Towards an Italian Energy Data Space

Chiara **Rucco**[1], Antonella **Longo**[1] and Marco **Zappatore**[1]

[1]*Dept. of Innovation Engineering, University of Salento, via Monteroni sn, 73100 – Lecce (Italy)*

Abstract

The efficient use and the sustainable production of energy are some of the main challenges to face the ever increasing request for energy and the need to limit the damages to the Earth. Smart energy grids, pervasive computing and communication technologies have enabled the stakeholders in the energy industry to collect large amounts of useful and highly granular energy data. They are generated in large volumes and in a variety of different formats, depending on their originating systems and prospected purposes. Moreover, the data type can be structured and unstructured, in open or proprietary formats. This work focuses on harnessing the power of Big Data Management to propose a first model of an Italian Energy Data Lake: the goal is to create a repository of national energy data that respects the FAIRness' key principles [1], aimed at providing a decision support system and the availability of FAIR data for open science. Starting from data of two thematic areas that are part of the nine common European Data Spaces identified in the European Data Strategy[2], namely the Green Deal data space and the Energy data space, an open and extensible platform to enable secure, resilient acquisition and sharing of information will be presented, for enabling the Green Deal priority actions on issues such as climate change, circular economy, pollution, biodiversity, and deforestation.

Keywords

Energy, Datalake, Open Data, Fairness

1. Introduction

Global energy requirements are continuously increasing. Conventional methods of producing more energy to meet this growth pose a great threat to the environment. CO_2 emissions and other bi-products of energy production have direct consequences on everyday life. Therefore, we need to understand and improve the energy efficiency at both producer and consumer sides. ICT-enabled smart energy grids and sensors are being installed globally to measure energy consumption and limit the environmental impact: these smart objects produce large volumes of data, generated by different devices and in different formats, so that they embody the concept of Gartner's 'Big Data 3Vs' [3] - volume, velocity and variety. For the purpose of knowledge discovery, this data needs to be collected and analyzed, and the extracted insights from the analysis need to be visualized for easy and effective understanding. To face these challenges, a highly scalable and flexible data analysis platform for automating the whole process is required. A first model that can meet these requirements is an architecture that draws elements from classic data warehouse systems on the one hand and from pure data lake systems on the other hand. This model, defined as Data Lakehouse [4], together with other paradigms like polystore

databases, can be implemented and tested with real life data from smart energy devices in order to contribute to the realization of a society that follows the innovative Circular Economy paradigm, a system where resource input and waste, emission, and energy leakage are minimized by slowing, closing, and narrowing material and energy loops [5]. The new schema will contain heterogeneous data sources and will be processed in order to be compliant to FAIR (Findable, Interoperable, Accessible, Reusable) principles: a well-documented and highly re-usable data set enables the ultimate aim to trusted, effective and sustained reuse of research resources [6]. This paper aims at identifying featuring architectural aspects and modelling challenges for an Energy Data Space to be adopted nationwide in Italy. The work is structured as in the following: the second section presents an overview of the background and the state of art, with reference to the rise of the new 'Energy of Things' characterised by heterogeneous large data sets, and to three cutting-edge projects on this topic. The main goal of the this architectural proposal is described in the third section. The last part is devoted to discussing the impact and results of the solution's development.

2. Background and state of art

Nowadays, the concept of "Data Lake" is popular for accumulating data from heterogeneous sources. Data lakes are used for storing large scale raw data as a single big data repository, providing ingestion, exploration, and monitoring functionality [7]. Data lakes, in contrast to data warehouses, are databases containing data from different sources in structured, unstructured and semi-structured formats, along with capabilities of handling batch and real-time streams. Moreover, data lakes exhibit different implementation forms (e.g., on premises, cloud or multi-cloud, and hybrid) [8]. Currently, data lakes have been exploited in several application domains, ranging from digital humanities [9] to power grid management [10]. In order to get a full insight on the scenario analyzed, a general overview has been built: at the beginning, the Circular Energy paradigm is discussed, with a view on its heterogeneous energy-related data and on data key principles for achieving *FAIRness* [1]. Then, as a starting point for this research, some existent projects for the creation of a National Energy data repository are explored: the first is an initiative for creating a digital twin of Earth, the second is a Danish proposal for renewable energies, the third is a novel US initiative to make energy data usable and discoverable by researchers.

2.1. Energy of Things

According to the United Nations Sustainable Development Goals agenda [11], energy efficiency is one of the key factors for sustainable development: "ensuring access to affordable, reliable, sustainable and modern energy for all by 2030 will open a new world of opportunities for billions of people through new economic opportunities and jobs"[1]. Furthermore, energy efficiency brings long-term economic benefits by reducing the cost of fuel import/supply, energy production and energy sector emissions. Effective analysis of real-time data in the energy supply chain plays a key role in improving energy efficiency and more optimal energy management [12]. Modern

---

[1]https://sdgs.un.org/topics/energy

technologies, such as the Internet of Things (IoT), offer a wide range of applications in the energy sector, i.e. in the areas of energy supply, transmission, distribution and demand. With the new surging of portable smart devices, consistently equipped with sensors, supported by more and more performing cloud computing solutions, and densely used for mobile social networking, *human as sensor* has become a promising sensing paradigm. For this reason, the term ENERNET (Energy of Things) was recently introduced by Steve Collier in an IEEE webinar on the future of energy: it is defined as a convergence and a marriage between Smart Grids and IoT [13]. Emerging ENERNET opens up other possibilities in order to have an affordable, reliable, secure and sustainable supply of electrical power and energy. The novel sensing technologies promote the data source into a new information space paradigm, which seamlessly integrates cyber-space (CS), physical space (PS) and social space (SS), namely Cyber-Physical-Social Systems (CPSS) [14]. CPSS has a crucial role in improving energy efficiency, increasing the share of renewable energy and reducing the environmental impact of energy use; this can be compliant to a context of Circular Economy, a concept that has been framed by the Ellen MacArthur Foundation as *an industrial economy that is restorative or regenerative by intention and design*[5]: it represents a system in which resource input and waste, emission, and energy leakage are minimized by slowing, closing, and narrowing material and energy loops. This can be achieved through long-lasting design, maintenance, repair, reuse, re-manufacturing, refurbishing, and recycling. The amount of available data for energy analysis is growing rapidly due to a large number of data sources, such as smart cities installing sensors, IoT and personal devices capturing regular behaviour, human curated datasets (e.g. Open Maps), large-scale collaborative data-driven research, satellite imagery, multi-agent computer systems, and open government initiatives. This abundance of data is diverse both in format (e.g. structured, images, graph-based, matrix, time-series, geo-spatial, and textual) and in types of analysis performed (e.g. linear algebra, classification, graph algorithms, and relational algebra). To help different types of data and analysis activities, scientists and analysts often rely on ad-hoc procedures to integrate various data sources. This typically means manually curating how to clean, convert and integrate data. Such approaches are delicate and time consuming. In addition, to perform the analysis, they require bringing both data and computation into a single architecture, which is typically a (distributed) system not suitable for all necessary computation. Most analysts and programmers, however, are not well prepared to handle a multitude of systems, handle transitions between systems robustly, or define the correct framework for the assignment.

2.2. Open initiatives in energy computing field

2.2.1. DestinE Data Lake

As part of the European Commission's Green Deal and Digital Strategy, Destination Earth (DestinE)[15] is a project focused on contribution to achieving the goals of the double transition, green and digital. DestinE is designed to unlock the potential of digital Earth system modelling. It will focus on the impacts of climate change, aquatic and marine ecosystems, polar regions, the cryosphere, biodiversity or extreme weather events, as well as possible adaptation and mitigation strategies. It will help predict major environmental disasters and environmental degradation with unprecedented accuracy and reliability. The heart of Destination Earth will

be a unified cloud-based modelling and simulation platform that will provide access to data, advanced computing infrastructure, software, artificial intelligence applications and analytics. As seen in figure 1, the project will integrate digital twins (DTs) - digital replicas of different aspects of the Earth system, such as weather and climate change projections, food and water security, global ocean circulation and ocean biogeochemistry, among others - and provide users with access to thematic information, services, models, scenarios, simulations, forecasts and visualisations. The platform will also allow the development of applications and the integration of user data.
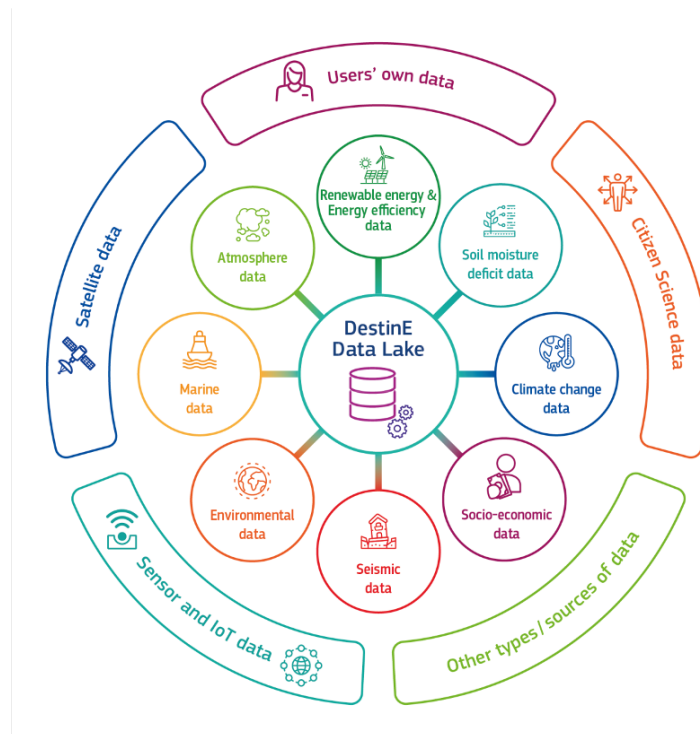

Figure 1: DestinE Data Lake as proposed in [15]

The project, which is currently only submitted as a proposal to the European Commission in line with the European Data Strategy, will be implemented gradually over the next 7-10 years starting in 2021. The basic operational platform, digital twins and services will be developed as part of the Commission's digital programme, while Horizon Europe will provide research and innovation opportunities that will support the further development of DestinE.

2.2.2. Flexible Energy Denmark

Flexible Energy Denmark (FED)[16] is a digitisation project that aims to make Danish electricity consumption flexible, so that it becomes possible to use excess electricity from wind turbines and solar cells. The project brings together leading researchers, organisations, utilities, software companies and numerous living laboratories in the country that provide real data for the project.

Specifically, FED collects data from a series of Living Labs (LLs) in physical environments representative of real life. Raw data on electricity, water and district heating consumption of many thousands of households, as well as indoor climate data of two primary schools and 155 households in Aalborg end up 1-4 times a day in a Data lake, called FED Data Lake (FEDDL), which is operated by the independent, non-profit national research centre *Center Denmark in Fredericia*[2] and enables efficient and advanced analysis. The FED ecosystem includes:

- A data ecosystem (the Datalake containing a variety of energy-related data that are mainly collected from the living labs in the project, but also from other sources such as BBR (the Danish Registry of Buildings and Houses) and DMI (Danish Meteorological Institute)
- An ecosystem for digital tools (tools based on artificial intelligence, are enabled by Big Data from the data ecosystem)
- An ecosystem for digitisation solutions combining some of the tools developed, with the aim of managing energy flexibility in Denmark.

FEDDL is built using only open source tools that can be run either on-premise or in cloud environments.

2.2.3. Open Energy Data Initiative

The Open Energy Data Initiative (OEDI)[17] is a centralized repository of valuable energy research datasets collected by U.S. Department of Energy programs, offices, and national laboratories. Designed to enable data discovery, OEDI facilitates access to a wide network of results, including data available in technology-specific catalogs such as the Geothermal Data Repository and the Marine Hydrokinetic Data Repository. The initiative aims to improve and automate access to high-value energy datasets across U.S. Department of Energy (DOE) programs, offices, and national laboratories. This platform is being deployed by the National Renewable Energy Laboratory (NREL) to make data usable and discoverable by researchers and industry to accelerate analysis and innovation development. Not only does the data lake provide tools to create actionable insights for analysts and to provide high-value open data, but it can also be used to conduct interesting data mashups or calculations to develop new and expanded data sets.

OEDI leverages on Amazon Web Services (AWS) to enable analytics capabilities, innovative dataset access and to trigger new relationships among cloud partners. The data lake is based on consolidated AWS storage solutions for datasets (i.e., AWS S3 buckets) with elastic load balancing, and AWS cloud-optimized analytics tools (e.g., AWS Glue, AWS Athena) that to help users consolidate data into non-standard formats, speed up analytics, and allow users to pull or move small parts of analytics into their AWS accounts. [3]
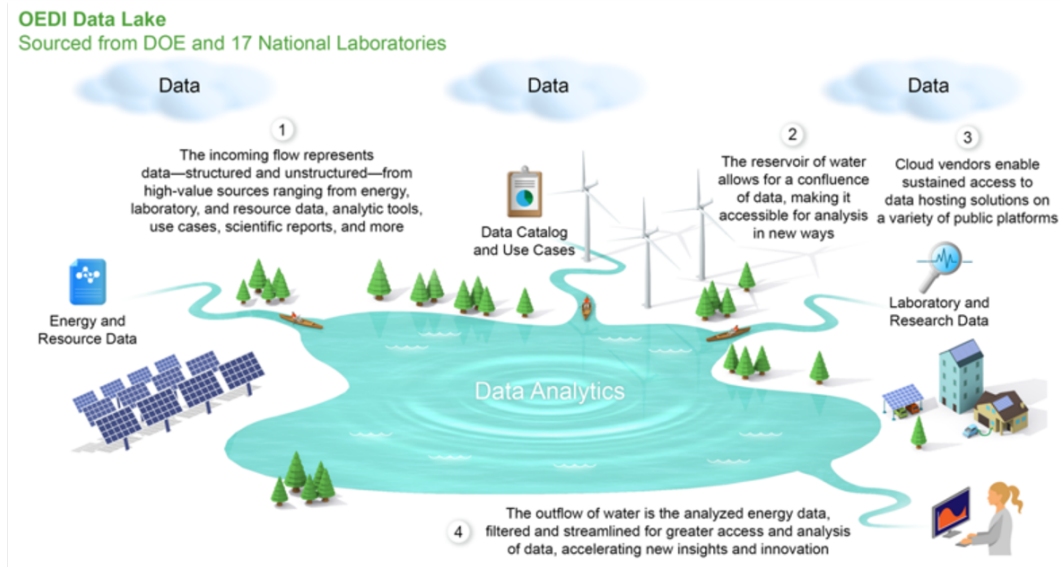
---

[2]https://www.centerdenmark.com/
[3]https://openei.org/wiki/

Figure 2: Open Energy Data Initiative as depicted in [17]

3. Design of the Italian Energy Data Space

3.1. Logical architecture

In this context, the aim of the work is to design a resource for the Internet of Energy, capable of collecting energy data from Italian agencies, consortia and research centres, in order to develop a "Google of Energy", a system capable of indexing and searching energy Big Data. It can be used to facilitate future studies in the energy sector and all reliable infrastructures. Developing and consolidating a new approach to energy management, throughout the analysis of data from institutional databases, sensors, IoT devices, Industry 4.0 infrastructures, in the field of energy and its eco-system, will help to use the Big Data potential to support the Green Deal's priority actions on issues such as climate change, circular economy, pollution, biodiversity and deforestation.

Based on the model developed in the Danish National Energy Data Lake[16], where a national repository for energy data is created, figure 3 gives an abstract overview of the proposed Data Lake logical architecture: it is composed of five separated layers, i.e., Data Sources, Data Collection/Ingestion, Data Storage, Data Exploration, and Data Consumers, and four cross-cutting layers, i.e., Privacy and Data Protection layer, Access Management, Meta Data Governance and Resources Management. Layers Data Sources and Data Consumers represent systems which are external to the Data LakeHouse structure.

Data Sources: Data sources considered for this purpose are Mobile Sensors and IoT sensors capable of collecting energy and environmental data, Open Data made available by Public Administrations or research centres and Living Labs. Some of the sources for collecting source data are as follows:
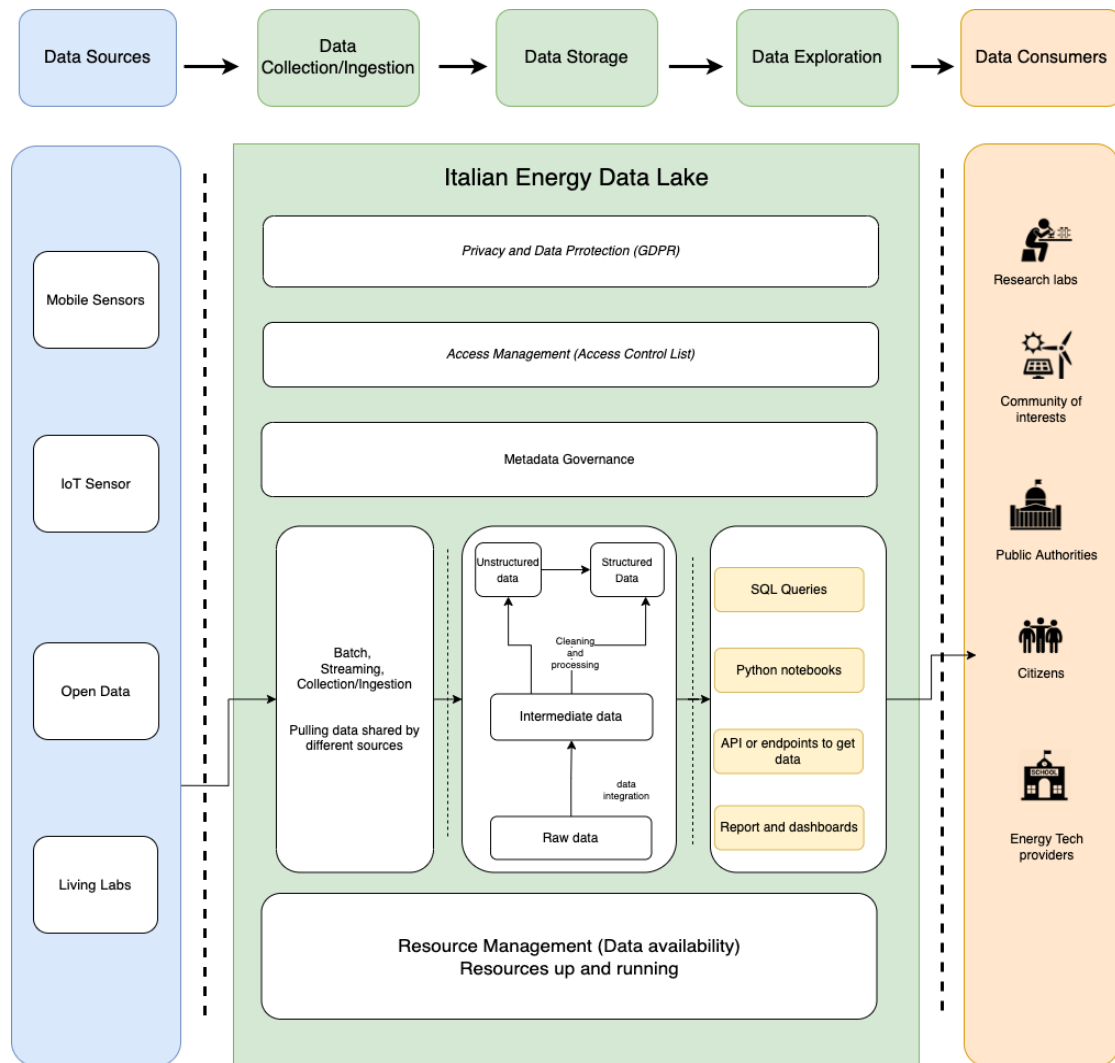
Figure 3: Logical view of the proposed architecture

- Open data: Energy Production and consumption
  - GSE: it provides data at national and regional data about renewable sources, transportation, energy counts;
  - ARERA: monitoring of novel generation plants at national level, Data about market, clients, production, consumption;
  - ISTAT: energy production from renewable sources at national level and consumption from families;
  - Terna: national data related to production, generation plants, international benchmarks, peaks, consumption;
  - Eurostat

* Energy statistics section: share of renewable energies, energy productivity, energy supply bu product, energy consumption by product;
* Sustainable development: primary energy consumption, population lacking energy due to poverty;
- Industry 4.0: smart meters, data coming from power generation plants.

Data Collection/Ingestion: The custom data collection enables data retrieval from data sources requiring custom scripts, e.g. if the data is embedded into HTML pages, APIs, or when files containing data are provided manually in CSV, TSV and PDF formats. Data ingestion is the transportation of data from assorted sources to a storage medium where it can be accessed, used, and analyzed by an organization. There are different ways of ingesting data, and the design of a particular data ingestion layer can be based on various models or architectures. The two considered kind of data ingestion are batch processing and streaming processing. In the first one, the ingestion layer periodically collects and groups source data and sends it to the destination system. Groups may be processed based on any logical ordering, the activation of certain conditions, or a simple schedule. Real-time processing (also called stream processing or streaming) involves no grouping at all. Data is sourced, manipulated, and loaded as soon as it's created or recognized by the data ingestion layer.

Data Storage: The data lake storage problem asks for selecting appropriate data stores to preserve ingested datasets. There are many solutions in the literature and they apply various relational and NoSQL databases [18], and present different manners of data storage organization. There are solutions considering heterogeneous data sources while others target at a particular type, e.g., relational tables. In order to host different types of data, the solutions could be a universal format or allowing multiple formats. Some approaches rely on the common relational or NoSQL stores while others have developed new storage systems. The data storage systems could be on-premise or cloud-based[8].

This layer can be divided in three different zones:

- *Raw data zone*: all types of data are ingested without processing and stored in their native format. This zone allows users to find the original version of data for their analytics to facilitate subsequent treatments. The stored raw data format can be different from the source format.
- *Intermediate zone*: after ingestion, the data lake is a vast collection of raw datasets with certain metadata. To make the data usable for querying, a number of solutions are proposed for further processing of the raw data, e.g., find more metadata, discover hidden relationships, and perform data integration, transformation or cleaning if necessary.
- *Structured and unstructured zones*: they stores all the available data for data analytics and provides the access of data. This zone allows self-service data consumption for different analytics (reporting, statistical analysis, business intelligence analysis, machine learning algorithms) according to their format.

Data Exploration: The top layer focuses at the interaction of users with the DataLake. It is important that useful information can be retrieved out of data lakes. However, this is

challenging due to a large number of ingested sources, and the heterogeneity of data. Given data lake systems with a large number of datasets, users may have knowledge for one or a few data sources, but rarely all the datasets. The query formulation component should support users in creating formal queries that express his information requirement. The data interaction should cover all the functionalities which are required to work with the data, including visualization, annotation, selection and filtering of data, and basic analytical methods. Users can first browse the existing data sources, including their description, statistics, and schema; then she can write a query (SQL or JSONiq [4]) for a single dataset, or use the user interface to make a keyword search over the schema or the data. Alternatively, with certain knowledge of the datasets, which could be learned through the previous exploration processes, they can choose to integrate a subset of relevant datasets, and query them using formal queries or keyword search [19].

An important kind of output that the architecture could provide, is the Fair Data API: according to the FAIR data principles, research outputs are shared in a way that enables and enhances reuse by humans and machines. The characteristics of these resources can be oriented to achieve compliance with FAIR guidelines. For example, output generated uses globally unique identifiers and can assign other identifiers. The data elements described in FAIR correspond to concepts and (meta)data objects modeled, as our DataLake resources and described with rich metadata and context information. In the output of the Data Lake, resources are retrievable via open APIs, that is, absolute URIs and standard Representational State Transfer (REST) protocols.

Data Consumers: Human users play an essential role in the management of data lakes. The users of a data lake are also data providers; the insight provided by the human helps the data in the data lake to mature over time. Data consumers range from communities of interest (e.g., citizens groups and associations interested in performing pollution measurement, and factories and industries interested in their level of environmental pollution) to public authorities, and from citizens (both single individuals and associations) to other end-user categories such as schools or research-labs.

3.2. The proposed development process

The steps for the development of the platform are based on the following phases.

1. The first part will focus on an in-depth study of the state of the art and analysis of existing architectures proposed in the second section. A first phase of heterogeneous data collection from the various energy sources will be carried out.
2. The second phase, aimed at scenario definition, will focus on interviews with SMEs and stakeholders that can help in the design of a use case to prototype the research results. This is aimed at an understanding to elicit the needs, the current state of the art in energy generation, distribution and use.
3. Design of pilot projects and use case, also creating living labs for involving prosumers and providers

---

[4]https://www.jsoniq.org/

4. Development of the digital platform for collecting data and providing data services and tools

5. Incremental extension of use cases and further involvement of new providers, consumers, stakeholders

4. Conclusion

Knowledge of consumers' energy consumption and indoor climate is worth its weight in gold to utilities, industry and researchers. They use the data to plan production and develop services and algorithms that control energy consumption so that it becomes more flexible and renewable energy is not wasted. Our ambition with the national platform is that it can form the basis for the release of data from electricity, water, heat and potentially also gas, so that the data can be used by commercial suppliers to develop new business models that support data-driven models for the green transition. Creating a repository of national energy data that respects the Fairness' key principles, is the starting point to provide an open and extensible platform to enable secure, resilient acquisition and sharing of information with the aim to improve the well-being and inclusion of citizens, produce a more effective response to pollution or other environmental emergencies, and make Smart Cities and extended urban areas feel more secure and safe to the citizens living in them. Further, endeavors from citizens and joint academic-community science can assist with distinguishing environmental health problems related with air quality in metropolitan regions. Unfortunately, there remains a gap between the development and the effective utilization of these cutting-edge technologies within communities of proactive decision-making [20]. The importance of this topic will help to raise public awareness of energy problems, to highlight the importance of citizens' engagement and to inspire citizens to adopt sustainable consumption habits and behavior patterns. These habits will promote new sustainable services, e.g. lengthening product life cycles through reuse, repair and refurbishment and encourage waste reduction, energy savings and circular thinking: the so-called 'citizen science' is emerging.

5. Acknowledgment

References

[1] M. D. W. et al., The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016). doi:`10.1038/sdata.2016.18`.

[2] Towards a common European data space, Technical Report, European Commission, COM 232, 2018.

[3] D. Laney, 3D data management: Controlling data volume, velocity and variety, META Group Research Note 6, 2001.

[4] P. G. Alonso, SETA, a suite-independent agile analytical framework, Master's thesis, Universitat Politecnica de Catalunya, 2016.

[5] E. M. Foundation, Towards the Circular Economy, Technical Report, EMF, McKinsey Company, 2013.

[6] F. H. Mardiansjah, Extended urbanization in smaller-sized cities and small town development in java: The case of the tegal region, IOP Conference Series: Earth and Environmental Science 447 (2020).

[7] C. Madera, A. Laurent, The next information architecture evolution: the data lake wave, MEDES: Proceedings of the 8th International Conference on Management of Digital EcoSystems (2016) 174–180. doi:`10.1145/3012071.3012077`.

[8] E. Zagan, M. Danubianu, Cloud data lake: The new trend of data storage, in: 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1–4. doi:`10.1109/HORA52670.2021.9461293`.

[9] J. Darmont, C. Favre, S. Loudcher, C. Noûs, Data lakes for digital humanities, in: Proceedings of the 2nd International Conference on Digital Tools amp; Uses Congress, DTUC '20, Association for Computing Machinery, New York, NY, USA, 2020. doi:`10.1145/3423603.3424004`.

[10] Y. Li, A. Zhang, X. Zhang, Z. Wu, A data lake architecture for monitoring and diagnosis system of power grid, in: Proceedings of the 2018 Artificial Intelligence and Cloud Computing Conference, AICCC '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 192–198. doi:`10.1145/3299819.3299850`.

[11] U. Nations, Progress towards the sustainable development goals, 2017.

[12] N. Hossein Motlagh, M. Mohammadrezaei, J. Hunt, B. Zakeri, Internet of things (iot) and the energy sector, Energies 13 (2020). doi:`10.3390/en13020494`.

[13] S. E. Collier, The emerging enernet: Convergence of the smart grid with the internet of things, IEEE Industry Applications Magazine 23 (2017) 12–16. doi:`10.1109/MIAS.2016.2600737`.

[14] P. Wang, L. T. Yang, J. Li, J. Chen, S. Hu, Data fusion in cyber-physical-social systems: State-of-the-art and perspectives, Information Fusion 51 (2019) 42–57. doi:`10.1016/j.inffus.2018.11.002`.

[15] Destination Earth (DestinE) Architecture Validation Workshop, Technical Report, European Commission, 2021.

[16] H. B. Hamadou, T. Pedersen, C. Thomsen, The danish national energy data lake: Requirements, technical architecture, and tool selection, 2020 IEEE International Conference on Big Data (Big Data) (2020) 1523–1532.

[17] re3data.org: Oedi, 2020.

[18] S. Khan, X. Liu, S. Ali, M. Alam, Storage solutions for big data systems: A qualitative study and comparison (2019).

[19] R. Hai, C. Quix, C. Zhou, Query rewriting for heterogeneous data lakes, Advances in Databases and Information Systems - 22nd European Conference (2018) 35–49.

[20] E. Bales, N. Nikzad, N. Q. et al., Citisense: mobile air quality sensing for individuals and communities design and deployment of the citisense mobile air-quality system., 2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops (2012) 155–158.