Data & Knowledge Engineering 148 (2023) 102236

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Challenges of a Data Ecosystem for scientific data

Edoardo Ramalli *, Barbara Pernici *

Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci, 32, Milan, 20133, Italy

ARTICLE INFO

Keywords:
Data Ecosystem
Scientific data
Data management
Data sharing
Data quality
Data transparency

ABSTRACT

Data Ecosystems (DE) are used across various fields and applications. They facilitate collaboration between organizations, such as companies or research institutions, enabling them to share data and services. A DE can boost research outcomes by managing and extracting value from the increasing volume of generated and shared data in the last decades. However, the adoption of DE solutions for scientific data by R&D departments and scientific communities is still difficult. Scientific data are challenging to manage, and, as a result, a considerable part of this information still needs to be annotated and organized in order to be shared. This work discusses the challenges of employing DE in scientific domains and the corresponding potential mitigations. First, scientific data and their typologies are contextualized, then their unique characteristics are discussed. Typical properties regarding their high heterogeneity and uncertainty make assessing their consistency and accuracy problematic. In addition, this work discusses the specific requirements expressed by the scientific communities when it comes to integrating a DE solution into their workflow. The unique properties of scientific data and domain-specific requirements create a challenging setting for adopting DEs. The challenges are expressed as general research questions, and this work explores the corresponding solutions in terms of data management aspects. Finally, the paper presents a real-world scenario with more technical details.

1. Introduction

Data play a central role as a critical resource, offering numerous benefits across diverse domains. The volume of data is steadily growing each day, therefore data platforms serve a fundamental purpose in managing otherwise inaccessible information [1]. Since their first formal definition [2], Data Ecosystems (DE) have been used in many different sectors. The most successful examples can be found in the industry sector [3], where a data-sharing platform creates a marketplace in which the data providers have a business benefit in sharing the data with the data consumer [4]. The DataONE usability and assessment working group[1] has studied data-sharing practices in research. It monitored the practices' evolution over a period of 3–4 years. The results of this study indicate a notable and general increase in the acceptance and willingness to share data, but the inhibitors and promoters of data-sharing practices vary across disciplines. The study also emphasizes the ongoing need to develop an infrastructure that promotes data sharing while accommodating the specific requirements of diverse research communities [5].

The limited access to data generated by other researchers or institutions is recognized as the primary obstacle to scientific progress in many scientific fields [5]. While DE technologies have demonstrated success in the industry sector, their application to scientific

* Corresponding authors.
E-mail addresses: edoardo.ramalli@polimi.it (E. Ramalli), barbara.pernici@polimi.it (B. Pernici).

[1] https://trace.tennessee.edu/utk_dataone/.

https://doi.org/10.1016/j.datak.2023.102236
Received 16 March 2023; Received in revised form 31 July 2023; Accepted 13 October 2023
Available online 19 October 2023

domains is gaining popularity [6,7], even if with fewer examples, likely due to the lower volume of data and a smaller user base in these sectors.

Scientific data are particularly rare and face well-known issues related to data quality [8]. Moreover, the unique data and domain characteristics make general guidelines to design a DE not easily applicable [9]. In many domains, a precise procedure to develop a DE in these sectors is still to be designed [10]. One of the most important uses of scientific data is the development of data-driven predictive models, also using machine learning methodologies. In this area, a DE can help enhance the amount of shared data available and, at the same time, increase the awareness of the importance of data quality [11].

Four types of Data Ecosystems are defined by the level of control over key data resources and participant interdependence [12]: Organizational, Distributed, Federated, and Virtual DEs. In this work, we propose a solution that is a hybrid configuration of organizational and federated DE where the management of the data is centralized, while the infrastructure is federated. This setting enables addressing the challenges that arise from scientific data properties and domain requirements when implementing a DE in scientific domains. Therefore, we first contextualize the different types of scientific data. Then, this work introduces the main challenges and presents them as general research questions. These research questions pose complex challenges because it is difficult to align individual interests with the data-sharing principles of a Data Ecosystem [13]. In order to address them, the research questions are decomposed as data management aspect actions. The paper concludes with a technical discussion based on a specific case study.

The paper is structured as follows. Section 2 discusses related work, while in Section 3, scientific data are introduced, showing how they interact at a business level in the predictive model development process. Section 4 presents the challenges of implementing a DE for scientific data in terms of research questions that result from the scientific data properties and the domain requirements. Section 5 shows the proposed solution to address the research questions, while Section 6 delves into the details of applying this approach in a real-world scenario. Finally, a more comprehensive discussion of the Data Ecosystem's challenges and future developments are discussed in Section 7.

2. State of the art

Data Ecosystems (DE) and, more generally, data spaces are key elements to facilitate knowledge discovery and deliver new technologies. For instance, artificial intelligence (AI) technologies play a significant role in driving the business growth of many companies. However, the development of AI products requires access to large and reliable volumes of data. Collaborative infrastructures like DEs can pave the way for the widespread adoption of AI across companies and all sectors of the economy [14]. Similarly, materials science researchers developed a DE to facilitate the use of machine learning within the community [15]. The DE discovers and collects data from different sources and automatically disseminates the new data across the ecosystem. Another recent application involves condensing heterogeneous COVID-19 data into a DE, building a Knowledge Graph (KG), and offering analysis services to improve the understanding of SARS-CoV-2 infection and its advancement [16].

Several studies have listed the design principles of data spaces [17], the architectural components [18], but in practice, each DE has its challenges, based on the application domain, that require specific customization [19]. If not properly addressed, these challenges can completely preclude the adoption of such a promising infrastructure. For instance, the medical sector requires a high level of trust and security [20], whereas the main challenges in the energy sector also impose to reach the fulfillment of regulations imposed by the data provider [21]. In general, the *fil rouge* between all DEs in terms of challenges consists in building trust between the stakeholders involved [22,23]. To achieve this, a proper data quality assessment and higher transparency are examples of possible solutions [24]. Other factors, such as openness and security, contribute to reaching a critical number of users necessary to keep running the platform. On the other hand, pricing and non-interoperable platforms are among the main failure factors [25]. In many cases, the presence of the role of a coordinator within the data space enhances the trust between the data consumer and the data provider [26]. Over time, four DE typologies have been defined based on the policy to manage the data, the DE goal definition, the degree of participant interaction, and data exchange within the ecosystem [12]. The first two typologies are Organizational and Distributed DEs. Both have a central control system to fulfill a predefined goal, but the DE participants can operate independently in the first one. In the latter, changes in the DE and pooled resources require participant collaboration. Meanwhile, federated and virtual DEs have no central management authority. In federated DEs, participants interact voluntarily to reach a predefined goal, while in the virtual DE, a coalition of participants can emerge to pool resources to achieve a specific goal. These four types of DE describe the edge cases of authority control, resources management, and participant interactions, but not all scenarios can be restricted to design a DE confined to only one of the previously mentioned DE categories.

3. Scientific data

Within this work, the term *scientific data* refers to four different types of data: experiments, predictive models, simulations, and analysis results. These data types are used or produced within a cyclic development process whose goal is to deliver a predictive model. The following paragraphs provide an intuitive explanation of each of these data types and how they interact with each other from a business perspective. In the next sections, we classify and discuss the properties that make them complex to be handled in a Data Ecosystem, together with possible specific domain requirements. Fig. 1 presents a high-level and non-domain specific class diagram of scientific data with their main attributes (or metadata) and their relationships. The detailed class diagram can involve additional classes, attributes, and relationships in a real-world scenario in a specific domain.

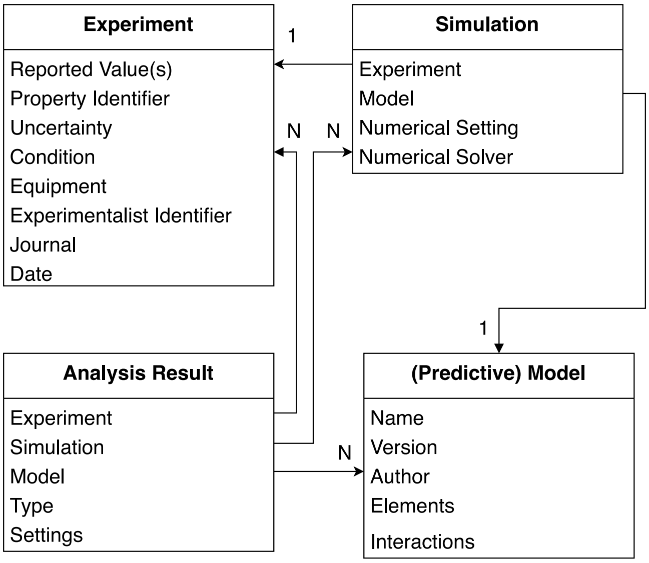

Fig. 1. General class diagram for scientific data.

Experiment. The term *experiment* represents data from an experimental campaign carried out by an experimenter. An experiment is a combination of chemical–physical measurements and metadata associated to it. The measurements account for quantifying a property under investigation, while the metadata provides essential details regarding how and when the measurement is carried out. Examples of this information are the technical instruments used, the environmental setting, the unit of measurement, the identification of the property of interest, the experimental procedure, and so on. In essence, the experiment metadata describes the experimental condition (or setting) of an experimental campaign. On the other hand, the measurement is the numeric value detected by the instruments and thus reported (*Reported value(s)*). Due to intrinsic measurement errors, usually an experimenter carries out an experimental campaign rather than single experiments, i.e., the same experiment is repeated multiple times in the same experimental *condition*, in order to be able to quantify and report the *uncertainty* on the measurements. The experimental condition expressed by the metadata does not vary between the multiple measurements. Usually, the experiments are published on a *date* with an associate publication in a *journal*. In most cases, the digital size of the data necessary to represent the essential digital information, i.e., the *experimental data* that consist of the reported value, together with the metadata, is quite tiny. It is rarely bigger than 10 MB, and often it is less than 1 MB, even if the entire material needed to derive the reported value can have a different order of magnitude in size. Therefore, with a moderate cost in terms of memory, it is possible to store many experimental data.

(Predictive) model. A predictive model, or model, in short, is the leading business driver of a Data Ecosystem for scientific data. The final purpose is to deliver an accurate model to predict unknown outcomes. Nowadays, a popular type of predictive model is neural networks (NN). NNs are black-box methods by design. On the other hand, in the scientific domain, predictive models usually embed chemical–physical laws into equations such as chemical reactions, which cannot be violated. Most of the time, chemical–physical equations can be translated into a set of interpretable differential equations. Thus, these are white-box methods. Black-box methodologies, such as physically informed NN (PINN), can also be employed to develop predictive models for scientific domains. PINNs incorporate chemical–physical laws in the loss function to learn a set of parameters (NN weights), but they are not interpretable. PINN, and in general NN, require much more data for the training and parameters to represent the domain, and they usually do not generalize as well as the white-box methodology. In scientific domains, in the case of white-box models, which elements, which equations, and how to include them in a scientific model is a design choice of the researchers, and it is usually referred to as model parameters. In general, the larger the number of equations in a scientific model, the more complex it is to resolve it and the more accurate. Simplifying, in the general case, a scientific model tries to predict how a chemical–physical system evolves starting from a particular initial condition, solving a set of equations that encode the elements and their interactions in a domain that the model designer decided to represent. Like neural networks, scientific models are also defined as data-driven since real-world observations, i.e., experiments, are used to validate and improve the predictive models.

Simulation. A predictive model, given a set of initial conditions, can forecast the future state of a system. This predicted system state is referred to as a *simulation*, which represents the solution to the model equations obtained through a numerical solver. Numerical solvers are generally very complex, thus time-consuming, since they need to resolve, for instance, differential equations and, in general, need numerical tweaks to address the problem correctly. A wrong or inappropriate configuration of the numerical settings can lead to incorrect results, even if the underlying model is accurate. Additionally, improper numerical parameter settings may lead to excessively long computational times and, in some cases, to failures to terminate the computation. The output file size of a simulation can vary based on the grain of the numerical settings and their complexity. For instance, their file size can range from less than a MB to several dozens of MBs.

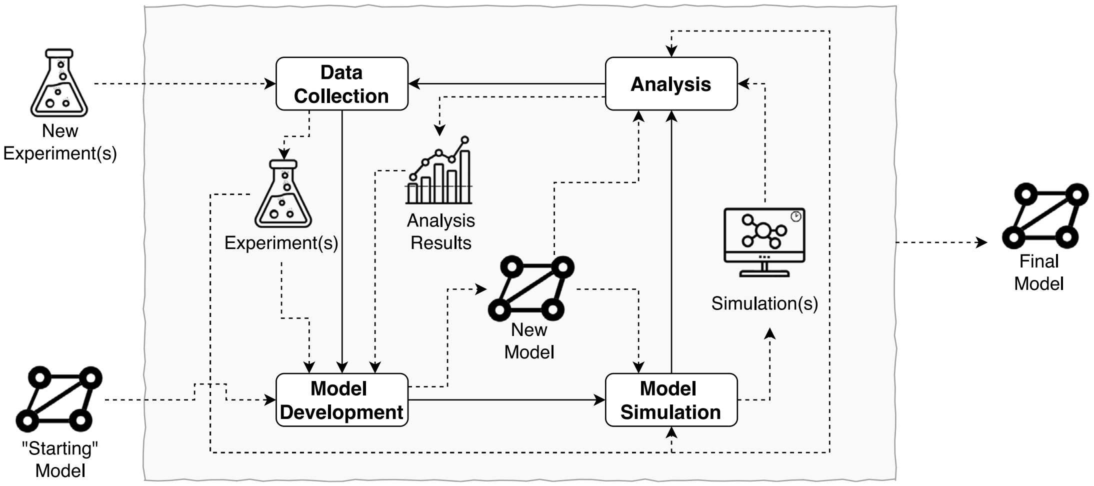

Fig. 2. Business process of the predictive model development loop.

Analysis result. The analysis results in the process of developing a predictive model serve the purpose of generating synthetic information on scientific data. More specifically, it is intended to provide aggregate information on the model's predictive capabilities in different domain settings. Which metric or procedure to employ during the analysis is a parameter of this type of data.

The typical business process of the model development loop in which the four types of scientific data interact is depicted in Fig. 2. This loop includes four main stages. It starts with the collection of experiments. Based on these experimental observations, a new model or an updated version of a pre-existing one (thus coming from outside the process or from a previous iteration of the loop) is generated or improved respectively to represent the new data, if necessary. Subsequently, the predictive model simulates the same domain condition of all available experiments. Finally, the analysis starts. Experiments, i.e., the real-world values, are used to compare the model predictions, i.e., the simulations. This kind of analysis is called model validation, and it is the most important analysis in this process. This data contains synthetic insights about the model performance to guide the next model improvement. This cycle is iterated until the analysis of the results is considered satisfactory or no more experimental data can be obtained.

4. Challenges

This section outlines the key properties of scientific data (Section 4.1) and the specific requirements of scientific domains (Section 4.2) that make the employment of a Data Ecosystem challenging. These challenges arise not solely from the unique characteristics of the data or specific domain requirements, but also from the combination of both. The challenges are then generalized as general research questions (Section 4.3).

4.1. Scientific data properties

This subsection discusses six properties (P) of scientific data that make their management and integration in a Data Ecosystem challenging. Table 1 qualitatively summarizes the high (H), medium (M), or low (L) impact or relevance of property on a specific type of scientific data, as described in Section 3.

P1: Low volume - High cost. Scientific data, unlike other types of data, such as social media, are significantly less available. The collection of experimental data involves on-field measurements using expensive equipment and materials, making it costly in terms of both financial resources and time. Consequently, experiments are often unique and not easily replicable. Similarly, the development of models is a complex process that demands extensive expertise, particularly in white-box approaches, and extensive computational resources and data for building the models in black-box approaches, resulting in only a limited number of models being available for certain scientific domains. Simulations are also available in a limited quantity since they are very pricey in terms of computational resources needed and, in some cases, space to store them. Consequently, the results of analyses based on the other three types of scientific data are limited too. Analyses are generally relatively inexpensive to compute. Although the volume of scientific data is lower compared to other types of data, manual management is still unfeasible and prone to human error. The low volume and high cost of scientific data highlight the need for a data management system, such as a Data Ecosystem, in the various scientific domains to promote the reuse of all types of data and related development and analysis services.

Table 1
Qualitative impact of a scientific data property on the corresponding scientific data type - (H) high, (M) medium, (L) low.

Property	Experiment	Model	Simulation	Analysis result
Low-volume high-cost	H	H/M	H/M	L
Uncertainty	H	H	M	M
Accuracy consistency	H	M	L	L
Heterogeneity	H/M	M	H/M	L
Completeness	M	L	M/L	M
Reproducibility transparency	H	M	M	H

P2: Uncertainty. Uncertainty can be classified into two macro-categories: epistemic and aleatoric. Experiments are real-world measurements, and they are intrinsically affected by aleatoric uncertainty. Repeating the same experimental measurement helps mitigate this issue, quantifying the uncertainty. The reported value for an experiment corresponds to the mean value of the measurements, and the standard deviation corresponds to the uncertainty [27]. The source of uncertainty in the experiment is not only due to measurement error, but there is a set of contributing uncertainty causes. For instance, another source of uncertainty for the experiments is the digitalization of plots from physical documents, such as published papers and reports, to extract the measurement values. Models, on the other hand, are mainly affected by epistemic uncertainty. A model is an approximation of a real-world system, inherently introducing errors. Simulations are, most of the time, deterministic. Repeating the exact simulation of an experiment with the same model leads to the same result. However, numerical errors can also affect the uncertainty of the model's predictions. Analyses are generally not uncertain, even though they are affected by the propagation of uncertainty from the models, simulations, and experiments. The uncertainties present in these underlying components can influence the overall uncertainty in the analysis results.

P3: Accuracy & Consistency. Experimental observations of a system should be close to the (unknowable) ground truth and consistent with each other. If multiple experimental measurements are available from different sources regarding the same experimental conditions, all the reported values should be consistent, also accounting for their uncertainty. In other words, the more the reported values are accurate, the lower the experiment's uncertainty is, the easier it is to detect inconsistencies. Nevertheless, in reality, it is hard to evaluate the consistency of the (many) experiments without uncertainty. Models represent the interaction of the system elements. However, when new elements are being investigated, they may not be standardized in the representation. For instance, models can use the same element name to represent different entities. As a result, it is not easy to compare the simulation results consistently. Numerical solvers differ from each other mainly for numerical implementation choices such as the number of digits or the employment of a particular library. Thus, giving the same model and conditions for forecasting can lead to different numerical solutions (simulations). Usually, the difference is marginal. Finally, concerning analysis data, consistency is almost guaranteed, assuming the analysis procedure is well-detailed and fixed on the same dataset.

P4: Heterogeneity. Scientific data exhibit heterogeneity from three distinct perspectives: type, source, and format. Considering that many scientific domains have been active for several decades, the sources and methodologies for collecting and generating experiments and models have evolved. At the same time, it is likely not to have a standardization by the scientific communities on the representation format. Therefore, there are (sometimes also multiple) de-facto representation formats for all types of scientific data.

P5: Completeness. Experiments are usually produced and collected over several decades. As the way of collecting them and scientific findings change over time, some additional information may become essential to include among the experiment metadata; however, some (old) experimental data may have incomplete information when more recent metadata are considered. Models are incomplete by definition since they simplify a real-world system. For instance, it may not include all the elements of the domain or all the interactions. The effects of such decisions are reflected in the model prediction accuracy. Simulations report all the elements and interactions described in a model, but are quantified along discrete and, thus, not continuous, dimensions.

P6: Reproducibility/Transparency. Experiments are challenging to replicate since it is practically impossible to reproduce the exact initial conditions of an experimental setting. Models and simulations, instead, if adequately documented, are easily reproducible. It is fundamental that models disambiguate the meaning of represented elements and, concerning the simulations, the numerical settings. Regardless of the methodology to derive a predictive model, white or black box, explaining the simulation results could be difficult. Analysis results must be transparent about the analysis process and the computational steps to avoid inappropriate conclusions.

4.2. Domain requirements

The development of a Data Ecosystem in a scientific domain faces not only the challenges posed by the unique properties of scientific data, but also the need to fulfill certain business requirements (R). This work explores these requirements and how they are critical to achieving a higher success rate for the Data Ecosystem within a scientific community.

R1: Keep user involved. Scientific domains are highly active research areas with a long history of studies. Over the years, the research and scientific processes have undergone continuous changes. Nowadays, the workflow is consolidated, but recent technological advancements present new opportunities to improve some aspects of the research process. New technologies such as machine learning promise to enhance the comprehension of phenomena, while data management systems are fundamental to automating and managing an increasing amount of data.

However, even if the new technologies are promising, changing the workflow that has guaranteed continuity of results over the years is problematic. Moreover, these technologies are often distant from a scientific community's expertise, thus, harder to understand and trust. In the end, proposing new technology, such as the Data Ecosystem, requires keeping the final user and the community involved.

In a Data Ecosystem where the central focus is data sharing, if users stop participating, it can lead to a decline in data sharing, ultimately discouraging others from taking part. Two factors require attention to break this negative vicious cycle. First, providing a system that offers all the traditional functionalities and the new ones in a user-friendly way. If some of the traditional functionalities are not present or not working properly, then the final user could not be willing to use multiple methodologies, thereby disrupting their workflow. Although it is acknowledged that software or information systems have a higher probability of failure in the early stages, on the other hand, as user numbers and usage increase, reliability improves. Thus, an increasing number of users translates into more data and greater trust in the shared data and platform. Second, a large initial investment is needed at the beginning to collect shareable data and to implement the Data Ecosystem services and functionalities that can attract users to start using the system and contribute additional data to increase the amount of shared data. Also, in this case, user involvement is important, both for data collection and for defining and testing the needed functionalities, as well as improving them.

R2: Breakdown the cost. Centralized management of data within a Data Ecosystem helps maintain user engagement and keeps them within the system by providing a single, reliable source of data and services. However, while a single organization may support the upfront costs of initial development and data collection, maintaining a data-centric system incurs in a number of operational expenses to guarantee the availability and adequate performance of the system. These costs involve not only maintaining the server infrastructure to provide services and data, but also managing the data itself. However, based on the availability of ad-hoc tools in some scientific domains, some costly and time-consuming data management aspects of the repository, such as data collection, can be automated or semi-automated. For instance, ChemDataExtractor [28] is an automatic tool to extract chemical information from the scientific literature. Additionally, certain simulations and analyses can be computationally intensive, making it impractical for a single organization to bear such expenses. An alternative is the creation of an ad-hoc organization and requesting a fee from its member to sustain these expenses. However, this approach is bureaucratically challenging and discourages the use and interest in adopting the Data Ecosystem. Another alternative solution is a federated or distributed Data Ecosystem. In these settings, each group interested in the Data Ecosystem services or data can create a copy of the repository or the system, eventually share data, manage them as preferred, and dedicate as many resources as desired. It is highly scalable, but it is particularly challenging to maintain all the databases synchronized and the software updated and ensuring a continuous willingness to share data among the participants.

R3: Confidentiality. New experiments and models, due to their development cost and their application in research and industry, are considered particularly precious. Companies and academia are unwilling to give away their work immediately after it is developed. These data give companies and research groups a technological advantage for the next generation of products. Usually, such data become available after publication in journals or patents. Therefore, a Data Ecosystem willing to host scientific data has to guarantee confidentiality while the data has not been published yet. Consequently, two factors have to be enforced. First, the user must be able to trust the infrastructure. Second, it is fundamental that all the data, the ones with confidentiality restrictions and the ones freely available, are in the same system but with the appropriate access rules. Research institutions will use such a system only if a data management system offers this feature. Otherwise, splitting the scientific workflows over different platforms, technologies, and methodologies based on the different confidentiality of the data leads to a heavier workload. Such a cost might not be sustainable to justify the advantages of starting to use a new platform provided by a Data Ecosystem, even with prominent features and services.

4.3. Research questions

The combination of scientific data properties and domain requirements determines the challenges of applying a Data Ecosystem in a scientific field. These challenges are summarized in three general research questions (RQ) that can be helpful when a Data Ecosystem is implemented in a specific scientific domain. Table 2 summarizes and represents the involvement of a property or requisite of a scientific domain within a research question.

RQ1: What services and functionalities should a Data Ecosystem implement when applied to scientific data?

RQ2: Which are the peculiar design choices in the architecture and in the workflow of a Data Ecosystem for scientific data?

RQ3: What are the prerequisites to facilitate adopting and retaining user engagement in a scientific Data Ecosystem?

Table 2
Mapping the involvement of the properties of the scientific data and the domain requirements onto the research questions.

Property	RQ1	RQ2	RQ3
Low-volume high-cost	X	X	X
Uncertainty	X		X
Accuracy consistency	X		X
Heterogeneity	X	X	
Completeness	X		X
Reproducibility transparency		X	X
Requirement	**RQ1**	**RQ2**	**RQ3**
Keep user involved	X	X	X
Breakdown cost		X	
Confidentiality	X	X	X

Table 3
Direct involvements of a data management aspect in the solution of a research question.

Aspect	RQ1	RQ2	RQ3
Data architecture	X	X	X
Data sharing	X	X	X
Data preparation		X	X
Data transparency			X
Data confidentiality	X		X
Data analysis	X	X	X

5. Proposed solutions

This section addresses the research questions that make the employment of a Data Ecosystem for scientific data challenging, proposing possible directions. Starting from the authors' previous work [29], the research questions are tackled by examining six key aspects of data management (Sections 5.1 and 5.6). The proposed solutions carefully balance the design principles of a general Data Ecosystem with the specific properties and requirements of a scientific domain. Table 3 reports a comprehensive overview of the direct involvement of each data management aspect in addressing a research question. It is important to mention that since the research questions are interconnected, the data management solutions are likewise interrelated. Therefore, each proposed solution indirectly influences resolution of other research questions.

5.1. Data architecture

A Data Ecosystem for scientific data requires specific design choices regarding the overall architecture illustrated in Fig. 3. First, it is fundamental to understand the business expectations when employing a Data Ecosystem in a scientific domain. It enables identifying the relevant scientific data and characteristics, required services, and actors involved in the business process. The unique characteristics of the data guide the definition of the database schema. The data management should be centralized. This architectural decision aims to enhance trust, transparency, traceability, and efficiency. The technological implementation of the database could be distributed, even if it is not recommended, due to the risk of consistency issues. Furthermore, central data management encourages users to share data, promoting collaboration and knowledge exchange within the Data Ecosystem. Fig. 3 depicts four types of entities: *data ecosystem*, *user*, *third-party service*, and *worker*. They communicate with each other through the internet, in particular with the HTTPS protocol. The Data Ecosystem offers its services to the users through microservices, through API endpoints. This service-oriented architecture allows flexibility, extensibility, and high maintainability, and users can request and combine services as preferred. All the services of the Data Ecosystem are available through authentication provided by the back end. Such authentication prevents malicious usage of the system and allows giving users different privileges and permissions. In particular, the Data Ecosystem includes four different types of privileges. The *reader* can only read data, while the *writer* can insert data into the system. The *researcher* can edit, delete, and verify the data; instead, the *executor* can use computational resources, hence perform analyses or simulations. These privileges are not exclusive; multiple privileges can be associated with a user. Through the back end, the Data Ecosystem can offer services with a combination of legacy and developed ad-hoc modules, both on-premises or in-cloud, with *third-party services*. Finally, to maintain the scalability of the system, the architecture foresees delegating the computational burden to external *workers* where the Data Ecosystem coordinates and distribute the workload. From this point of view, the infrastructure is provided as a DE federated configuration.

5.2. Data sharing

Data are the key element in a Data Ecosystem, and its sharing functionalities help keep the system active within a community. In the case of scientific data, their value and scarcity make sharing even more crucial. For this reason, a Data Ecosystem for scientific

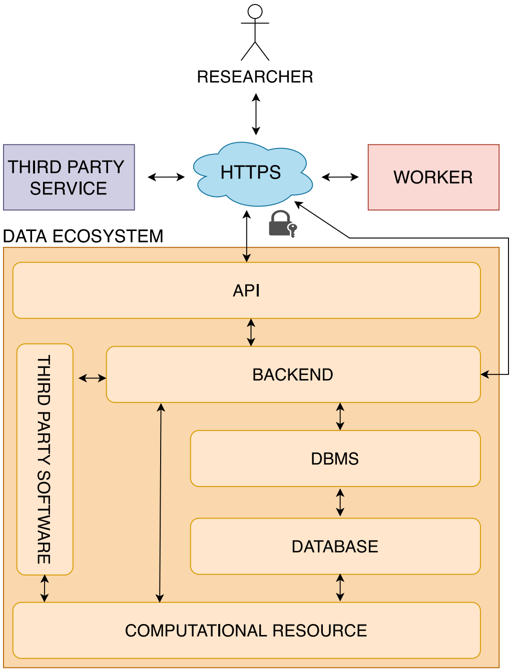

Fig. 3. General overview of the architecture and the main actors of a Data Ecosystem for scientific data.

data has to embrace the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [30], while ensuring confidentiality, as discussed later. Scientific data are inherently uncertain and often come with data quality issues, such as completeness, accuracy, and consistency. A Data Ecosystem, as a centrally managed repository, facilitates information sharing and addresses these problems. It also prevents unnecessary data replication, promoting scientific data reusability. Experiments, models, and simulations in some scientific domains are accumulated over decades, during which the representation format evolves. With a relational database, the Data Ecosystem repository becomes independent of any specific format, and specific translation engines of the different formats can be used to transform the scientific data as desired, making the Data Ecosystem *interoperable*. The service-oriented architecture envisioned for the Data Ecosystem makes it accessible and flexible to be easily integrated with different workflows. The Data Ecosystem should implement controlled access to its data and services. The authentication ensures the traceability of any database change and the trustworthiness of the overall system and avoids malicious uses. However, this level of security does not permit the data to be *findable* from outside the system. Therefore, the Data Ecosystem should create a representation of the scientific data in a well-known representation format and upload it to platforms, such as for instance Zenodo,[2] that associate a Digital Object Identifier (DOI) to data items. By assigning a DOI, scientific data is accessible and searchable from outside, but without providing the full range of services offered within the Data Ecosystem.

5.3. Data preparation

A Data Ecosystem gathers experiments, models, simulations, and analysis results in the same database. This repository enables cross-analyses, leading to new insights about the data. On the other hand, it also introduces the risk of incorrect data quickly influencing other data. To mitigate this risk, identifying user roles and assigning appropriate privileges within the Data Ecosystem helps limit human errors. Moreover, tracing the activity of the users makes it easier to roll back the action and contain the propagation of the error. However, the data preparation phase is fundamental because all data-driven outcomes are highly sensitive to data quality. Therefore, ensuring a certain level of data quality within the Data Ecosystem enhances overall trustworthiness and engagement. Data quality assessment is a precursor to data cleaning and enrichment activities. In the case of scientific data, following the fitness for use concept [31], completeness, consistency, and accuracy are the primary data quality dimension to be investigated. In particular, in the case of experiments, uncertainty is correlated to these three data quality dimensions [32]. Timeliness is not a primary data quality dimension. Experiments are expensive, and replicating them is inconvenient. Thus, updating the old reported values with new ones is rare, even if performed with newer and more precise instruments. On the other hand, simulation and analysis results are deterministic, while models are versioned. In order to guarantee a certain data quality for scientific data, it is necessary to combine automatic and manual controls. The Data Ecosystem should also automatically check the fulfillment of data quality rules

[2] https://zenodo.org/.

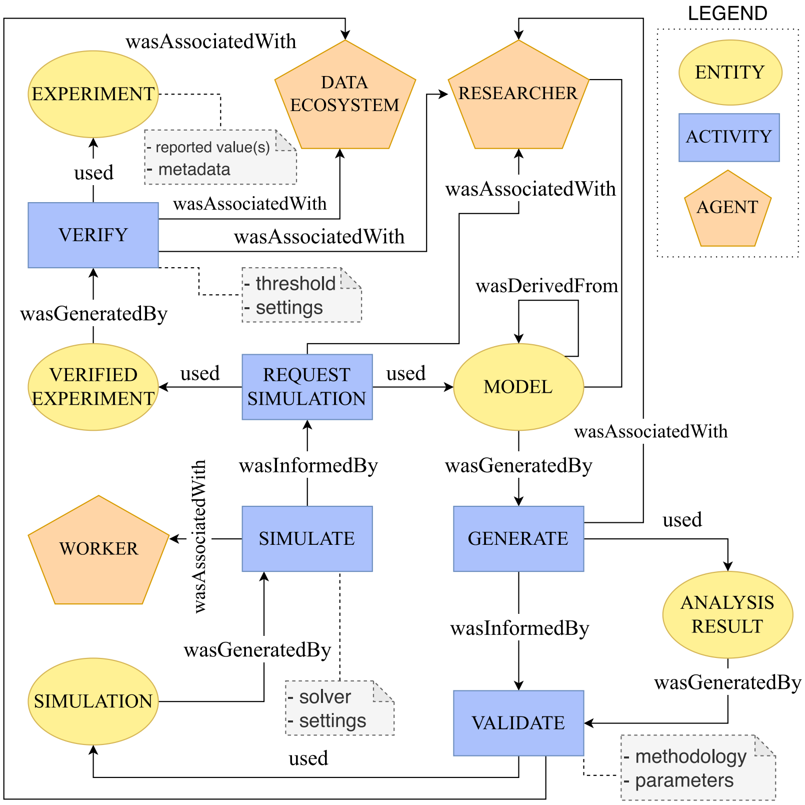

Fig. 4. Representing provenance of the model development process in the context of scientific data within a Data Ecosystem with the W3C PROV-Data model.

for each dimension; however, when it comes to the experiment's accuracy, it is challenging to assess since, by definition, there is no ground truth to compare. However, a peculiarity of scientific data is that its models follow chemical–physical laws, as they represent more or less precisely a domain. Consequently, multiple models, developed independently by different groups, can be used to verify the consistency of the experimental data based on the models' predictions. A large discrepancy between the predictions and the experimental data suggests that the expertise of a designated user is necessary to resolve the ambiguity.

5.4. Data transparency

Data provenance, also known as data lineage, is a branch of data management and data preparation whose scope is to account for the reproducibility and transformation history of the data. Data provenance studies how and what to represent regarding the stages that alter raw data. Usually, provenance enriches the final data with a collection of metadata, named provenance metadata, that encodes all the information about entities, actions, and subjects involved in the different stages of transformation. Provenance is not only fundamental to backtrack the data transformation, but it can also be used to store information about the replicability of the data processing and track down errors in data-driven applications [33,34]. In the end, provenance metadata is designed to ensure data transparency and trustworthiness. The provenance data model defines the structure of the provenance metadata. It specifies how to express the entities, activities, and agents with their interactions that participate in the creation or transformation of data.

This work employs the W3C PROV data model[3] to design the provenance data model. According to the data sheets directives [35], the provenance data model should represent only what is strictly necessary for the subject of the study. In general, a provenance data model can be more or less specific [34], changing the verbosity level.

The W3C PROV data model leverages three concepts to represent the provenance metadata: *Entity*, *Activity*, and *Agent* [36]. An *Entity* is the subject of the provenance; it is something whose evolution is tracked. An *Activity* is an action performed on an entity to produce a new version of itself or another entity. Instead, the *Agent* represents who is responsible for an action or an entity. The W3C data model presented in Fig. 4, using a combination of the "data" and "workflow" levels of detail, proposes the provenance metadata necessary to describe the stages of the business process described in Fig. 2, not including the procedure of external collection of experiments and models. On the other hand, it includes the new components included in the proposed solution in this work, such as

[3] https://www.w3.org/TR/prov-dm/.

the verification step to enhance the quality of data in the repository and the use of a coordinator-worker configuration to distribute the computational workload.

Multiple experiments and models are stored in the Data Ecosystem. An *experiment*, before entering the loop of the model development process, is submitted to a validation process. The *Data Ecosystem* and *Researcher* automatically and manually, respectively, *verify* whether a new *Experiment* is compliant with the established quality thresholds in the policies defined by the Data Ecosystem community. As a result, a *Verified Experiment* is derived and stored within the *Data Ecosystem*. At the first iteration of the loop, a *Model* is given, and the *Researcher requests* the *Simulation* of a set of *Verified Experiments* with the *Model*. At this point, a computational task is assigned to the *Workers* that *simulate* the *Verified Experiment* generating a *Simulation*. Once all the *Simulations* are terminated, the *Data Ecosystem* can perform an analysis of the results, such as model *validation*, thus providing *Analysis Result*, fundamental for the *Researcher* to *generate* a new, improved version of the *Model* and start the model development process again.

5.5. Data confidentiality

A trustworthy Data Ecosystem is one of the requirements for achieving high system engagement. From the security perspective, it means ensuring authentication and confidentiality. Authentication is more related to the business process organization, user roles, and privileges. On the other hand, data confidentiality is related to the management of user interactions across different organizations within the Data Ecosystem in the different stages of the business process. Data confidentiality addresses the domain requirement for scientific data that need to use the Data Ecosystem services and ultimately share the data, but, on the other hand, for some time, during model development, it also needs not to be publicly accessible to guarantee academic and industrial advantages. Requiring all data to be accessible to any user of the Data Ecosystem at all times could potentially disrupt the workflow of a research team based on the usual scientific data management policies. Poor data management policies could ultimately discourage the adoption of the Data Ecosystem. In the scientific data domain, experiments and models are the main types of data that require confidentiality. In particular, new experiments are confidential for the reported values, while their metadata do not have confidentiality constraints. The metadata of the scientific data should be accessible without any restriction. For instance, the metadata of an experiment describe the experimental setting. Knowing all the experimental settings discourages other researchers from investing in performing an experiment that others have already been investigated, even if it has yet to be made publicly available, optimizing the resources and reducing duplicates. Within organizations, groups of users that work together and thus share the same resources can be defined. Examples of organizations are universities, research groups, departments, or research centers. Every user in the Data Ecosystem is affiliated with at least one organization. Some examples of data confidentiality policies are presented in the following. All the scientific data published or generated by a user belong to his/her organization(s). The user, during the data publication, has to specify whether the data are open or closed and under which conditions. All the closed-content data will become open, for example, after an embargo of one year from its insertion in the Data Ecosystem. Within the Data Ecosystem, each user could access all open-content data of all organizations and all the closed-content data belonging to his/her organization(s). Such flexible configuration allows the definition of fine-grained (closed content) sharing policies. For instance, a single experiment or a collection of them can be shared with another organization, or an organization can share all the closed-content data omni- or bi-directional. In terms of implementation, data confidentiality can be ensured by encrypting the confidential information with the private key of the owner of the confidential data.

5.6. Data analysis

Data analysis can extract hidden insights about the analyzed data and enhance the understanding of complex chemical–physical phenomenologies that underlie the model predictions. In the context of this work, data analysis is a fundamental step for the further improvement of a predictive model. A Data Ecosystem facilitates data analysis by providing a centralized repository where all data are readily accessible. Moreover, the data within the Data Ecosystem have undergone a data preparation process, ensuring a certain level of data quality. This addresses two key challenges in data analysis: the quantity and quality of data. Since most data analysis is often domain-specific, it is important to underline two aspects of the design of a Data Ecosystem for scientific data. First, to retain the users within the system, a Data Ecosystem has to offer the same analysis services as the usual – manual – workflow. The Data Ecosystem, in addition, should support the automation of the workflow and be able to mine big amounts of data to extract new insights that were not possible to derive before. Second, the Data Ecosystem should have a micro-services architecture where a web service (HTTPS API) offers each functionality. In such a configuration, the infrastructure is flexible and language-independent. In the majority of the experimental measurement, the final reported value is not a storage burden, but some analysis activities are computationally expensive. Therefore, in the Data Ecosystem design, it is important to identify such functionalities and decide whether to outsource (to the workers) the computational resources or not.

6. Case study

SciExpeM (Scientific Experiments and Models)[4] is a Data Ecosystem developed with the aim to automatize the development process of chemical kinetics predictive models [37]. Currently, it hosts more than 30K experimental data points regarding 2030

[4] https://sciexpem.polimi.it.

different experiments collected by six research groups. It has 206 chemical kinetics models that have generated almost 10K simulations. The database size is around 1 GB.

Not all services and features were implemented immediately in SciExpeM, but it was a gradual process. We have identified three main phases, each with specific goals: prototype, framework, and Data Ecosystem [29]. This step-by-step development methodology, in which features and services, thus complexity, were gradually integrated, not only expedited the delivery of SciExpeM, but also allowed for continuous feedback collection throughout all stages of development. However, this approach is not risk-free. It was necessary to design and develop each version of SciExpeM considering the end goal of delivering a Data Ecosystem. Otherwise, there is a risk of reimplementing or redesigning components that have become central to the project.

The prototype phase aims to identify the domain requirements, the business process, and the types of users and data involved. In the chemical kinetic domain, we observed that the model development process can be outlined as a sequence of small operations that sometimes are repeated. For this reason, a microservices-oriented architecture was a suitable solution for us. During this phase, also the main technological choices were made. In particular, Django is used as the back end since it is versatile and offers functionalities such as authentication and PostgreSQL as a database engine. In the case of SciExpeM, the final database schema is much bigger and more complex than the one presented in Fig. 1 [37], but its core is unchanged. The additional complexity is due to the specific scientific domain's level of detail and complexity, but the database schema presented in Fig. 1 aims to be a template for similar domains and applications.

During the framework stage, SciExpeM evolved from a proof-of-concept into a fully functional system that can be utilized daily by a small group of users. As the system accommodates multiple users, it became crucial to address various data-related challenges, including data preparation, sharing, and analysis aspects. Automatic data quality checks are immediately performed after data insertion or creation to maintain high data quality standards. Data failing these checks are promptly refused. Over time, as the repository accumulated a substantial volume of reliable data, additional services were introduced, such as data cleaning and enhancement using machine learning techniques. For instance, knowledge graph embedding facilitated experimental uncertainty prediction [32]. Within the Data Ecosystem and the organization, different users with varying levels of expertise interact with the platform. Some data quality dimensions can be objectively assessed for scientific data, whereas others, such as the accuracy of the experiments, may not have simple automated detection methods. However, as we explained previously, we can leverage cross-validation of model prediction and experiments to report ambiguous cases automatically. The ambiguity is resolved by the verification of the experiment by an experienced user, and authentication can be used to check if a user has the privileges for this kind of operation. An ideal validation process for the experiments involves multiple validation iterations by users with diverse backgrounds, i.e., from different research groups. For efficient data analysis and improved user interaction, the system stores analysis results for reuse. By doing so, users can access and reference previous analyses, making their interaction with the Data Ecosystem more responsive and efficient.

Finally, in the last stage of the development of the system, SciExpeM becomes a Data Ecosystem, since it implements also data transparency and data confidentiality functionalities. Regarding the former, since all the functionalities and data are offered through the APIs, we can easily trace all the events that modify the data and assemble the corresponding provenance records. For the latter, our deployment uses a database that is not distributed. Django interacts with the DBMS, and the query results are filtered and decrypted according to the data confidentiality policies with respect to the user requesting the data.

In the end, implementing a Data Ecosystem in the scientific domain starts with analyzing a business process. However, SciExpeM not only incorporates such a process speeding it up and bringing new functionalities, but it also demonstrates that the Data Ecosystem functionalities are fundamental in a proper research procedure, suggesting changes in the business process itself [38].

7. Concluding remarks

This work discusses the adoption of a Data Ecosystem (DE) solution for scientific data to facilitate research workflows. It examines the challenges and the corresponding proposed solutions.

Four types of scientific data are considered: experiments, predictive models, simulations, and analysis results. These data types are the main data entities involved in the predictive model development process.

A DE in scientific domains can bring tremendous benefits for three reasons: first, due to the scarcity and cost of scientific data, a DE can enhance data sharing within the research community. Second, since the predictive model development process foresees time-consuming steps and error-prone sequences of tasks on scientific data, the DE can transform the tasks into services that can help automate and speed up research workflows. Finally, collecting and organizing the information in a DE can open new frontiers to discovering hidden insights in analyzing large amounts of data.

Although using a DE in this field promises great results, some challenges may arise and impede successful accomplishment. The challenges to adopting a DE in a scientific domain result from the combination of specific domain requirements from the scientific research community and distinctive properties of scientific data. These properties vary from high heterogeneity and uncertainty to the low volume of data that complicates the assessment of accuracy and consistency. The domain requirements can be summarized as confidentiality, affordability, and trustworthiness. Experiments and models are very confidential contents that can give a considerable industrial and academic advantage. Therefore, the developers are not willing to share data immediately but only after some time. As a consequence, a Data Ecosystem has to provide a complete set of services that do not split the workflow based on data confidentiality. In fact, as long as the final user has to use multiple workflows because the DE only includes some of the usual functionalities, no one has sufficient incentives to switch to a new technology. Therefore, the DE services have to be complete and reliable. The effects of the challenges can be perceived in terms of user engagement levels. Low engagement determines fewer shared data and, thus, again, a lower number of users and level of trust in the DE platform. In addition, sustaining the costs of a Data Ecosystem, both in terms of infrastructure, management, and maintenance, require a scalable and sustainable solution.

The paper agglomerates these challenges as research questions that can be used in other domains or scientific applications. Subsequently, it addresses these research questions by examining solutions for different aspects of data management. The literature has examples of centralized, federated, or distributed DE. In a centralized or distributed DE, there is a central data management authority, but the centralized one allows for the independent participation of the DE users, thus not requiring coordination between the participants to keep the DE running with all the available data. A centralized data management system has major control over the data and makes it easier to track the use, misuse, and eventually, the right to be forgotten of data with simpler, less expensive, and binding technologies such as the blockchain. At the same time, such DE can also easily perform other analyses on the repository, such as data diversity, which is a critical property for data ethics. Ideally, the data management authority should be a no-profit superparty organization, such as IEEE, well-known in the field and trusted, whose ultimate goal is promoting cooperation and knowledge. A centralized data control system also allows for more reliable findability. In a distributed DE, even if the data management is centralized, the availability of some functionalities or the entire repository requires coordination and reliable participants. For instance, if one participant became offline, not all the data could be available to the remaining participants. Therefore, assessing some analyses, such as the data diversity of the repository, could lead to incomplete or unreliable results. A similar situation occurs in a federated DE with no central data management authority. For instance, the data quality policy to accept the data in the repository, such as the required metadata to describe an experiment, could differ based on the participant. In a centralized DE, the availability of services and data does not depend on the participants, and the data management policies are the same among all the participants. On the other hand, this approach is easier to scale up. In this paper, to conciliate the design principles of DE and the challenges of scientific data, we propose a hybrid DE configuration: central data management with federated computational resources. This architecture is a good trade-off between the centrality of an organizational DE that has control over the data, but simultaneously encourages and promotes data sharing and the scalability of the system. In our setting, DE is the coordinator, and each organization can decide to make available one or more workers for the computational tasks requested by its users. The paper concludes with a final discussion of a real-world application of a DE in the chemical engineering domain. Future research directions concern studying a more formal definition and application of the data confidentiality policies and guaranteeing intellectual property that, from our experience interacting with the domain stakeholders, are the primary "deal-breaker" and the limitation factor of applying a DE in a scientific domain.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of Edoardo Ramalli is supported by the interdisciplinarity Ph.D. project of Politecnico di Milano, Italy. The work of Barbara Pernici is supported by the PNRR-PE-AI "FAIR" (Future Artificial Intelligence Research) project funded by the NextGenerationEU program, Italy.

References

[1] V. Stodden, The data science life cycle: a disciplined approach to advancing data science as a science, Commun. ACM 63 (7) (2020) 58–66.
[2] M.I.S. Oliveira, B.F. Lóscio, What is a data ecosystem? in: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, 2018, pp. 1–9.
[3] M. Jarke, B. Otto, S. Ram, Data sovereignty and data space ecosystems, Bus. Inf. Syst. Eng. 61 (5) (2019) 549–550.
[4] B. Otto, M. ten Hompel, S. Wrobel, Designing Data Spaces: The Ecosystem Approach To Competitive Advantage, Springer Nature, 2022.
[5] C. Tenopir, E.D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, K. Dorsett, Changes in data sharing and data reuse practices and perceptions among scientists worldwide, PLOS ONE 10 (8) (2015) e0134826.
[6] J. Gelhaar, T. Groß, B. Otto, A taxonomy for data ecosystems, in: Proceedings of the 54th Hawaii International Conference on System Sciences, 2021, pp. 6113–6122.
[7] J. Gelhaar, T. Gürpinar, M. Henke, B. Otto, Towards a taxonomy of incentive mechanisms for data sharing in data ecosystems, in: Pacific Asia Conference on Information Systems, 2021, p. 121.
[8] C. Batini, M. Scannapieco, Data and Information Quality: Dimensions, Principles and Techniques, Springer, 2016.
[9] L. Nagel, D. Lycklama, How to build, run, and govern data spaces, in: Designing Data Spaces: The Ecosystem Approach To Competitive Advantage, Springer International Publishing Cham, 2022, pp. 17–28.
[10] M.I. S. Oliveira, G.d.F. Barros Lima, B. Farias Lóscio, Investigations into data ecosystems: a systematic mapping study, Knowl. Inf. Syst. 61 (2019) 589–630.
[11] C. Cappiello, W. Samá, M. Vitali, Quality awareness for a successful big data exploitation, in: Proceedings of the 22nd International Database Engineering & Applications Symposium, 2018, pp. 37–44.
[12] E. Curry, A. Sheth, Next-generation smart environments: From system of systems to data ecosystems, IEEE Intell. Syst. 33 (3) (2018) 69–76.
[13] I. Jussen, J. Schweihoff, V. Dahms, F. Möller, B. Otto, Data sharing fundamentals: Definition and characteristics, in: Proceedings of the 56th Hawaii International Conference on System Sciences, 2023, pp. 3685–3694.
[14] D. Hecker, A. Voss, S. Wrobel, Data ecosystems: A new dimension of value creation using AI and machine learning, in: Designing Data Spaces, 2022, p. 211.
[15] B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, I. Foster, A data ecosystem to support machine learning in materials science, MRS Commun. 9 (4) (2019) 1125–1133.

[16] A. Sakor, S. Jozashoori, E. Niazmand, A. Rivas, K. Bougiatiotis, F. Aisopos, E. Iglesias, P.D. Rohde, T. Padiya, A. Krithara, et al., Knowledge4COVID-19: A semantic-based approach for constructing a COVID-19 related knowledge graph from various sources and analyzing treatments' toxicities, J. Web Semant. 75 (2023) 100760.

[17] B. Otto, S. Lohmann, S. Auer, G. Brost, J. Cirullies, A. Eitel, T. Ernst, C. Haas, M. Huber, C. Jung, et al., Reference Architecture Model for the Industrial Data Space, Fraunhofer-Gesellschaft, 2017, http://dx.doi.org/10.24406/publica-fhg-298818, URL https://publica.fraunhofer.de/handle/publica/298818.

[18] Y. Demchenko, C. De Laat, P. Membrey, Defining architecture components of the Big Data Ecosystem, in: 2014 International Conference on Collaboration Technologies and Systems (CTS), IEEE, 2014, pp. 104–112.

[19] C. Cappiello, A. Gal, M. Jarke, J. Rehof, Data Ecosystems: Sovereign Data Exchange among Organizations (Dagstuhl Seminar 19391), Dagstuhl Rep. 9 (9) (2020) 66–134, http://dx.doi.org/10.4230/DagRep.9.9.66, URL https://drops.dagstuhl.de/opus/volltexte/2020/11845.

[20] T. Berlage, C. Claussen, S. Geisler, C.A. Velasco, S. Decker, Medical data spaces in healthcare data ecosystems, in: Designing Data Spaces: The Ecosystem Approach To Competitive Advantage, Springer International Publishing Cham, 2022, pp. 291–311.

[21] V. Janev, M.-E. Vidal, D. Pujić, D. Popadić, E. Iglesias, A. Sakor, A. Čampa, Responsible knowledge management in energy data ecosystems, Energies 15 (11) (2022) 3973.

[22] J. Gelhaar, B. Otto, Challenges in the emergence of Data Ecosystems, in: Pacific Asia Conference on Information Systems, 2020, p. 175.

[23] D. Lis, B. Otto, Data governance in data ecosystems–insights from organizations, in: Proc. AMCIS, 2020, p. 20.

[24] S. Geisler, M.-E. Vidal, C. Cappiello, B.F. Lóscio, A. Gal, M. Jarke, M. Lenzerini, P. Missier, B. Otto, E. Paja, et al., Knowledge-driven data ecosystems toward data transparency, ACM J. Data Inf. Qual. (JDIQ) 14 (1) (2021) 1–12.

[25] L. Özcan, C. Koldewey, E. Duparc, H. van der Valk, B. Otto, R. Dumitrescu, Why do digital platforms succeed or fail? – A literature review on success and failure factors, in: Proc. AMCIS, 2022, p. 15.

[26] B. Otto, The evolution of data spaces, in: Designing Data Spaces: The Ecosystem Approach To Competitive Advantage, Springer International Publishing Cham, 2022, pp. 3–15.

[27] R.J. Moffat, Using uncertainty analysis in the planning of an experiment, Trans. ASME J. Fluids Eng. 107 (2) (1985).

[28] M.C. Swain, J.M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, J. Chem. Inf. Model. 56 (10) (2016) 1894–1904.

[29] E. Ramalli, B. Pernici, From a prototype to a data ecosystem for experimental data and predictive models, in: Proc. of the First International Workshop on Data Ecosystems (DEco'22), CEUR-WS, 2022, pp. 18–26.

[30] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (1) (2016) 1–9.

[31] R.Y. Wang, D.M. Strong, Beyond accuracy: What data quality means to data consumers, J. Manage. Inf. Syst. 12 (4) (1996) 5–33.

[32] E. Ramalli, B. Pernici, Knowledge graph embedding for experimental uncertainty estimation, in: Information Discovery and Delivery, 2023, http://dx.doi.org/10.1108/IDD-06-2022-0060, ahead-of-print.

[33] K. Cranmer, L. Heinrich, R. Jones, D.M. South, et al., Analysis preservation in ATLAS, in: Journal of Physics: Conference Series, Vol. 664, IOP Publishing, 2015, pp. 1–5.

[34] M. Herschel, R. Diestelkämper, H.B. Lahmar, A survey on provenance: What for? What form? What from? VLDB J. 26 (6) (2017) 881–906.

[35] T. Gebru, J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H.D. Iii, K. Crawford, Datasheets for datasets, Commun. ACM 64 (12) (2021) 86–92.

[36] K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, et al., PROV-DM: The PROV data model, W3C Recomm. 14 (2013) 15–16.

[37] E. Ramalli, G. Scalia, B. Pernici, A. Stagni, A. Cuoci, T. Faravelli, Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering, Front. Big Data 4 (2021) 663410.

[38] E. Ramalli, T. Dinelli, A. Nobili, A. Stagni, B. Pernici, T. Faravelli, Automatic validation and analysis of predictive models by means of big data and data science, Chem. Eng. J. 454 (2023) 140149.

Edoardo Ramalli received his master degree in computer science and engineering from Politecnico di Milano in 2020. From 2020 he is a Ph.D. candidate in information technology under the supervision of prof. Pernici. His main research interests include knowledge extraction and management in data ecosystems, mainly for developing predictive models.

Barbara Pernici is full professor of Computer Engineering at the Politecnico di Milano. She has published more than 70 articles in international journals and about 350 papers at international level. Her research interests include adaptive information systems, data quality, IS energy efficiency, and social media analysis. She is a member of the Editorial Board of the IEEE Transactions on Services Computing and ACM Transactions on the Web.