Journal of Big Data

RESEARCH Open Access

An exploratory approach to data driven knowledge creation

Costantino Thanos, Carlo Meghini, Valentina Bartalesi* and Gianpaolo Coro

*Correspondence:
valentina.bartalesi@isti.cnr.it

Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" (ISTI), CNR, Pisa, Italy

Abstract

This paper describes a new approach to knowledge creation that is instrumental for the emerging paradigm of data-intensive science. The proposed approach enables the acquisition of new insights from the data by exploiting existing relationships between diverse types of datasets acquired through various modalities. The value of data consistently improves when it can be linked to other data because linking multiple types of datasets allows creating novel data patterns within a scientific data space. These patterns enable the *exploratory data analysis*, an analysis strategy that emphasizes incremental and adaptive access to the datasets constituting a scientific data space while maintaining an open mind to alternative possibilities of data interconnectivity. A technology, the Linked Open data (LOD), was developed to enable the linking of datasets. We argue that the LOD technology presents several limitations that prevent the full exploitation of this technology to acquire new insights. In this paper, we outline a new approach that enables researchers to dynamically create data patterns in a research data space by exploiting explicit and implicit/hidden relationships between distributed research datasets. This dynamic creation of data patterns enables the exploratory data analysis strategy.

Keywords: Data exploration, Data relationships, Data patterns, Data analyzer, Data publication

Introduction

Data Intensive Science is considered to be the fourth paradigm of science, after the empirical, theoretical and computational paradigms [1]. It is seen as a data driven science where a set of Information Technology (IT) tools and technologies support scientists in analyzing huge volumes of scientific data to discover new insights. There are two different scientific scenarios where the scientific data analysis is carried out. One scenario is characterized by the presence of discipline-specific large databases/data warehouses (big data); the other scenario is characterized by the presence of small research groups, worldwide distributed, that produce and store large volumes of data. The first scenario is typical of some scientific disciplines like astronomy, high-energy physics. To retrieve the data to be analyzed, the scientist has to deal with a big challenge, that is, the inadequacy of the conventional query processing technology as the huge volumes of data have outgrown the capabilities of this technology. In addition, the scientist has to explore the whole database even though the needed data is a small part of it. Indeed, the database management systems are conceived in such a way that in order to provide a complete answer to a query they have to explore the whole database. Querying a very large database is a very expensive and time-consuming task. Moreover, the scientist, frequently, does not know exactly what kind of queries to issue. The second scenario is characterized by the presence of small research groups worldwide distributed. These groups accumulate and store large volumes of data. In such a scenario the scientist, has to interact with several database systems in order to retrieve disparate data sets to be analyzed. The challenge here is the need to interact with database systems that adopt different data models, query languages and protocols. To alleviate the problems that arise in the second scenario, the data integration technology is employed [2]. However, the traditional approach to data integration, based on the design of a unified view (global schema) is technically unfeasible. First, due to the extreme heterogeneity of the local schemata to be integrated, the design of a global schema is a very complex task. Second, the local schemata evolve over time and therefore the global schema should undergo continuous restructuring. Another approach, which is alternative to the data integration, has been proposed the "Dataspace" approach [3]. This approach is a data co-existence approach. The goal is to provide through a platform base functionality over a large number of diverse, interrelated data sets, but having no means of managing them in a convenient, integrated fashion. However, this approach has never been implemented. In conclusion, in both scenarios a scientist has to deal with big technological challenges that make her/his task, i.e. locating interesting data for his research activity prohibitive. To support scientists in overcoming these complex technological problems, a new paradigm of query processing has emerged: data exploration [4]. The data exploration approach enables the incrementally and adaptively exploration of a database until a subset of it that satisfies an information need is obtained. Based on this new paradigm of data seeking also a new approach to data analysis has emerged: Exploratory Data Analysis (EDA). In essence, EDA is a strategy of data analysis that emphasizes maintaining an open mind to alternative possibilities in order to gather as many insights as possible from the data [5]. In the first scenario, an exploratory approach to extracting knowledge from data, known as Knowledge Discovery in Databases (KDD), has been defined and employed. It is a process that aims at identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6]. An important step of this process is Data Mining. Data mining can be defined as the computational process of analyzing large amounts of data in order to extract patterns and useful information [7]. This approach can be successfully employed in discovering patterns of data contained in large databases or data warehouses where plans and protocols can be put in place in order to efficiently manage and facilitate accessing the data. However, some challenges must also be faced, for example, the dynamicity of data that can invalidate previously discovered patterns, as well as the existence of complex dynamic data relationships. In the second scenario, an exploratory approach to extracting knowledge from data, can be conducted based on the Linked Open Data (LOD) technology [8]. This technology supports identification of data resources via HTTP Internationalized Resource Identifiers (IRIs), making them web resources. These IRIs can then be used to express knowledge about these resources via one of the recommended Semantic Web languages (RDF,[1] OWL[2]), and to create the links between them that constitute the patterns/graphs as the basis of the above described exploratory approach. In other words, RDF/OWL graphs have the potential to constitute data patterns needed to implement the exploratory approach to the creation of knowledge, based on the interconnection of research datasets. However, there remain some limitations that prevent the full exploitation of this technology in its current state. We have identified four main limitations.

1 From a conceptual point of view, the number of existing types of relationships between datasets is far more than the relationships defined and axiomatized by existing ontologies. This is a significant limitation to the expressiveness of any linking mechanism.
2 From a practical point of view, a considerable number of relationships between datasets are unknown or hidden/implicit. Therefore, to be expressed in an RDF/OWL graph, they must be inferred (discovered). We still lack a systematic study of the algorithms to carry out this kind of discovery, which further limits the potential of the linking mechanism.
3 The dynamic nature of the relationships existing between research data sets. In essence, an existing relationship between two datasets can disappear or change its type as a consequence of the dynamicity of the "intention" of the datasets involved in the relationship.
4 A domain-specific data space organized according the LOD technology is composed of a number of pre-established patterns/graphs that a researcher has to follow in order to obtain the information she/he is looking for. In essence, she/he has to traverse a static data structure. On the contrary, the exploratory activity requires a researcher be able to dynamically create links between datasets on the basis of her/his cognitive state. The researcher's cognitive state is continuously updated as the investigation proceeds and, therefore, new information needs that require the discovery of additional relationships can arise. In essence, a researcher needs some tools that enable her/him to discover relationships among datasets dynamically.

In this paper, we present a new exploratory approach to knowledge production that is different from the previously mentioned approaches, i.e., data mining and LOD. Data mining aims to detect data patterns that are already present in the data. LOD establishes data patterns by connecting discipline-specific data sets by exploiting known relationships between them that are defined by discipline-specific ontologies. The proposed approach has been fostered by some characteristics of big research data, i.e., high dynamicity, high dimensionality and high relationality. In fact, in many scientific domains (e.g., genomic data) the number of attributes associated with the entities described in a dataset can become very large (high dimensionality). The proliferation of the attributes of data entities causes also the proliferation of relationships among attributes of data entities. This proliferation increases the connectivity of distributed datasets. Our

[1] https://www.w3.org/TR/rdf-schema/

[2] https://www.w3.org/TR/owl2-overview/

approach enables the researcher to discover existing implicit relationships (causal/semantic/temporal/spatial) as well as correlations between scientific data sets. Then the researcher, by materializing them through links, creates data patterns. These data patterns are dynamically created as the data exploration process proceeds. By following these patterns new insights can be produced. In essence, our approach can be considered as a preliminary step of the EDA. In fact, a data pattern that evidences some new insights would be regarded as a starting point for hypotheses generation. In essence, the proposed approach aims at enabling researchers to create insightful data patterns that allow them to build mental models of the phenomenon being studied. Finally, it must be stressed that an exploratory approach to knowledge creation is feasible only in the context of Open Science. That is, a policy framework that prescribes the open sharing of all research outcomes as early as is practical in the discovery process. The European Commission has endorsed the Open Science policy and all the projects funded by it must observe this policy.

The paper is organized as follows: in "The scientific data space" section, the characteristics of the scientific data space that mainly influence the proposed exploratory approach are described. In "Research data relationships" section, the different types of relationships, both explicit and hidden (implicit), that exist between datasets and that allow the creation of a linked scientific data space are described. In Exploring the scientific data space section, different approaches to exploring the linked scientific data space are illustrated. In "Data analyzers" section, the characteristics of the data analyzers, that is, the software that discovers relationships between datasets are described. In "An example of discovering an existing extensional datarelationship between two datasets" section, a use case is illustrated. Finally, in "Concluding remarks" section, some concluding remarks are given. In what follows, we will consider “dataset” as a synonym of “big dataset”, for generality.

The scientific data space

The current Scientific Data Space is composed of a large number of research data sets and of relationships between them. A dataset has an intension and an extension. The intension of a dataset describes the structure of the dataset and is expressed as a schema of the particular data model that the dataset conforms to. The extension of a dataset is a set of data that are structured according to the dataset intension. The research datasets are organized and managed in various ways, depending on the scientific context within which these data sets have been collected or created: some are traditional relational databases, others are XML document repositories, others are Linked Data, and so on. In many cases, such as institutional repositories, more than one technology is used to manage the same data collection.

Research Datasets. Research data can be classified in a variety of ways. Classifying types of research data can be helpful for understanding the similarities and differences as well as the intended and potential use of data over time. The US National Science Board (NSB) [9] classifies digital research data based on way the data was collected or generated: observational, computational, or experimental. According to the NSB, observational data cannot be recollected and are archived indefinitely. These data are typically time and/or location dependent. The observational context, including time, location, and method of collection, is essential to facilitating data reusability. Data that is the result of computer models or simulations (computational data) can be reproduced if adequate information is provided about the computer hardware, software, and inputs. Experimental data can often be reproduced, provided that the experimental conditions are known. These data are associated with a particular methodology or instrument. It must be emphasized that research datasets are highly dynamic; their dynamicity can be both extensional as well as intentional. By extensional dynamicity it is intended the changes in the extension of a dataset caused by a number of operations like the acquisition of more research data, or the modification of existing data or the elimination of some data. By intentional dynamicity it is intended the change of the structure of real world objects represented in research datasets as new insights are gained in the scientific domain.

Dataset Identity. Each of above dataset classes has distinguished characteristics that contribute in defining the identity of a dataset. By dataset identity it is intended a number of characteristics that make a dataset definable and recognizable allowing, thus, to distinguish it from other datasets but also to discover relationships between different datasets [10]. Identity must be an intrinsic characteristic of the dataset and, therefore, independent from its structure/format which may change over time. Several characteristics concur to establish the identity of a dataset [11]; for the purpose of identifying relationships between research datasets, we consider the following characteristics of datasets as very important:

- Class: as said, three main dataset classes have been identified based on the collection method: observational, computational and experimental.
- Relatedness: datasets are collections of data that are related to each other in several ways. Four types of relatedness are important for our study: circumstantial relatedness, temporal, spatial and semantic relatedness. Circumstantial relatedness refers to the context within which the dataset has been created; temporal relatedness refers to the time interval during which the dataset has been produced; spatial relatedness refers to the location where the dataset has been produced; and semantic relatedness refers to the fact that the data contained in a dataset concerns the same subject or has a common theme.
- Purpose: research datasets are created in order to support a scientific investigation. The purpose of this investigation constitutes a distinguished characteristic of a dataset.

Metadata Schemes for Research Datasets. A metadata scheme "is a logical plan showing the relationships between metadata elements, normally through establishing rules for the use and management of metadata specifically as regards the semantics, the syntax and the optionality of values" (ISO 23081). The metadata scheme of a research dataset must, formally, define those elements that concur to establish the dataset identity [12]. For each dataset class, the metadata scheme will contain elements that characterize this particular dataset class, for example, time, location, context, procedures, theme, purpose. Having classified research datasets into three classes implies that, also, the associated metadata will have different features related to each category [13]. For example,

- metadata schemes of observational datasets should include temporal information, spatial information as well as information about the observational context and the collection method;
- metadata schemes of experimental datasets should include information about the adopted methodology as well as the instrument employed;
- metadata schemes of computational datasets should include information about the data service (software) used as well as the necessary input in order to produce the dataset described by the metadata.

Several metadata schemes have been formally defined and some of them have been adopted as domain-specific standards; some other standards are under development.

Database Abstractions/Views. Research datasets contain huge amounts of data. Usually, researchers are interested only in some parts of a dataset. These parts (called sub datasets) are known as dataset views. Dataset views can be considered as data abstractions of an epistemological nature [14]. The epistemological approach to abstraction is concerned with the different levels of observation or interpretation at which a dataset can be studied. For example, a dataset can be observed and analyzed at different levels of abstraction, with regard to time, place, instrument, or object of observation. Examples of epistemological levels of abstraction are spatial and temporal data abstractions. A dataset view can also be defined as a function [15] that when applied to a dataset produces a subset of that dataset. Obviously, each dataset view must have a well-defined identity. We think that each large dataset should be endowed with a number of (possibly overlapping) views. In summary, we foresee that in the near future the scientific data space will be constituted by a large number of widely distributed dataset views interconnected by several kinds of relationships (explicit or hidden). A step towards this direction is the formal definition of domain and/or class specific metadata standards.

Data Publication. Finally, an emerging approach in the scientific communication that is instrumental in the discovery of datasets/views and therefore, in their interconnection is Data Publication. By Data Publication, we mean a process that allows researchers to discover, understand, and make assertions about the trustworthiness and fitness for purpose of the datasets/views in a data space. The ultimate aim of Data Publication is to make scientific data available for reuse both within the original disciplines and the wider community. Among the main functions that the data publication process performs, we distinguish the following two that are of paramount importance for the creation of data patterns: data registration and data semantic enhancement. The purpose of registration is to make a dataset/view citable as a unique piece of work, while the purpose of semantic enrichment is to make it understandable. We expect that, in the near future, domain-specific registries will be developed where the datasets/views produced by research activities will be published. Once accepted for deposit, a dataset/view should be assigned a “Digital Object Identifier” (DOI) for registration. A DOI [16] is a unique name (not a location) within the scientific data universe and provides a system for persistent and actionable identification of data. In addition, the dataset/view should be assigned appropriate metadata. An emerging best practice that supports the re-usability of research data is the FAIR principles [17] that aim at making research data findable, accessible, interoperable and re-usable.

Research data relationships

Research data relationships are of paramount importance for the implementation of the exploratory approach to knowledge production. Relationships between datasets can exist both at the extensional and at the intentional level. Relationships at the extensional level have been studied for the purpose of taking under control the proliferation of datasets produced in the context of a research project. By an extensional relationship between two datasets we mean a property that depends solely on the current data of these datasets In particular, assuming that two datasets are implemented as tables (relations), a number of relationships between data in spreadsheets have been identified: row containment, column containment, containment, sub-containment, complementation, equal, incompatible, and others [18]. By intentional relationship between two datasets we mean a relationship that exists between elements of the corresponding metadata schemes. This relationship can be explicit or implicit.

Explicit Dataset Relationships. An explicit relationship between two datasets exists when it is represented by common elements in their respective metadata schemes. For example, in a relational database an explicit relationship between two relations/tables exists when one table has a foreign key that references the primary key of the other table. Explicit relationships are intentionally created by the designers of database schemes. The establishment of explicit relationships between datasets is facilitated by the adoption of standard metadata schemes for each dataset class. Datasets belonging to the same class share the same metadata scheme and therefore, the same metadata elements. In this case, a query processor is able to identify a relationship between two datasets. For example, in a relational database a query processor, based on the relational calculus, is able to identify existing relationships between datasets. In the case of two observational datasets, the existence of the element time in the corresponding metadata schemes enables a query processor, based on temporal logic, to discover a temporal relationship between the two datasets. More problematic is the discovery of relationships between datasets belonging to different classes and therefore having different metadata schemes. In this case, a query processor, in order to identify an existing relationship between two datasets must be supported by domain-specific ontologies and their alignment.

Implicit Dataset Relationships. An implicit relationship between two datasets exists when there are no common elements in the metadata schemes of the two datasets, but there exist a relationship (for example semantic, causal, spatial) between elements of the corresponding metadata schemes. This type of relationship can be discovered by a query processor based on a logic that depends on the type of the sought relationship, for example, modal logic, causal logic, etc. Obviously, hidden relationships are not intentionally created but they arise in a continuously way as the research datasets produced by research teams are dynamically created.

Here below we, briefly, describe some important intentional relationships between datasets.

Semantic relationships. A semantic relationship between two datasets is the association that exists in the domain of discourse between the objects that the datasets represent. Typically, these objects are represented in words/phrases contained in the metadata associated with the datasets, so that we can simply say that the semantic relationship concerns the meanings of words/phrases contained in the metadata associated with these datasets. In the literature [18] a list of 31 semantic relationships has been provided. Among them it is worthwhile to mention:

Inclusion relationship that describes situations where one entity comprises or contains other entities. Three different types of inclusion have been identified: class, meronymic, and spatial.

- Class inclusion is the standard subtype/supertype relationship often expressed as is-a, (A is-a B, where A is referred as the specific entity type of B). Other examples include: relationships of classification, generalization, and specialization.
- Meronymic inclusion is the relationship between something and its parts. Examples include the relationships: component-object, member-collection, phase-activity, and place-area.
- Spatial inclusion is the relationship between an object and another object that surrounds it without being part of the surrounding object.

Some other relationships have been identified that are similar to meronymic:

- Possession is the owner ship relationship.
- Attachment is the relationship in which one entity is attached or joined to another.
- Attribution is the relationship between one entity and its attributes.
- Antonyms is the relationship that indicates the mutual exclusivity between two attributes/entities/relationships.
- Synonyms is the relationship that indicates two attributes/entities/relationships are the same or nearly the same.

Correlation. Correlation is a statistical measure that indicates the extent to which two datasets or variables fluctuate together. A positive correlation indicates the extent to which those datasets/variables increase or decrease in parallel. A negative correlation indicates the extent to which one dataset/variable increases as the other decreases and vice versa. Many times, the correlation is close to 0; this means that there is no obvious relation between the two datasets/variables. It is worthwhile to note that correlation does not imply causation. Correlation is very important in science [19] as it let researchers know if two datasets/variables are related to each other; it is becoming the workhorse of quantitative data analysis. Indeed, correlation analysis is arguably the most important technique that enables the definition of trends and making predictions. Therefore, it is important to be able to measure the correlation between two datasets. The statistical measure that indicates the degree to which two datasets vary together or oppositely is called correlation coefficient. Several algorithms have been proposed for measuring the correlation coefficient. These algorithms will enable the creation of "correlation data patterns", that is, patterns whose links between datasets are constituted by correlations.

Causal relationship. Causal relationship is the relationship between cause and effect. In essence, causality is what connects a dataset with another dataset, where the objects represented in the former are partly responsible for the objects represented in the latter, and the latter is partly dependent on the former. In order to identify a causal relationship between two datasets, first, a variation of the dataset assumed to cause the change in the other dataset must be observed, and then measure the change in the other dataset. Different approaches and systems have been proposed in literature in order to identify causal relationships. These systems should be guided by a causal logic. Such systems will enable the creation of "causal data patterns", that is, patterns whose links between datasets are constituted by causal relationships.

Temporal relationship. Temporal relationship is the relation between two datasets that indicates the ordering in time of these objects represented in the datasets. Temporal information is very important as research datasets are time varying. Examples of temporal relationships include: antecedent-forerunner relationship; synchronicity relationship; asynchrony relationship; sequential relationship. Therefore, modeling temporal information is of paramount importance. There are two mechanisms for including temporal information in a dataset, depending whether a diachronic or a synchronic approach is followed in the representation of data.

- In the diachronic approach, the same dataset contains data collected at different time units, a fact that is reflected by labelling data in the dataset with time information. The labeling mechanism is based on time-stamping. By time-stamping it is intended the addition of a temporal entity t, that labels data in the dataset.
- In the synchronic approach, a dataset is organized in a number of snapshots each containing data collected at the same time unit. Each time the dataset is changed (for example when new data are inserted/deleted) a new snapshot, or version, of the dataset is created and the previous snapshot is stored somewhere.

There are two main dimensions of temporal information: valid and transaction times. Valid time is the time when a dataset is valid; transaction time is the time when a dataset is actually created. The versioning approach is more appropriate for capturing the transaction time, while the labeling approach is used for representing valid time. Adding time in data models and implementing them in temporal DBMSs is an active research area. Temporal graphs, based, for example, on labeling mechanisms, can be created where the arcs are labeled with their interval of validity.

Spatial relationship. A spatial relationship is the relationship between two datasets that the objects represented in the datasets are connected by a topological, or a distance, or a directional relation, amongst the others. A topological relationship describes a relationship between datasets in space. For example, the relationship, between two marine datasets collected in the Aegean and Tyrrhenian seas.

Data Assimilation. Data Assimilation is a set of statistical techniques whereby datasets collected by observations are combined with datasets produced by simulation or from numerical models, to estimate better the evolving state of a complex system such as, for example, the atmosphere. In essence, data assimilation enables to improve the knowledge of the future states of a system by jointly using experimental data and the theoretical (a priori) knowledge on the system. Several methods, including the statistical one, have been proposed in literature and several data assimilation systems have been implemented.

Obviously, the type of data relationships between data sets is domain-specific. Finally, it must be emphasized that also the dataset relationships are dynamic. In fact, both the extensional dynamicity as well as the intentional dynamicity of a dataset could impact on existing relationships between this dataset and other datasets. In essence, an existing relationship between two datasets can disappear or change its type as a consequence of the intentional/extensional dynamicity of one of the datasets involved in the relationship. A new relationship could also be established between this dataset and another dataset. Therefore, upon an intentional/extensional change of a dataset, if it is linked with another dataset by a specific relationship, the validity of this relationship has to be checked.

Data patterns. Discovering inter-dataset relationships, and making them explicit for instance through a linking mechanism, allows the creation of an interconnected data space. In fact, the existence of relationships between datasets enables the establishment of data patterns. By a data pattern we intend a directed graph whose nodes are datasets and whose arcs represent relationships between datasets. A data pattern may be cyclic or acyclic, depending on the relationship represented by the arcs. For instance, causal or mereonymic relationships typically give raise to acyclic data patterns, while intrinsically symmetric relationships, such as proximity or similarity relationships may give raise to cyclic data patterns. These data patterns contain implicit and often previously unknown information, i.e., knowledge. In essence, they constitute knowledge patterns [7]. It could be possible to create data patterns that are characterized by the type of relationship represented by the links between the datasets involved in the patterns.

Exploring the scientific data space

The information exploration (data seeking) in a scientific data space can be carried out in two modes: navigational querying or navigational browsing [20].

- In the navigational querying mode, the data seeking occurs in an intentional way, that is, the researcher has a specific target in mind that is described via a linguistic expression, known as query; the query is submitted to the system that manages a dataset of the data space; by processing the query, the system produces a subset of the queried dataset containing all and only the data of the dataset that satisfy the given description. Successively, the user can refine her/his query, based on the information contained in the subset so far obtained. This refined query can be issued against the same dataset or any other dataset of the data space obtaining, thus, another subset that is more closed to her/his information needs. This mechanism can be iterated until the researcher succeeds to obtain the exact information she/he is looking for. This mode of data exploration is known as "navigational querying".
- In the navigational browsing mode, the researcher is moving through the data space without a clear target in mind. Actually, the researcher is not able to formulate her/his information need as a query, but she/he can recognize relevant information when find it. In the browsing mode the data seeking occurs in an extensional way. The researcher navigates in the data space following different data patterns in the hope that she/he might find datasets that contain relevant information.

In the proposed exploratory strategy, a researcher is enabled to explore the scientific data space in the hope that she/he might find relevant information, i.e., she/he has not a specific target in mind. The exploration activity is performed by navigating in the scientific data space and by dynamically choosing the route of navigation. Here there is an apparent contradiction between the responsibility for selecting routes, implied by the concept of navigation, and the absence of a definite target. Actually, the choice of the route depends on the cognitive state of the researcher that is continuously updated as the exploratory action proceeds. A route of navigation is a data pattern that the researcher dynamically creates by linking different datasets on the basis of relationships that exist among them and that are of interest for her/his research activity. In essence, the researcher starts the exploration by establishing a link between a starting dataset and another dataset. The link materializes a specific relationship that exists between the two datasets. If the information gained by accessing the linked dataset is relevant for the research activity conducted by the researcher, then she/he can iterate this process creating, thus, a data pattern functional for her/his research activity. If the information contained in the linked dataset is not relevant, then the researcher has the option to activate another link between the starting dataset and another dataset on the basis of another type of relationship between these datasets. Again, the researcher can iterate this process creating a different data pattern.

We can call this information exploration mode mediated browsing. In essence, the researcher is not able to formulate queries but she/he is able to formulate request for different types of relationships between datasets. Once a link between two datasets has been established, the researcher can browse the linked dataset for relevant information.

Data analyzers

The automatic discovery of relationships between research datasets is of paramount importance for the successful implementation of an exploratory approach to knowledge production [21]. Therefore, the development of software able to discover data relationships to establish interconnections between datasets must be hastened. A data analyzer should calculate a measure of dependence between variables in pairs of datasets. Most of the data relationships can be modeled as functions, but not all are well modeled by a function. The modeling of data relationships is a domain-specific task and it must be supported by domain-specific vocabularies. Some prototypes have already been implemented [22]. We envision the development, in the near future, of software analyzers specific for each type of relationship. This kind of software will enable the creation of "specialized "data patterns.

Data Analyzers must be adequately described in order to enable potential users to find them. The data analyzers should be described at three distinct levels [23]: the computational, the algorithmic and the implementation levels. At the computational level, the logic of the abstract computational model is described. In essence, at this level, the goal of the computation is described as the identification of a certain type of relationship between variables contained in the schemata of two dataset views. As said in "Research data relationships" section, several types of relationships can exist between these variables. The computational model, in essence, implements an appropriate logic that must guide the discovery of a particular type of relationship sought by a user. Examples of logics, that can be adopted, include conventional, modal, causal, temporal, etc. At the algorithmic level, the representation of certain variables of dataset/view schemata are described (the values of these variables constitute the input to the analyzer) as well as the representation of the output (existence or absence of a certain type of relationship between the input variables). At the implementational level, the data analyzer is described as a software with a discoverable and invocable interface. All these three levels of description are included in the metadata of the analyzer.

As for the datasets/views, also the data analyzers must be published in order to make them discoverable. This means that domain-specific data analyzer catalogues have to be developed. These catalogues should include, at least, for each data analyzer:

- a description that is contained in the metadata;
- an identifier DAI (Data Analyzer Identifier);
- the type of the data analyzer;
- how to request the data analyzer;
- how the data analyzer delivery is fulfilled.

An example of discovering an existing extensional data relationship between two datasets

Let's consider two datasets containing information about marine species: IUCN RedList [24] and Global Record of Stocks and Fisheries (GRSF) [25]. For the sake of the example, we assume that the two datasets are implemented as tables and that fish species are identified in the same way in them. These assumptions will simplify the example and are without loss of generality: in fact, (a) file-based scientific data are typically organized as tables, such as spreadsheets, and (b) common identifiers for species have been specified by standardization bodies and are largely in use.

IUCN RedList provides a wealth of useful information on marine species including their vulnerability status. Far more than a list of species and their status, it is a powerful tool to inform and catalyze action for biodiversity conservation and policy change, critical to protecting the natural resources we need to survive. It provides information about range, population size, habitat and ecology, use and/or trade, threats, and conservation actions that will help inform necessary conservation decisions. Global Record of Stocks and Fisheries provides the fishing areas where certain marine species are fished. Its main purpose is to provide registered users with an environment and tools for accessing stocks and fisheries information.

As already mentioned in "Research data relationships" section, several relationships can be identified between these two datasets. For the purpose of the proposed example, the following relationships between the two datasets are of interest:

- Column containment: because the data contained in a GRSF column is a subset of the data contained in a RedList column.
- Complement: because the unmatched columns of GRSF and RedList provide complementary information about the marine species.

In the literature, algorithms for the identification of the containment relationships between datasets have been proposed and implemented [26]. Once discovered, these relationships can be represented as statements in some data description language (for instance, as data dependencies in SQL or as axioms in OWL). As a case in point, suppose that a researcher has the following information need:

"Find the vulnerability status of the marine species present in a given fishing area X"

This need can be clearly satisfied by appropriately combining the information in the two datasets [27]. However, it can be expressed as a query only if case two conditions are met:

1 species are identified in both tables in the same way, so that the identifiers selected in one dataset can be used to access the other;
2 the researcher performing the query knows exactly the attributes of each table and their semantics.

If these conditions are satisfied, it is not difficult to derive the query Q that expresses the above information need and obtain the sought answer by executing Q.

However, it is well-known that the above situation is very hard to find in reality, since each dataset is the product of an investment made by a community with specific objectives, and this condition results in a great heterogeneity of the data space.

Under these circumstances, we can safely assume only that each dataset provides an API conforming to some standard, so that it is possible to develop a specific procedure that takes the name of the fishing area (X) as the input and interacts with the dataset APIs' as follows:

1 the procedure processor extracts the set I of the identifiers of the marine species fished in X by executing a specific query on the GRSF API.
2 from the containment relationship between the two datasets (GRSF, RedList), it infers that these species are a subset of species listed in RedList
3 for each species s in I, the procedure uses the RedList API to extract the vulnerability status vs of s. Each pair (s, vs) is part of the answer to the information need above.

This procedure can be standardized as a SPARQL query over an OWL 2 DL ontology by modeling each dataset as a class, endowed with properties that reflect the structure of the dataset. In this case, the semantic containment relationship between the two datasets can be captured as a subclass axiom given by (in the functional OWL 2 DL [28] notation):

SubClassOf(GRSF, RedList).

By virtue of this axiom the superclass, RedList, inherits all species of the subclass, GRSF. It then suffices to query RedList to obtain the desired information. In essence, a so enriched query processor is able to exploit an implicit data relationship represented as a logical axiom and answer a query that depends on this relationship.

Concluding remarks

In this paper, we have outlined a new approach to the knowledge creation based on the exploitation of the knowledge hidden in huge data volumes of research data. This data is the outcome of several domain-specific research activities carried out by scientists. The proposed approach is framed within a scientific context that has been revolutionized during the last years. The main characteristics of this new scientific context are: (i) the widespread digitization of the research results; (ii) the production of big research data; (iii) an increasingly data intensive science; (iv) an increasingly multidisciplinary science; (v) an increasingly e-science; and (vi) an increasingly open science. All these characteristics have motivated our effort for a new approach to knowledge creation based on the exploration of the scientific information space (data patterns driven). Realizing this approach implies the implementation of data infrastructures and the development of tools for the automatic discovery of hidden data relationships. The data infrastructures should provide: (i) linking services to allow the creation of linked information spaces; (ii) intermediary services to make the holdings of data centers, digital libraries, institutional repositories, etc. discoverable, accessible, understandable and reusable; (iii) navigational services to allow researchers to navigate the linked scientific information space; and (iv) workflow services to draw patterns of interest within the linked scientific information space. Concerning the tools for the automatic discovery of data relationships, we have already discussed in "Data analyzers" section the need for specialized data analysis software. We envision that in the near future these pre-conditions will be fully implemented enabling, thus, an exploratory approach to the knowledge creation.

Abbreviations

IT	Information Technology
EDA	Exploratory Data Analysis
KDD	Knowledge Discovery in Databases
LOD	Linked Open Data
IRI	Internationalized Resource Identifier
DOI	Digital Object Identifier
NSB	National Science Board
GRSF	Global Record of Stocks and Fisheries
DAI	Data Analyzer Identifier

Acknowledgements
Not applicable.

Author contributions
C.T. conceived the main ideas behind the manuscript. All authors contributed equally to writing and reviewing the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Availibility of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 15 April 2022 Accepted: 23 February 2023
Published online: 06 March 2023

References

1. Hey T, Tansley S, Tolle KM. Jim Gray on eScience: a transformed scientific method. The Fourth Paradigm; 2009.
2. Bernstein PA, Haas LM. Information integration in the enterprise. Commun ACM. 2008;51(9):72–9.
3. Halevy A, Franklin M, Maier D. Principles of dataspace systems. In: Proceedings of the Twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2006:1–9.
4. Idreos S. Big data exploration. Big Data Computing. Taylor and Francis; 2013:3.
5. Yu CH. Exploratory data analysis in the context of data mining and resampling. In J Psychol Res. 2010;3(1):9–22.
6. Fayyad UM, Piatetsky-Shapiro G, Smyth P. *et al.* Knowledge discovery and data mining: Towards a unifying framework. In: KDD. 1996;96:82–88
7. Gullo F. From patterns in data to knowledge discovery: What data mining can do. Phys Procedia. 2015;62:18–22.
8. Auer S, Bryl V, Tramp S. Linked Open Data-Creating Knowledge Out of Interlinked Data: results of the LOD2 Project. Cham: Springer; 2014.
9. Simberloff D, Barish B, Droegemeier K, Etter D, Fedoroff N, Ford K, Lanzerotti L, Leshner A, Lubchenco J, Rossmann M. *et al.* Long-lived digital data collections: enabling research and education in the 21st century. National Science Foundation N/A. 2005.
10. Wynholds L. Linking to scientific data: identity problems of unruly and poorly bounded digital objects. digital curation conference, chicago. INTERNATIONAL JOURNAL OF DIGITAL CURATION, 2011. 6:214–225.
11. Renear AH, Sacchi S, Wickett KM. Definitions of dataset in the scientific and technical literature. Proc Am Soc Inf Sci Techno. 2010;47(1):1–4.
12. Farnel S, Shiri A. Metadata for research data: current practices and trends. In: International Conference on Dublin Core and Metadata Applications, 2014:74–82.
13. Willis C, Greenberg J, White H. Analysis and synthesis of metadata goals for scientific data. J Am Soc Inf Sci Technol. 2012;63(8):1505–20.
14. Floridi L, Sanders JW. Levellism and the method of abstraction. In: IEG (ed.) IEG Research Report, 2004.
15. Buneman P, Davidson S, Frew J. Why data citation is a computational problem. Commun ACM. 2016;59(9):50–7.
16. Paskin N. Digital object identifiers for scientific data. Data sci J. 2005;4:12–20.
17. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Santos LBDS, Bourne PE, et al. Addendum: The fair guiding principles for scientific data management and stewardship. Scientific data. 2019;6:6.
18. Storey VC. Understanding semantic relationships. VLDB J. 1993;2(4):455–88.
19. Malik F. Understanding Value of Correlations in Data Science Projects. 2019.
20. Waterworth JA, Chignell MH. A model for information exploration. Hypermedia. 1991;3(1):35–58.
21. Alawini A. Identifying relationships between scientific datasets. PhD thesis, Portland State University. 2016.
22. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011;334(6062):1518–24.
23. Thanos C, Klan F, Kritikos K, Candela L. White Paper on Research Data Service Discoverability. Publications. 2016;5(1):1.
24. International Union for Conservation of Nature, et al. IUCN Red List categories and criteria. IUCN; 2001.
25. Marketakis Y, et al. On the evolution of semantic warehouses: the case of global record of stocks and fisheries. In: Metadata and Semantic Research: 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020, Revised Selected Papers 14. Springer International Publishing, 2021. p. 269–281.
26. Alawini A, Maier D, Tufte K, Howe B. Helping scientists reconnect their datasets. In: Proceedings of the 26th International Conference on Scientific and Statistical Database Management. 2014. 1-12.
27. Coro G, Ellenbroek A, Pagano P. An open science approach to infer fishing activity pressure on stocks and biodiversity from vessel tracking data. Ecol Inform. 2021;64:101384.
28. World Wide Web Consortium: OWL 2 web ontology language document overview. 2020.

Publisher's Note