicts the Conclusion and the possible future works.

2. The HVDs promoted by European Commission and relation with BeOpen project

The concept of High Value Datasets is central to the EU's strategy for leveraging data as a key asset in the digital economy. Certain datasets (such as geospatial, environmental, statistical, and other crucial domain) hold significant potential for creating economic value, driving innovation, and improving public services. The need for HVDs arises from the increasing demand for data-driven decision-making across various sectors. Ensuring that these datasets are open, standardized and interoperable across member states can enhance their usability, promote cross-border applications, and ultimately support the EU's broader goals of digital transformation and sustainable growth, as stated in the implementing act [8] issued by EC in 2022.

The term "High Value" was adopted because they offer substantial benefits when made widely available and easily accessible. It's also an EU policy, since public data is in some degree funded by citizens. The European Commission defined six main domains to have datasets of high value [9]: 1) Statistic, 2) Earth observation and environment, 3) Meteorological, 4) Geospatial, 5) Companies and company ownership, and 6) Mobility. For these domains, six thematic macro characteristics were identified (economic benefits, environmental benefits, social benefits, generation of innovative AI-based services, reuse, and the improvement, strengthening and support of public authorities in carrying out their missions) that can create wide benefits for public use. The HVDs are also marked by legal and technical requirements. For example, they must be free of charge, have a machine-readable format, provide extensive metadata, API and bulk download, publicly available documentation, and open-source license.

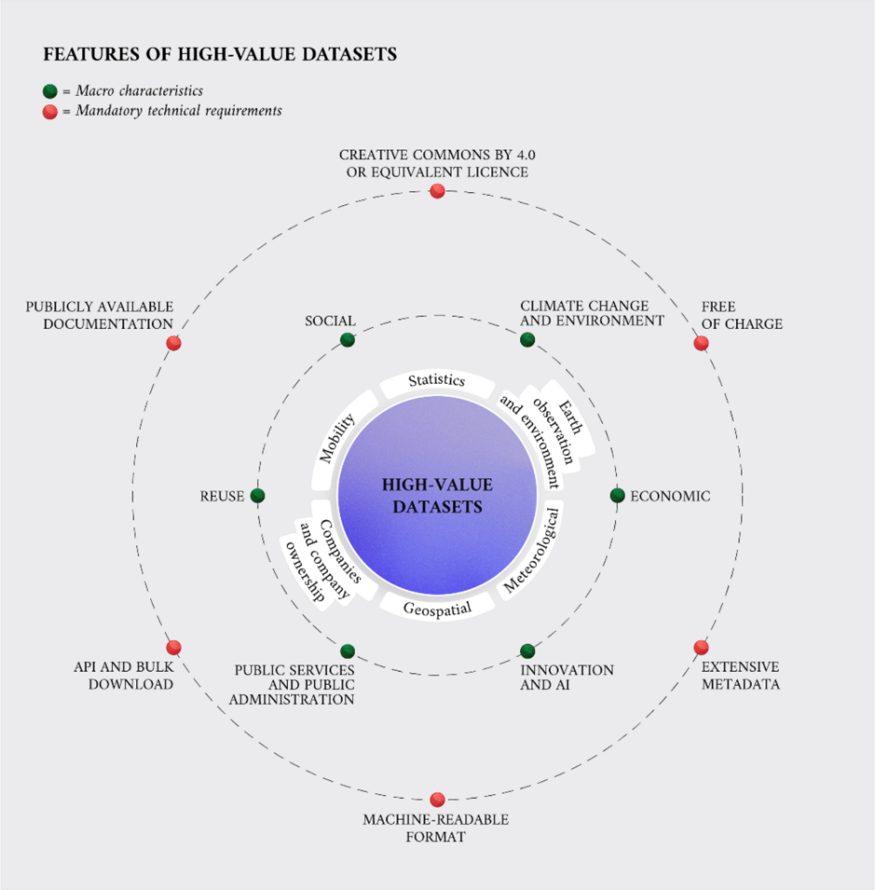
FEATURES OF HIGH-VALUE DATASETS
● = Macro characteristics
● = Mandatory technical requirements
CREATIVE COMMONS BY 4.0 OR EQUIVALENT LICENCE
PUBLICLY AVAILABLE DOCUMENTATION
FREE OF CHARGE
SOCIAL
CLIMATE CHANGE AND ENVIRONMENT
Statistics
Earth observation and environment
Mobility
HIGH-VALUE DATASETS
REUSE
ECONOMIC
Companies and company ownership
Meteorological
Geospatial
API AND BULK DOWNLOAD
PUBLIC SERVICES AND PUBLIC ADMINISTRATION
INNOVATION AND AI
EXTENSIVE METADATA
MACHINE-READABLE FORMAT

Figure 1: Futures of High-Value Datasets [9].

Nevertheless, a specific and standardized approach to support data officers does not exist so far [10]. BeOpen project addresses this gap, developing methodologies, technical framework and guidelines to facilitate the creation and use of the HVDs. These features were tested in 8 use cases in 6 different countries (Greece, Italy, Spain, Portugal, Germany, and Lithuania). Not all data is useful, some are good to have, others are lifesaving, helping decision making. In the BeOpen project, pilots address floods, wildfire, climate change, urban mobility and other natural threats to society. Several domains have been identified, specifically location or geospatial data, also combined with mobility, traffic, health related, and data coming from other sensors, such as lightning. We need to ensure that data is accessible and open, using standard technologies and methods to promote data interoperability.

The current challenges in the public sector are opportunities for the BeOpen project to act as a catalyst for Open Data in public administration, improving data sharing, interoperability, availability

& usability of datasets in different sectors. BeOpen offers a technical framework for data interoperability which consists of a set of technical tools, that is adopted by the different Public Administration involved in the BeOpen project in order to manage and improve their High Value Datasets. The final aim is to publish them via standard and interoperable formats and interfaces. Among the main technical tools of the framework there is the *Data Model Mapper,* which allows to obtain interoperable datasets, since it enables the harmonization of their structure using common data structures and schemas. Then, data thus structured are stored in the *FIWARE Orion Context Broker*, a crucial tool that promotes interoperability between datasets. The web-based environment that is the final access point for users, facilitating easy access to all tools and services, is the *BeOpen dashboard;* it allows for a continuous transfer and monitoring of the data as well as an easy way to exchange data with application and digital services and facilitates the publication and reuse of HVDs in line with the EU Implementing regulation on High-Value-Datasets. The framework covers the complete pipeline from data collection, quality improvement, semantic enrichment, to data provisioning; it promotes the publication of a catalogue of HVDs, publicly available and accessible for the creation of new services and AI applications, a set of guidelines, recommendations and legal interoperability aspects of data. All these results allow public authorities to create new digital services leveraging the improved HVDs.

Interoperability must be improved to encourage reuse of HVDs. To achieve it, it is needed to be compliant with international standards or to adopt common data structures. Ensuring interoperable HVDs means adhering to common schemes and standards known at European level. To address the need for data homogenization, many initiatives and organizations have been actively working towards the development of data models, APIs and (meta)data standards. The directive adopted as a schema for many of the HVDs produced in BeOpen is *INSPIRE* [15] which seeks to establish a European spatial data infrastructure by harmonizing and standardizing geospatial information for the purposes of EU environmental policies. Inspire addresses 34 spatial data themes needed for environmental applications and provides a common framework for interoperability, facilitating data sharing and integration across national borders.

3. How the HVDs can be the food for AI-based digital services

Excellence and trust are the two key pillars on which the European approach to AI is based. Regarding excellence, the EU wants to enable the development and adoption of AI within the European Union by maximizing and promoting resources, including the adoption of HVDs and coordinating investments. The EU will ensure this by issuing regulations that provide the right infrastructure for building robust and high-performance AI systems. To build trustworthy AI, the European Commission has launched the AI Act, with rules on general-purpose AI coming into force in August 2025. The AI Act sets out a clear set of risk-based rules for AI developers and deployers regarding specific uses of AI. Among the high risks, AI systems are subject to strict obligations before they can be placed on the market, including the importance of providing and ensuring high-quality datasets that feed into AI systems to minimize risks of discriminatory and wrong outcomes.

HVDs are essential for the operation of AI-based decision-making systems (AI-based Digital Services). AI-based Digital Services (DS) is a set of technologies based mainly on machine learning and deep learning techniques, used for data analysis. To generate correct and sure predictions, HVDs provide complete, accessible and high-quality information. The quality of HVD data avoids errors and biases in AI models and ensures that decisions taken are based on solid and precise information, establishing a robust foundation basis for developing intelligent systems in various sectors. HVD is an essential element on which the AI decision-making system is based; its use is a notable advancement toward a future in which decisions are increasingly accurate, transparent and reliable. The quality of the data provided improves the reliability of the results, both in terms of technological innovation and economic growth.

Characteristics that make HVDs suitable as an entry for AI-based DS:

- **Reliability**: data is the cornerstone of any AI application, as it defines both the quality and effectiveness of the model. For this reason, it is clear that adopting a high-quality dataset is crucial to train precise models, guaranteeing reliable results and making correct decisions, which is also based on clear data governance, such as that provided by the EU through the European Data Governance Act. It is a key pillar of the European strategy which seeks to increase trust in data sharing, strengthen mechanisms to increase data availability and overcome technical obstacles to the reuse of data [11].
- **Structure**: HVD is well structured as it follows precise standards and common data models. The HVDs produced within the BeOpen project, for example, are datasets published in an open data portal according to the DCAT-AP standard [12], which stands for DCAT Application profile for data portals in Europe for describing public sector datasets, to enable cross-data portal search for data sets and make public sector data better searchable across borders and sectors. This ensures that the training process and access to information are carried out without incomplete, disorganized and poorly structured data.
- **Availability**: high-value datasets are made available under optimal conditions built on the principles of findability, accessibility, interoperability and reusability (FAIR principles). In particular, an HVD is public and easily accessible. It can be provided via standard and application programming interfaces (APIs) and via bulk download. Furthermore, the dataset is available in a machine-readable format, suitable and easily understandable for AI-based DS. In addition, such datasets support meeting the general reuse requirement, since they are made available through Creative Commons BY 4.0 license or any equivalent or less restrictive open license.

Using HVDs in the context of AI will bring value and benefit to both the economy and society, for example, by helping to reduce urban air pollution, optimizing urban transport networks and finally, monitoring the environment by predicting extreme weather events with the development of disaster mitigation strategies.

4. The real case of Herne

The city of Herne is a major city with 150,000 inhabitants. It is located in the center of the Ruhr region in the German state of North Rhine-Westphalia. The Ruhr region has 5.1 million inhabitants,it is a major transportation hub, and a key economic center with companies in the fields of industry, education, and research.

The digital service, which uses high-value datasets as part of the BeOpen project, focuses on improving public safety at the Cranger Kirmes. The Cranger Kirmes takes place every August in Herne. It is one of the largest folk festivals in Germany, attracting over 4 million visitors over a 10-day period each year. The Cranger Kirmes has a long tradition, is an important part of local culture, and is a significant economic factor.

The fair takes place in the heart of the city on public streets and squares. Since there are no designated entrances and exits in this public area, the flow of visitors cannot be controlled. Therefore, the fairgrounds are always very crowded.

Another problem is the many visitors arriving by car. There is only one main access road to the fairgrounds, and parking is relatively limited.

Two major dangers can arise from this situation:

- Large traffic jams in the city area, which extend to the motorways
- A mass panic on the fairground due to overcrowding

The goal is to enable responsible regulatory authorities to identify dangerous situations more quickly than before so they can take timely countermeasures. A digital service has been developed for this purpose that provides real-time data on traffic and crowd density captured by sensors. A forecast for the future development of the measured values is also provided.

A convolutional neural network model is used for prediction. To ensure the best possible results, the data sets of the individual features used to train the model must be of the highest possible quality. The BeOpen framework was used to normalize this training data and guarantee the required quality. High-value datasets from past events on crowd density, traffic, weather, and social media data were generated for training.

In productive operation, a new forecast is calculated and provided hourly based on the trained model. The events in 2022 and 2023 served to generate historical data. The first productive and successful use took place at the 2024 event.

One challenge in implementing the case studies was the technical transmission of the data. The reception quality of the LPWAN networks was severely compromised by the movement and electronics of the rides. Due to the high number of visitors, the mobile network is also regularly overloaded. A dedicated Wi-Fi network with directional antennas provided a solution. Another challenge was the use of camera-based systems in public areas in compliance with data protection regulations.

Validation of the sensor data resulted in an accuracy of over 90%. This is particularly positive when measuring crowd density, as people obscure each other in busy areas and visibility conditions for the cameras are poor in the evening.

5. The real case of Vilnius

The European green capital of 2025 along with other cities is fighting a phototoxic invasive plant, Sosnowski's hogweed [13], which not only encroaches on native species, but also poses a health risk with the capability of skin burn. Vilnius city municipality updates sightings information, and often district elders go through known areas to update information on these sites, helping with monitoring. Representatives of other municipalities have stated that the biggest challenges are related to human resources and coverage. There are hard to reach places, not enough manpower to scan large areas, leaving patches of plants unnoticed and proceeding to spread. To tackle these problems, we decided to use drones and machine learning.

The main fuel for contemporary AI and machine learning models is data. In our case, when building the “Sosnowski identifier”, we used labeled drone images, where human experts drew polygons around the identifiable Sosnowski hogweeds. Even if the coverage at once is not as big as it would be from a satellite or a fixed wing UAV, using drones still aids in covering remote areas, saving on-foot inspection time. Also, the design decision of using RGB was motivated by the fact that the solution can be more largely adoptable for other entities, meaning no need to buy expensive multispectral cameras by investing in technology that is still suitable for many other cases. Although we need to emphasize that spotting Sosnowski plants in an RGB image is difficult, since the weed is green and usually very hard to spot in grasslands or lawns. Data collection also heavily relies on the plant’s phenology and coordination with eradication work. If the collection process is late, there is a possibility of missing out on new training data for model improvement. After the labelling process, we have chosen to fit a YOLO [14] model (YOLO11 iteration fits perfectly with our need for faster time to market) on the photos and labels, following inference.

During the first version of the model, we only had 163 training images for a season, leaving us with around 0.4 precision and recall. Plants that are in less urbanized areas are mixed in with lifeless vegetation as well. This has caused the model to not only pick up tree branches in low confidence but also identify browned young hogweeds that have already gone through one of many extermination measures.The creation of labels has taken the most amount of time in the project – more than 80%. As we approach a new season for the plant, we will be able to create new improved versions of our identifier. Model predictions are collected and transformed into two separate geospatial polygonal datasets. One for identified invasive species and the other showing the shifted spread through an analysis using high value meteorological wind data. Most of the attributes are filled in automatically (id, probability, name, detection date etc.) while some are left to be edited by the people involved in their management and eradication. This new dataset would come in handy analyzing how effective the methods used are with lifespan date fields and concrete eradication method records.

Multiple existing datasets have been selected to become HVDs for further improvement of the digital service. Datasets regarding proximity to green spaces, intensely used green spaces and natural framework are guides to assess susceptibility and priority areas when allocating resources to fight this invasive plant. Furthermore, the datasets can serve as input features in analyzing what areas to cover, checking more of the city to identify overlooked patches.

The datasets will have been made available in the city's maps portal, where citizens and stakeholders could interact with the improved HVDs as well as other available datasets. The platform is made so that everyone can log in and save their own map themes. Users can add any of the available datasets to the map and save it. This way, you can easily access not only thematic maps already prepared but also customize them to your own interests. The city's maps portal integrates outputs from various sources, is easy to use and serves as a great visual tool for viewing, sharing and interpreting data for all. As for AI, ID Vilnius is developing a GitHub repository to increase reproducibility.

6. Conclusions

BeOpen project, following the Open Data Directive, increased the availability, quality and usability of HVDs. This encourages the easy development of new services & AI applications to improve the quality of citizens' lives. One of the benefits of exploiting HVDs and creating new digital applications is the reduction of the market fragmentation for digital services, leveraging on standard and trusted technologies coming from the EU project results. The BeOpen Framework can be easily replicated in all the cities that want to improve their HVDs and create digital services by exploiting them. The solution is open source and it can be integrated in the open data portal of the public administration.

In the future, one of the main challenges is to convince the public sector to invest in the openness and quality of data, since data is the fuel of the actual economy. Furthermore, AI without data and humans is not intelligent, cities cannot meet the definition of a SmartCity without quality data and where there is an absence of data, there can be no Big Data. Who produces data? Humans. That means we, as humans, are still at the center of revolution.

Acknowledgements

This project has received funding from the European Union's Digital Europe Programme under grant agreement No 101100807.

Declaration on generative AI

One author used GPT-4o to select and adapt the appropriate citation format.

References

[1] L. Floridi, Children of the fourth revolution. *Philosophy and Technology* 24 (3):227-232. (2011)

[2] Michael J. Paulus, Jr., 2018. URL: https://michaelpaulus.org/libraryfutures/the-four-information-revolutions/

[3] J. Weiner, Why ai/data science projects fail: how to avoid project pitfalls. Synth Lect Comput Anal 1(1):77, (2020)

[4] Ue. Habiba, M. Haug, J. Bogner, et al., How mature is requirements engineering for AI-based systems? A systematic mapping study on practices, challenges, and future research directions. Requirements Eng 29, 567–600 (2024). https://doi.org/10.1007/s00766-024-00432-3

[1] A. Vogelsang, M. Borg, Requirements Engineering for Machine Learning: Perspectives from Data Scientists. 2019 IEEE 27th International Requirements Engineering Conference Workshops (REW). doi:10.1109/rew.2019.00050

[2] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, The ML test score: A rubric for ML production readiness and technical debt reduction, in Proceedings of IEEE Big Data, 2017.

[3] European Commission, Open Data Directive, 2019. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433&uri=CELEX:32019L1024

[4] European Commission, Implementing regulation (EU) of 21.12.2022022, URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=PI_COM:C(2022)9562

[5] European Commission, High-value datasets – an overview through visualisation, 2022. URL: https://data.europa.eu/en/publications/datastories/high-value-datasets-overview-through-visualisation

[6] A. Nikiforova, N. Rizun, M. Ciesielska, C. Alexopoulos, A. Miletić, Towards High-Value Datasets Determination for Data-Driven Development: A Systematic Literature Review. In: Lindgren, I., et al. Electronic Government. EGOV 2023. Lecture Notes in Computer Science, vol 14130. Springer, Cham. https://doi.org/10.1007/978-3-031-41138-0_14

[7] European Commission, 2024. European Data Governance Act. URL: https://digital-strategy.ec.europa.eu/en/policies/data-governance-act

[8] European Commission, 2024. DCAT Application profile for data portals in Europe (DCAT-AP). URL: https://interoperable-europe.ec.europa.eu/collection/semic-support-centre/solution/dcat-application-profile-data-portals-europe

[9] GBIF Secretariat, GBIF Backbone Taxonomy, 2023. URL: https://doi.org/10.15468/39omei

[10] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 779–788. doi:10.1109/CVPR.2016.91.

[11] Joint Research Centre, INSPIRE Infrastructure for Spatial Information in Europe, 2025. URL: https://knowledge-base.inspire.ec.europa.eu/index_en