publications

MDPI

Article

Research Data Reusability: Conceptual Foundations, Barriers and Enabling Technologies

Costantino Thanos

Institute of Information Science and Technologies, National Research Council of Italy, 56124 Pisa, Italy; Costantino.Thanos@isti.cnr.it

Academic Editor: Isabel Bernal
Received: 17 October 2016; Accepted: 4 January 2017; Published: 9 January 2017

Abstract: High-throughput scientific instruments are generating massive amounts of data. Today, one of the main challenges faced by researchers is to make the best use of the world's growing wealth of data. Data (re)usability is becoming a distinct characteristic of modern scientific practice. By data (re)usability, we mean the ease of using data for legitimate scientific research by one or more communities of research (consumer communities) that is produced by other communities of research (producer communities). Data (re)usability allows the reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others. It has four main dimensions: policy, legal, economic and technological. The paper addresses the technological dimension of data reusability. The conceptual foundations of data reuse as well as the barriers that hamper data reuse are presented and discussed. The data publication process is proposed as a bridge between the data author and user and the relevant technologies enabling this process are presented.

Keywords: data reuse; data discoverability; data understandability; relational thinking; data abstraction; data representation; metadata; explicit knowledge; tacit knowledge; data publishing

1. Introduction

New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks and running simulations are generating massive amounts of scientific data. Often referred to as a data deluge, massive datasets are revolutionizing the way research is carried out, which results in the emergence of a new fourth paradigm of science based on data-intensive computing [1]. This data-dominated science will lead to a data-centric way of thinking, organizing and conducting research activities that could lead to new approaches to solve problems that were previously considered extremely hard or, in some cases, even impossible to solve and also lead to serendipitous discoveries [2]. Today, one of the main challenges faced by researchers is to make the best use of the world's growing wealth of data.

By data (re)usability, we mean the ease of using data for legitimate scientific research by one or more communities of research (consumer communities) that is produced by other communities of research (producer communities). We use the term data reusability to mean the ease of use of data collected for one purpose to study a new problem [3]. This term denotes the reutilization of existing datasets in significantly different contexts. Data reusability is becoming a distinct characteristic of modern scientific practice, as it allows the reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others.

Data (re)usability can be effectively implemented in the Open Science framework, as the ultimate goal of the Open Science is to make research data publicly available and (re)usable. The European Commission is moving decisively towards the implementation of an Open Science framework in Europe: In 2012, the European Commission encouraged all European Union EU Member States to

put public-funded research results in the public sphere in order to make science better and strengthen their knowledge-based economy, via a *Recommendation* [4]. A recent document *The Amsterdam Call for Action on Open Science* advocates "full open access for all scientific publications" and endorses an environment where data sharing and stewardship is the default approach for all publicly funded research. This document was produced at an Open Science meeting organized by the Dutch Presidency of the Council of the European Union (4–5 April 2016) [5].

Another initiative of the European Commission that is worthwhile to mention is the publication of *Guidelines on FAIR Data Management in Horizon 2020*, that is, a set of guiding principles to make data Findable, Accessible, Interoperable and Reusable [6].

Data reusability has four main dimensions: policy, legal, economic and technological. A legal and policy framework should favor the open availability of scientific data and allow legal jurisdictional boundaries to be overcome; the economics concern how the costs associated with the process of making scientific data reusable are distributed among the stakeholders; and technology should render physical and semantic barriers irrelevant. In this paper, we will concentrate on the technological dimension of data reusability.

The paper is organized in the following way: Section 2 describes the research-data universe composed of different types of research data, different kinds of data collections, of many data actors and different data uses. Section 3 introduces the conceptual foundations of data reusability, i.e., relational thinking, knowledge boundaries, data abstraction/levellism and representation. Section 4 discusses the barriers that hamper data reuse. In Section 5, the data publication process, that spans the distance between the data author and the data user, is described. In Section 6, the technologies that enable this process are briefly described. Section 7 stresses the important role of standards in making data usable. Finally, Section 8 summarizes the main points to be taken into consideration when addressing the pressing need to reuse large datasets produced by the research communities.

2. The Research Data Universe

The research data universe is complex, involving many actors using many types of data for many different scientific purposes. Recent years have witnessed the rise of a multitude of data collections that are robust and flexible, while allowing for heterogeneous data types and associated metadata developed to satisfy the wide range of requirements of diverse research communities.

Research Data

By research data, we mean scientific or technical measurements, values calculated, and observations or facts that can be represented by numbers, tables, graphs, models, text, or symbols and that are used as a basis for reasoning or further calculation [7]. Such data may be generated by various means, including observation, computation, or experimentation. Scientists regard data as accurate representations of the physical world and as evidence to support claims [8].

Data can be distinguished by their **origins**—whether they are **observational**, **computational**, or **experimental**.

Observational data are collected by direct observations and a particular feature of these data is that they cannot be recollected.
Computational data are produced by executing a computer model or simulation; their feature is that they can be reproduced.
Experimental data are collected by conducting experiments; in principle, data from experiments can be accurately reproduced. In practice, however, it may not be possible to reproduce precisely all of the experimental conditions.

Data can be referred to as **raw**, **derivative**, or **verified** [8].

Raw data consist of original observations, such as those collected by satellite and beamed back to earth or generated by an instrument or sensor or collected by conducting an experiment.

Derivative data are generated by processing activities. The raw data are frequently subject to subsequent stages of refinement and analysis, depending on the research objectives. There may be a succession of versions. While the raw data may be the most complete form, derivative data may be more readily usable by others as processing usually makes data more usable, ordered or simplified, thus increasing their intelligibility.
Verified data are generated by curatorial activities. Their quality and accuracy have, thus, been assured.

Data Collections/Databases

Scientific data are stored into managed data collections/databases. Data collections fall into one of three functional categories as reported in [9]:

Research Data Collections are the products of one or more focused research projects and typically contain data that are subject to limited processing or curation. They may or may not conform to community standards, such as standards for file formats, metadata structure, and content access policies.
Resource or Community Data Collections serve a single science or engineering community. These digital collections often establish community-level standards either by selecting from among preexisting standards or by bringing the community together to develop new standards where they are absent or inadequate.
Reference Data Collections are intended to serve large segments of the scientific and educational community. Characteristic features of this category of digital collections are the broad scope and diverse set of user communities including scientists, students, and educators from a wide variety of disciplinary, institutional, and geographical settings. In these circumstances, conformance to robust, well-established, and comprehensive standards is essential, and the selection of standards by reference collections often has the effect of creating a universal standard.

Data Actors

The main actors in the scientific data universe are [9]:

Data Authors are individuals or teams involved in research activities that generate digital data that are subsequently deposited in a data collection. Their interests lie in ensuring that they enjoy the benefits of their own work, including gaining appropriate credit and recognition, and that their results can be broadly disseminated and safely archived.
Data Users are representatives of the scientific communities. Their interests lie in having ready access to data sets that are discoverable and intelligible, i.e., well defined and well documented.
Data Managers are individuals responsible for the operation and maintenance of the data collections/databases.
Data Scientists are information and computer scientists developing innovative concepts in database technology and information sciences, including scientific data modeling, data discovery, data visualization, etc., and applying these to the fields of science relevant to the data collection/database.

Data Uses

Data are used in different ways according to their contexts. Two broad categories of data use can be defined:

End Use is defined as the ability of accessing a dataset to verify some fact or perform some job-related or personal task.
Derivative Use is defined as the ability of building on a preexisting dataset by extracting information from one or more datasets in order to create a new dataset that can be used for the same, similar, or an entirely different purpose with respect to the original dataset(s).

Diversity in the Research Data Universe

In conclusion, we can affirm that research data exist in many different types and formats subject to varying legal, cultural, protective, and practical constraints. Data authors, managers, and users often come from different disciplinary, professional, cultural, and other settings with different needs, expectations, responsibilities, authorities, and expertise. These experts are subject to varying legal, physical, scientific, cultural, and other constraints.

The diversity in data, individuals, disciplines, contexts, and cultures is the big challenge faced by researchers in order to harness the accumulating data and knowledge produced by the research communities and make them reusable.

3. The Conceptual Foundations of Data Reusability

In this section, we introduce a conceptual framework within which to address the data (re)usability problem from a more theoretical point of view. We have borrowed some concepts from other theoretical fields, i.e., relational thinking, levellism and knowledge sharing in order to create this conceptual framework. We have identified these three fields because data (re)usability implies a relationship between the data author and data user; it requires that data must be made available at different levels of abstraction and representation, and it also requires that data be semantically enriched in order to be able to cross semantic barriers without semantic distortions.

3.1. Relational Thinking

The definition of data usability assumes that the two entities, data author and data user, are neatly separated from one another and considers the properties attached to these entities as independent of the relationships with which they interact and exist. Therefore, it tends to reify the attributes of these entities by detaching them from their scientific context. This often takes place, as substantive attributes are easier to identify or more convenient to count and so are assumed to be more concrete or "real" than relational attributes. However, such a substantialist approach is not appropriate to address data reusability issues. We think that an approach based on relational thinking is more appropriate. By relational thinking, we mean a loosely structured framework or scaffold around which various practice theories and methods are being developed (10).

In **relational thinking** found in practice theory, subjects, social groups, networks, or even artifacts develop their properties only in relation to other subjects, social groups, or networks. Social objects derive their significance from the relations that link them, rather than from the intrinsic features of individual elements.

A dataset cannot be understood and used in and of itself (isolation), and cannot be transferred from one scientific context to another without changes to its properties. Relational thinking entails that a dataset produced by a community of practice in order to be used by another community of practice must be endowed with properties (auxiliary information) that take into consideration the characteristics of the "**usability relation**" between the two communities.

Several kinds of usability relations can be established between two communities of practice. For example, a "**confirmation relation**" is established when the consumer community tries to find a confirmation of some scientific expectation by gathering enough evidence from a data set produced by the producer community. Another kind of usability relation is the "**reproduction/verification relation**" that is established when the consumer community tries to reproduce/verify a scientific result by using a data set produced by the producer community. One more kind of usability relation is the "**discovery relation**" that is established when the consumer community tries to discover new insights from a data set produced by the producer community.

Therefore, a community of practice that produces a data set, in order to make it (re)usable by another community of practice, must complement it with appropriate metadata information. The properties of the metadata information (provenance, context, quality, uncertainty, etc.) heavily

depend on the "usability relation" established between the producer and consumer communities of practice.

Thus, if a dataset is to be used by different communities of practice, different metadata information must be provided to these diverse communities of practice depending on the characteristics of the "usability relations" that link the producer community of practice with them. For example, for one consumer community, it could be enough to know who, where and when a given dataset was produced; for another, it could be important to know how this dataset was produced.

As a consequence, a data producer community of practice must define metadata models based on the usability relations established between this community and the communities that consume the data produced by it.

Relational thinking makes it possible to choose and organize the metadata information so as to overcome the semantic and pragmatic boundaries between communities of research and, thus, increase the understandability and reusability of the data.

In order to apply the relational thinking approach to improving data reusability, we have to characterize the "usability relation" between data producer and data consumer communities. In particular, we have to consider [10]:

(a) **Differences characterizing the relation**. A first characterization entails delineating what are the differences associated with a "*usability*" relation between the data author and data user. It would be important, for example, to be able to characterize the differences in the knowledge and perspectives of a data author and a data user when working in the context of a multidisciplinary/interdisciplinary collaborative research activity.

(b) **Dependencies characterizing the relation**. A second characterization entails delineating what are the dependencies associated with a "*usability*" relation between the data author and data consumer. The knowledge developed by the data producer is not inconsequential to the data consumer but develops in dependency of the perspectives promoted by the data consumer. It would be important to be able to delineate the dependencies that characterize the "usability" relation between data producer and data consumer.

(c) **Changes characterizing the relation**. Differences and dependencies characterizing a "usability" relation between data producer and data consumer change over time. We must assume that the "usability" relation undergoes continuous refinement and/or revision through interactions between the two entities (data author, data user).

The above characterizations should be taken into consideration when defining metadata models. They should guide the definition of good quality metadata models (purpose-oriented, community-specific) that can increase data reusability.

As the quality of metadata is probably the most important factor that determines the reusability of data, we can affirm that the relational thinking approach is decisive in achieving a good level of data reusability.

3.2. Knowledge Boundaries

In all the research activities, experimental, observational, or computational together with the production of scientific data, a rich body of knowledge is also created. This knowledge can be of two types: explicit knowledge and tacit knowledge.

In order to make the scientific data effectively reusable, the underpinning explicit and tacit knowledge also has to be made reusable. The notion of knowledge reuse refers to the concepts of transferring and reutilizing existing knowledge bases in significantly different contexts. Ideally, it would be desirable to be able to handle these two types of knowledge as a commodity that can be extracted, represented and packaged within a given context (data producer context) and transferred and easily inserted in another context (data user context) [11].

This means that both types of knowledge should be part of the data publishing process, that is, the process through which scientific data are made sharable and usable.

The difficulty in making knowledge reusable consists in the fact that what is codified in one discipline may not be understood to those in other fields because of the intellectual content and amount of background needed.

3.2.1. Explicit Knowledge Reuse

By explicit knowledge, we mean knowledge that can be encoded in some language and exchanged between distributed research teams.

Building a knowledge base also implies to endow it with a number of components that help to generate knowledge from the knowledge base. Part of the complexity of reusing knowledge stems from the multiple components of knowledge that should be reused. In fact, making a knowledge base reusable implies that these components should also be reusable. Among these components, we identify two that are particularly important [11]:

Reusable Lexicons: In building a knowledge base, an important step is the establishment of the domain of discourse. It consists of identifying the objects in the world about which an inference engine will reason and the set of linguistic terms, which have a precise and invariant meaning, by which both the engine and the users will refer to those objects. A lexicon is reusable if it contains a set of reusable terms. By reusable terms, we mean that an equivalence can be established among these terms and the terms of other different lexicons.

Reusable Ontologies: In many cases, it is important to share more than a common vocabulary; it is required to specify also the relationships among the objects in the world to which the term refers, to understand how classes of objects can be defined and what are the rules that allow the assignment of individual objects to particular classes. In essence, it is necessary to create ontologies. A domain-specific ontology is reusable if it can be aligned with other domain-specific ontologies (see Section 6.2).

3.2.2. Tacit Knowledge

By tacit knowledge, we mean knowledge that is confined within specific practices and interpersonal exchanges and bound up with a set of communications, tools, etc. The main characteristic of this type of knowledge is its embeddedness [12]. This characteristic of tacit knowledge makes its codification in some language very difficult.

In order to make tacit knowledge reusable, we must transform it into a "mobile knowledge", that is, a knowledge that can be codified in some language and easily transported or translated from one working context to another one.

Unfortunately, there are several difficult problems that hinder the knowledge transformation from tacit into mobile.

The main conceptual problem is how to transform knowledge that is embedded within highly specific scientific domains into mobile knowledge that can cross several scientific domains. The literature in many scientific fields addresses the tension between rich knowledge that is embedded in interpersonal contexts, and the need to make knowledge mobile when it must be shared and reused by distributed teams of researchers. Factors that can influence the effectiveness and efficiency of the knowledge transformation from tacit to mobile include [13]:

(i) the characteristics of the knowledge,
(ii) the functionality of the data and communications infrastructures that support the data publishing process and the mobilization of knowledge, and
(iii) the characteristics of the working contexts involved in a distributed collaborative effort (data producer context, data consumer context).

The mobile knowledge derived from tacit knowledge is highly contextualized. Therefore, in order to make the mobile knowledge shareable among different teams, it is essential to create an interpretative context shared by all the actors involved in collaborative efforts.

As a conceptual framework within which knowledge can be embodied, mobilized and shared has been proposed, the concept of "boundary object" [13] is relevant. Boundary objects are those objects that both inhabit several communities of practice and satisfy the informational requirements of each of them. Boundary objects are thus both plastic enough to adapt to local needs and constraints of several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use and become strongly structured in individual-site use. These objects may be abstract or concrete. Such objects have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation.

Boundary objects could play a key role in the successful translation of knowledge between different communities. Unfortunately, boundary objects are not well understood or easily identified, so their use as a translation tool is not widely implemented.

3.3. Data Abstraction and Data Representation

Research data provide an account of the results of a scientific work and therefore they must be intelligible to those wishing to understand or scrutinize them. Therefore, data communication must be differentiated for different categories of audiences with different scientific and cognitive backgrounds. This means that effective data communication should enable recipients to scrutinize a dataset at a level of abstraction and at a level of representation that are more appropriate for their scientific background and research interests.

3.3.1. Data Abstraction

There are two main varieties of abstraction: *ontological* and *epistemological* [14]. The ontological approach to abstraction, is concerned with the different levels of organization of a system that can be identified and defined. For example, a database can have conceptual, semantic, syntactic and physical levels of organization.

The epistemological approach to abstraction is concerned with the different levels of observation or interpretation at which a system can be studied. For example, a database can be observed and analyzed at different levels of abstraction, consisting of data related by time, place, instrument, or object of observation. Examples of epistemological levels of abstraction are spatial and temporal data abstractions.

Basic Concepts: In order to be able to describe the "method of abstraction", three concepts are introduced [14]:
Typed Variable: A "typed variable" is a uniquely-named conceptual entity (the variable) and a set (called its type), consisting of all the values that the entity may take. Two typed variables are regarded as equal if and only if their variables have the same name and their types are equal as sets. A variable that cannot be assigned well-defined values is said to constitute an ill-typed variable.

The degree to which a type is appropriate depends on its context and use.

Observable: An "observable" is an interpreted typed variable, that is, a typed variable together with a statement of which feature of a scientific data collection (for example, spatial, temporal, graphical, visual) under consideration it represents.

The definition of an observable reflects a "particular view" or "attitude" towards the data collection being studied. Most commonly, it corresponds to a simplification.

Level of Abstraction: A level of abstraction (LoA) is a finite but non-empty set of observables. No order is assigned to the observables. Different LoAs may be appropriate for different purposes. The definition of observables is the first step in studying a data collection at a given LoA. The second step consists in deciding what relationships are held between the observables.

The Method of Abstraction

As scientific databases should be studied at different levels of abstraction, a method for specifying these different levels of abstraction must be defined [14].

In order to be able to specify a level of abstraction, first, the range of queries which can be meaningfully asked by the target audience, and that are answerable in principle, must be identified. The input of a level of abstraction consists of the scientific database under analysis; its output is an abstract view of the database. The type and amount of data vary with the level of abstraction: a lower level of abstraction produces a view that contains more data than a view produced at a higher, or more abstract level. Therefore, type and quantity of data that must be taken into consideration when specifying a level of abstraction is predetermined by the choice of this level. In essence, a given level of abstraction provides a quantified commitment to the kind and amount of data that can be extracted from a scientific database. The observables at a given level of data abstraction can be obtained as a result of a query issued against a database.

The method of abstraction is ideally suited to the study and analysis of large and complex databases derived from experiments and from upcoming petascale and exascale simulation systems. They are best understood stepwise, that is, by their gradual disclosure at increasingly fine levels of abstraction.

Several data abstraction approaches are currently used by data scientists in order to improve data accessibility and understandability; among them, we list the most relevant:

Metadata. An ontological data abstraction level that is of paramount importance in the domain of scientific data is the metadata abstraction level. This abstraction level captures the information content of the underlying data independent of representational details. Metadata descriptions enable representation of domain knowledge describing the information domain to which the underlying data belong.

Data Virtualization. An important ontological data abstraction level is data virtualization; it hides all the technical aspects of data storage; the data users do not have to know where all the data have been stored physically, where the database servers run, what the source Application Programming Interface API and database language is, and so on.

Data Clustering. An epistemological data abstraction approach is data clustering. It allows the grouping of the data into clusters; the data contained in a cluster are similar to each other while data belonging to different clusters are dissimilar.

In conclusion, we argue that (i) epistemological abstraction should be retained as a proper abstraction method for increasing data reusability as it supports the definition of several levels of explanation and interpretation of a scientific database; and (ii) ontological abstraction should be retained as a proper abstraction method for increasing data accessibility and understandability as it supports the definition of several levels of organization of a scientific database.

Finally, we argue that the right level of abstraction to be communicated to a given data consumer community should be based on the "usability" relation established with the data producer community.

3.3.2. Data Representation

Appropriate data representation is essential for enabling scientists to correctly interpret data and use them appropriately as the same information content can be represented differently in different data description languages. A major problem is that we have no shared formal conceptual model of data representation that is both accurate and sufficiently detailed to support the data needs of scientists

belonging to different scientific disciplines [15]. The traditional relational data model is not adequate to represent the data needs of most of the scientific disciplines [16]. For some scientific disciplines (astronomy, oceanography, fusion, and remote sensing), an array data model is more appropriate. Some other disciplines, i.e., biology and genomics, consider graphs and sequences more appropriate for their needs. Lastly, solid modelling applications want a mesh data model. In the big data era, pictorial representation of data is of paramount importance. It makes the presentation of data more intelligible, and allows investigators to easily see the salient features of the data, and bring out the hidden pattern and trends of the complex datasets. Two of the main approaches to the pictorial representation of data are shown below:

Visual Representation of Data. Effective data visualization improves interpretation of data and helps scientists in analyzing and reasoning about data and evidence. Visual data analysis enables the detection and validation of expected results while also enabling unexpected discoveries in science. Data visualization makes complex data more accessible, understandable and usable. In the big data era, data visualization is an indispensable technique for extracting meaning from large and complex scientific datasets.
Graphical Representation of Data. Graphical methods are also well suited for digesting great amounts of data. Investigators can have a better look at the information collected and the distribution of data. The graphic method of the representation of data enhances our understanding, makes the comparisons easy and creates an imprint on the mind for a longer time.

4. Barriers That Hamper Data Reuse

Despite the importance, it is not easy to reuse data. There are several obstacles. We have identified five main obstacles:

Heterogeneity of Representations

There are four critical impediments to data reuse due to the heterogeneity of representations [17].

Heterogeneous Data Representations: there is a wide variety of scientific data models and formats and scientific information expressed in one formalism cannot directly be incorporated into another formalism.
Heterogeneity of Query Languages: Data collections are managed by a variety of systems that support different query languages. It is difficult to share data if they are encoded in different dialects.
Lack of Communication Conventions: Data reuse does not necessarily require a shared database. If separate systems can communicate with one another, they can benefit from each other's database without sharing a common database. Unfortunately, this approach is not generally feasible for today's scientific database systems and file repositories as these systems are not based on formal data models and thus, making them interoperable is very difficult. We lack an agreed-on protocol specifying how these systems are to query each other and in what form answers are to be delivered. Similarly, we lack standard protocols that would provide interoperability between research data infrastructures.
Vocabulary Mismatching: another barrier to data reuse is when a common vocabulary and domain terminology is lacking.

Discovering Data

Researchers must be aware of who has the data they need or where the data are located. In a networked scientific multidisciplinary environment, pinpointing the location of relevant data is a big challenge for researchers. A data discovering capability requires the support of appropriate metadata descriptions and registries, data classification/categorization schemes, as well as definitions of researcher profiles and goals.

In addition, after finding appropriate data, researchers must often negotiate with the owner or develop trusting relationships to gain access.

Understanding Data

Once in possession of a data set, the next problem regards the capacity of the data user to understand the information/knowledge embodied in it. Data understandability must be built on a fundamental premise: a data set is intelligible only when its metadata relates to its intended use. An additional difficulty arises when providing the same data set for different user communities. In this case, appropriate abstractions of the data set must be created for the different communities.

To make data understandable, they must be endowed with auxiliary information, including metadata, community-specific ontologies or taxonomies, and terminologies.

However, much of the knowledge needed to make sense of a data set is tacit. Scientists are not necessarily able to explicate all of the information that is required to understand someone else's work.

Moving Data

In the scientific data universe, actors and data collections inhabit multiple contexts. There is the risk, when data are moving across contexts, of interpreting data representations in different ways caused by the loss of the interpretative context. This can lead to a phenomenon called "ontological drift" as the intended meaning becomes distorted as the data move across semantic boundaries (semantic distortion). This risk arises when a shared vocabulary and domain terminology are lacking.

Data Mismatching

There are several data mismatching problems that hamper data reusability:

Quality mismatching occurs when the quality profile associated with a data set does not meet the quality expectations of the user of this data set.

Data-incomplete mismatching occurs when a data set is lacking some useful information (for example, provenance, contextual, uncertainty information) to enable a data user to fully exploit it.

Data abstraction mismatching occurs when the level of data abstraction (spatial, temporal, graphical, etc.) created by a data author does not meet the expected level of abstraction by the data user.

5. Data Publishing: A Process for Bridging the Gap between Data Author and Data User

An emerging approach in the scientific communication is Data Publication. By Data Publication, we mean a process that allows the research community to discover, understand, and make assertions about the trustworthiness and fitness for purpose of the data. In addition, it should allow for those who create data, to receive academic credit for their work [18,19]. The ultimate aim of Data Publication is to make scientific data available for reuse both within the original disciplines and the wider community. Many of the issues regarding data availability and usability can be addressed if the principles of *publication* rather than *sharing* are applied [20]. The Data publication approach imitates the scholarly literature publication and generally emerges from the culture of academic research and scholarly communication [21].

The Data Publication process should perform the following main functions:

- **Data Peer-Reviewing**
  The purpose of peer review is to ensure a certain level of quality assurance. In fact, a peer-reviewed dataset can be considered to have been through a process of scientific quality assurance.
- **Data Registration**
  The purpose of registration is to make data citable as a unique piece of work and to allow claims of precedence of a scholarly finding. Data registration should facilitate data discoverability.
- **Data Semantic Enrichment**
  The purpose of semantic enrichment is to make data understandable. The published data should be endowed with appropriate discipline-specific metadata information. The metadata information improves data understandability.

- **Data Archiving**
  The purpose of archiving is to preserve data over time.
- **Awareness**
  Publishing data allows scholars to remain aware of new claims and findings.
- **Rewarding**
  The purpose of rewarding is to bring scholarly credit to the data authors.

In conclusion, we can say that Data Publication is a process that guarantees the "*right to know*" of scholars and the "*right to be known*" of the data authors [22].

Data publication is enabled by **data curation** and **data stewardship** that are two fields of practice of paramount importance for the data (re)usability.

Data curation is the active and on-going management of data through its entire lifecycle of interest and usefulness to scholarship [23]. Data curation activities maintain data quality, add value, and provide for reuse over time and also include authentication, archiving, management, preservation, and representation. Curation, in essence, is concerned with availability and future use of data, including the enhancement, extension, and improvement of data for reuse beyond a single scientific community.

Data stewardship is concerned with the management of shared data collections. It is essential to their preservation and persistence. Stewardship is the process of overseeing and enforcing these activities in accordance with policies defined by data collections' owners. The stewardship function is often primarily an administrative workflow [24].

The focus of data curation is on the "*interest and usefulness*" of data to scholarship; in essence, it addresses the data quality criterion of relevance, while data stewardship is mainly concerned with data *trustworthiness* [24].

Both data curation and stewardship address the critical function of helping users take confidence in data usability based on various criteria of its quality, and thus, are instrumental to the data publication. In fact, they constitute the two pillars that bear data publication.

An instrument that effectively supports the Data Publication process and therefore the data reusability is the Data Management Plan (DMP). A DMP is a formal document that states what data will be created and how, and outlines the plans for sharing and preservation. In addition, it also states any restrictions that may need to be applied on the collected data. All the data organizations that maintain data collections as well as the research projects that create data collections must be endowed with a DMP.

6. Enabling Technologies

There are several technologies that can be employed to effectively implement the Data Publication procedures and, thus, to overcome the impediments to data reuse. Some of them enable data discoverability, some others data understandability, and others make data assessable. Altogether, these technologies contribute to make scientific data reusable. We have identified seven main enabling technologies: modeling metadata information, building and aligning domain-specific ontologies, discovering data, enabling data exchangeability, linking data to publications, linking open data, and applying standards.

6.1. (Meta) Data Modeling

In order to facilitate data understandability, it is necessary to define and develop formal models that adequately describe:

- data representation needs of a given scientific discipline;
- data provenance information;
- data contextual information;
- data uncertainty;

- data quality information.

All this information is collectively called metadata information. If scientists are to reuse data collected by others, then the data must be carefully documented. Metadata is the descriptive information about data that explains the measured attributes, their names, units, precision, accuracy, data layout and ideally a great deal more. Most importantly, metadata should include the data lineage, i.e., how the data was measured, acquired, or computed. The use of purpose-oriented metadata models is of paramount importance to achieve data reusability. Data is incomprehensible and hence useless unless there is a detailed and clear description of how and when it was gathered, and how the derived data was produced [25]. The type of descriptive information to be provided by the data author depends very much on the usability relations established between the data authors and users.

Data Provenance Modeling: In its most general form, provenance (also sometimes called lineage) captures where data came from, and how it has been updated over time. Provenance can serve a number of important functions [26]: explanation, verification, re-computation and repeatability. In the long-term, a standard open representation and query model is needed. A promising example is the "Open Provenance Model" [27], a community-driven model, which allows provenance to be exchanged between systems.

Data Context Modeling: Context is a poorly used source of information in our computing environments. As a result, we have an impoverished understanding of what context is and how it can be used.

Contextual information is any information that can be used to characterize the situation of a digital information object. In essence, this information documents the relationship of the data to its environment. *Context* is the set of all contextual information that can be used to characterize the situation of a digital information object.

Several context modelling approaches exist and are classified by the scheme of data structures which are used to exchange contextual information in the respective system [28]: Key-value Models, Mark-up Scheme Models, Object Oriented Models, Logic Based Models, and Ontology Based Models.

Data Uncertainty Modeling: As models of the real world, scientific datasets are often permeated with forms of uncertainty. Uncertainty is the quantitative estimation of error; all measurements contain some uncertainty generated through systematic error and/or random error. Acknowledging the uncertainty of data is an important component of reporting the results of scientific investigation.

There has been a significant amount of work in areas variously known as "*uncertain, probabilistic, fuzzy, approximate, incomplete and imprecise*" data management.

Unfortunately, current data management products do not support uncertainty [16]. Undoubtedly, the development of suitable database theory to deal with uncertain database information remains a challenge that has yet to be met.

Data Quality Modeling: The quality of data is a complex concept, difficult to define. There is no common or agreed upon definition or measure for data quality, apart from such a general notion as *fitness for use*.

The consequences of poor data quality are often experienced in all scientific disciplines, but without making the necessary connections to its causes [29]. Awareness of the importance of improving the quality of data is increasing in all scientific fields.

In order to fully understand the concept, researchers have traditionally identified a number of specific quality *dimensions*. A dimension or characteristic captures a specific facet of quality. The more commonly referenced dimensions include *accuracy, completeness, consistency, currency, timeliness and volatility.*

For specific categories of data and for specific scientific disciplines, it may be appropriate to have specific sets of dimensions.

Metadata is as valuable as the data itself [30]. The use of metadata and their accuracy have increased over the past several decades. The quality of metadata is probably the single most important factor that enables the reusability of scientific data.

Data Paper. Recently, a mechanism, the *data paper,* able to improve data understandability and, thus, data reusability has been proposed. A data paper can be defined as a *scholarly publication of a searchable metadata document describing a particular on-line accessible dataset, or a group of datasets*, published in accordance to the standard academic practices [31]. In essence, a data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. As such, it contains facts about data, not hypotheses, and arguments in support of those hypotheses based on data, as found in a conventional research article. Its purpose is threefold: (i) to provide a citable journal publication that brings scholarly credit to data authors; (ii) to describe the data in a structured human-readable form; and (iii) to bring the existence of the data to the attention of the scholarly community.

A data paper should describe how the data sets were collected/created, who collected/created them and who owns these data sets, which software was used to create the data sets, the spatial and temporal coverage of the data sets, etc. It could also include sections summarizing the history of the data set, e.g., original purpose, funding body, etc., as well as its perceived value and usefulness to scientific research (fundamental and/or applied).

An important feature of data papers is that they should always be linked to the published datasets they describe, and that this link (an URL, ideally resolving a digital object identifier, DOI) should be published within the paper itself.

6.2. Domain-Specific Ontologies

Ontologies constitute a key technology enabling a wide range of data services [32]. The growing availability of data has shifted the focus from closed, relatively data-poor applications, to mechanisms and applications for searching, integrating and making use of the vast amounts of data that are now available. Ontologies provide the semantic underpinning that enables reuse of research data [33,34]. Current research is exploring the use of formal ontologies for specifying content-specific agreements for a variety of data/knowledge reuse activities.

A community of practice has to establish its own domain of discourse and choose a formalism, i.e., a knowledge representation language, in order to create its own domain-specific ontology. In addition, a set of linguistic terms by which the members of the community will refer to these objects must be identified. Building this set of terms is difficult because words often have multiple synonyms and because the meanings of words in natural language always depend heavily on the contexts in which the words are used. To overcome this difficulty, explicit lexicons should be created which offer the members of a community of practice a set of terms with which to refer to specific concepts.

In the context of a networked scientific world, domain-specific ontologies are not standalone artifacts. They relate to each other in ways that might affect their meaning, and are inherently distributed in a network of interlinked semantic resources, taking into account in particular their dynamics, modularity and contextual dependencies. The alignment of domain-specific ontologies is crucial for data reusability. It is achieved through a set of mapping rules that specify a correspondence between various entities, such as objects, concepts, relations, and instances. Several concept and relation constructors are offered to construct complex expressions to be used in mappings [35].

6.3. Data Discovering

By *Data Discovery,* we mean the capability to quickly and accurately identify and find data that supports research requirements. The process of discovering data that exist within a data collection/database is supported by *search* and *query* functionality which exploits data registration and

citation capabilities; and metadata descriptions contained in data *categorization/classification* schemes, data *dictionaries*, data *inventories*, and metadata *registries*.

Data Registration. By Data Registration capability, we mean a capability enabling researchers to make data citable as a unique piece of work. Once accepted for deposit, data should be assigned a "Digital Object Identifier" (DOI) for registration. A DOI [36] is a unique name (not a location) within the scientific data universe and provides a system for persistent and actionable identification of data. Identifiers should be assigned at the level of granularity appropriate for an envisaged functional use. The Data Registration capability should include a specified numbering syntax, a resolution service, a model and an implementation mechanism determined by policies and procedures for the governance and application of DOIs.

Data Citation. Data can also be identified and accessed through a publication by means of a data citation capability. By data citation capability, we mean a capability providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. A Data Citation capability should include a minimum of five components [37]: the author of the dataset, the date the data set was published, the data set title, a Unique Global Identifier system (Life Science Identifiers (LSID), Digital Object Identifier (DOI), Uniform Resource Name (URN), etc.) and a Universal Numeric Fingerprint (UNF). The UNF is a short, fixed-length string of numbers and characters that summarize the content of the data set, such that a change in any part of the data would produce a completely different UNF. The fifth component is necessary because unique global identifiers do not guarantee that the data do not change in any meaningful way when the data storage formats change. Together, the Global Unique Identifier and UNF ensure permanence, verifiability, and accessibility, even in situations where the data are confidential, restricted, or proprietary.

Data Classification. Data Classification is the categorization of data for its most effective and efficient use. Data can be classified according to any criteria. A well-planned data classification system makes essential data easy to find. This can be of particular importance in data discovery. A classification scheme should allow/help scientists to effectively answer the following questions:

- What data types are available?
- Where are certain data located?
- What access levels are implemented?
- What protection level is implemented and does it adhere to compliance regulations?

Although data classification is typically a manual process, there are many tools from different vendors that can help gather information about the data. They help to categorize data for several purposes.

Data Dictionary. Data Dictionaries contain the information about the data contained in large data collections. Each data element is defined by its data type, the location where it can be found, and the location where it came from. Often, the data dictionary includes the logic when a field is derived. Typically, each data collection has its own data dictionary.

Metadata Registry. By domain-specific Metadata Registry, we mean a registry used to describe, document, protect, control and access informational representations of a scientific domain. There are two types of metadata registry: (i) metadata schema registries which are databases containing metadata schemas relative to the data collections/databases of a scientific domain; (ii) metadata registries that hold metadata and reference information, a kind of index of terms regarding the data stored in the data collections/databases of a scientific domain. These two types of registry can be components of a 2-tiered metadata registry architecture.

A Metadata Registry supports data reuse as it:

- holds precise data definitions and descriptions;
- holds documentation of data characteristics;
- provides guidance for the identification of data elements stored in data collections/databases;

- provides means for organizing standard shareable data elements; and
- sets up common standards between communities of practice

6.4. Data Exchangeability

By data exchangeability we mean the ability of two entities, i.e., data author and data user, to exchange meaningful data sets. Data exchangeability is a prerequisite for data reuse. During the data exchange process, especially when data are moving between scientific disciplines, three types of "*heterogeneity*" must be addressed.

First, heterogeneity between the native data/query language (of the data author) and the target data/query language (of the data user). When this heterogeneity is resolved, we say that *syntactic exchangeability* between the two entities has been achieved.

Second, heterogeneity between the data models adopted by data author and user to represent information objects. When this heterogeneity is resolved, we say that *structural exchangeability* between the two entities has been achieved.

Third, heterogeneity between the domain of discourse of data author and user. When this heterogeneity is resolved, we say that *semantic exchangeability* between the two entities has been achieved.

These three levels of exchangeability, i.e., syntactic, structural, and semantic allow a meaningful exchange of data between the data author and user. However, the three levels of exchangeability do not guarantee that the data user is able to reuse the exchanged data; they only constitute a *necessary* but not *sufficient* condition for effective data reuse.

The main concept enabling data exchangeability is mediation [38]. This concept has been used to cope with many dimensions of heterogeneity, spanning data language syntaxes, data models and semantics. The mediation concept is implemented by a mediator, which is a software device capable of establishing exchangeability by resolving heterogeneities.

6.5. Linking Data to Publications

In data-dominated science, scientific communication undergoes a significant change. Modern scientific communication should support the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. The need to cite data is starting to be recognized as one of the key practices underpinning the recognition of data as a primary research output rather than as a by-product of research. Data will, thus, become a first-class citizen of the scientific communication. Linking scientific data to publications will produce significant benefits as publications: (i) facilitate data findability; (ii) facilitate data interpretability; and (iii) provide the data author better credit for the data.

As a consequence, accessing a data set through a scientific publication will increase the usability of this data set.

Linked Open Data

The usability of scientific data could be greatly increased by the adoption of the "Linked Data" technologies as they provide a more generic, more flexible data publishing paradigm that makes it easier for data producers to interconnect their data with those produced in other scientific disciplines and for data consumers to discover and integrate data from large numbers of data sources. The term Linked Data refers to a set of best practices for publishing structured data on the Web [39]. In particular, Linked Data provides [40] (i) a *unifying data model*. Linked Data relies on Resource Description Framework RDF as a single, unifying model; (ii) *a standardized data access mechanism*. Linked Data commits itself to a specific pattern of using the HTTP protocol; (iii) *hyperlink-based data discovery*. By using URIs as global identifiers for entities, Linked Data allow hyperlinks to be set between entities in different data sources; and (iv) self-descriptive data.

Linked Data have gained significant uptake in several scientific domains as a technology that allows to connect the various data sets that are used by researchers in different scientific domains and

to navigate along the RDF links between different scientific data sets as well as between publications and supporting data.

Recently, a grassroots effort, the Linked Open Data, is aiming to publish and interlink open license data sets from different data sources as Linked Data on the Web.

7. Standards

The role of standards in increasing data understandability and reusability is crucial. Standardization activities characterize the different phases of the scientific data life-cycle. Several activities aim at defining and developing standards to represent scientific data, i.e., standard data models; standards for querying data collections/databases, i.e., standard query languages; standards for modeling domain-specific metadata information, i.e., metadata standards; standards for identifying data, i.e., data identification standards, standards for creating a common understanding of a domain-specific data collection, i.e., standard domain-specific ontologies/taxonomies and lexicons, standards for facilitating the transfer of data between domains, i.e., standard transportation protocols, etc.

A big effort has been devoted to creating metadata standards for different research communities. Metadata standards vary in terms of their specificity, structure, and maturity largely because each standard has been developed on the basis of the needs of a particular user community.

Given the plethora of standards that now exist, some attention should be directed to creating crosswalks or maps between the different standards.

In [41], the standardization is considered to be particularly important for the reuse of data across distance, where the use of data outside their original context implies distance. The word distance is subject to a variety of interpretations. Most commonly, distance is used to refer to something outside the local sphere of activity. An example of this definition is the space between the assumptions and methods of one discipline and another. Distance can also exist within a community, for reasons such as personal or institutional status, subspecialty, or epistemological view. Additionally, the word distance can be defined in a temporal sense. For example, there can be a time lag between the original data collection and reuse.

Standards are important because they can help to span all kinds of distance (spatial, temporal, cultural, etc.) as they have the capability to transform local knowledge into public knowledge and thus avoid that epistemological differences due to distance can lead to different interpretations of the same data.

8. Concluding Remarks

Research data reuse is the quintessence of the open science and open data principles on which modern science is based. To make it feasible, first, all potential barriers, technological, political, legal and economics must be identified. In a previous paper [42], we have described the technological barriers that hinder research data (re)usability. In this paper, we have described the conceptual foundations of research data (re)usability. This does not mean that we underestimate the importance of the policy and legal aspects and their power to hinder data reuse.

At present, the trend in the research data management practices is the creation of domain-specific data centers [43] designed to ensure the stewardship and provision of quality-assessed data and data services to the international science community and other stakeholders. Each domain-specific data center has the responsibility for defining an appropriate Data Management Plan. In this DMP, a conceptual data usability scenario must also be included; it should contain:

- identification of the "usability relationships" between the data author and potential data users;
- transformation of the explicit and tacit knowledge accumulated during the data production process into mobile knowledge in order to be transferred and translated from the data author context to the data user contexts; and

- definition of appropriate levels of data abstraction and data representation to be communicated to the potential users.

In essence, in the context of a DMP, a Data Publication process must be put into action that will perform all the functions listed in Section 5.

There are three main contributions in this paper: first, it presents the problem of scientific data (re)usability in a structured and comprehensive form; second, it sets the conceptual foundations of data usability by borrowing basic concepts from other theoretical fields, i.e., relational thinking, levellism, and knowledge representation and applying them in the data usability field; and third, it identifies the data publication process as the enabler of data usability.

Therefore, in making scientific data reusable first, a conceptual data usability scenario must be defined; it includes:

- identification of the "usability relationships" between the data author and potential data users;
- transformation of the explicit and tacit knowledge accumulated during the data production process into mobile knowledge in order to be transferred and translated from the data author context to the data user contexts; and
- definition of appropriate levels of data abstraction and data representation to be communicated to the potential users.

Then, a Data Publication process must be put into action that will perform all the functions listed in Section 5 and will be implemented within the conceptual data usability scenario identified beforehand.

In conclusion, we foresee that the well-established model of scientific publishing will be increasingly complemented by a system of data publication and that many of the issues regarding research data sharing and reuse could be effectively addressed if the principles of Data Publication are applied within a Data Usability Conceptual Framework.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Hey, T.; Tansley, S.; Tolle, K. (Eds.) *The Fourth Paradigm: Data Intensive Scientific Discovery*; Microsoft Research: Redmond, WA, USA, 2009.
2. Thanos, C. Global Research Data Infrastructures: Towards a 10-Year Vision for Global Research Data Infrastructures—Final Report. 2011. Available online: http://www.grdi2020.eu/repository/filescaricati/e2b03611-e58f-4242-946a-5b21f17d2947.pdf (accessed on 6 January 2017).
3. Zimmerman, A. Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists. Thesis, Degree of Doctor of Pholosophy Information and Library Studies, University of Michigan, Ann Arbor, MI, USA, 2003. Available online: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/39373/ann_zimmerman_dissertation_2003.pdf?sequence=2 (accessed on 6 January 2017).
4. European Commission. Commission Recommendation on Access to and Preservation of Scientific Information. 2012. Available online: https://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf (accessed on 6 January 2017).
5. Amsterdam Call for Action on Open Science. 2016. Available online: https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science (accessed on 6 January 2017).
6. European Commission; Directorate-General for Research & Innovation. H2020 Programme, Guidelines on FAIR Data Management in Horizon 2020. 2016. Available online: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed on 6 January 2017).
7. National Research Council. *Bits of Power: Issues in Global Access to Scientific Data*; National Academy Press: Washington, DC, USA, 1997.

8. National Research Council—Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*; National Academy Press: Washington, DC, USA, 1999.
9. National Science Board. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. 2005. Available online: https://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf (accessed on 6 January 2017).
10. Osterlund, C.; Carlile, P. Relations in practice: Sorting through practice theories on knowledge sharing in complex organizations. *Inf. Soc.* **2005**, *21*, 91–107. [CrossRef]
11. Musen, M. Dimensions of Knowledge Sharing and Reuse. *Comput. Biomed. Res.* **1992**, *25*, 435–467. [CrossRef]
12. Kanfer, A.G.; Bruce, B.C.; Haythornthwaite, C.; Burbules, N.; Wade, J.; Bowker, G.C.; Porac, J. Modeling distributed knowledge processes in next generation multidisciplinary alliances. *Inf. Syst. Front.* **2000**, *2*, 317–331. [CrossRef]
13. Star, S.; Griesemer, J. Institutional ecology, translations, and coherence: Amateurs and professionals in Berkeley's museum of vertebrate zoology. *Soc. Stud. Sci.* **1989**, *19*, 387–420. [CrossRef]
14. Floridi, L.; Sanders, J. *Levellism and the Method of Abstraction*; Research Report 22.11.04; Information Ethics Group (Oxford University and University of Bari): Oxford, UK, 2004.
15. Wickett, K.; Sacchi, S.; Dubin, D.; Renear, A. Identifying Content and Levels of Representation in Scientific Data. In Proceedings of the American Society for Information Science and Technology, Baltimore, MD, USA, 28–31 October 2012.
16. Stonebraker, M.; Becla, J.; Dewitt, D.J.; Lim, K.T.; Maier, D.; Ratzesberger, O.; Zdonik, S.B. Requirements for Science Data Bases and SciDB. In Proceedings of the CIDR 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, 4–7 January 2009.
17. Neches, R.; Fikes, R.; Finin, T.; Gruber, T.; Patil, R.; Senator, T.; Swartout, W.R. Enabling Technology for Knowledge Sharing. *AI Mag.* **1991**, *12*, 36–56.
18. Lawrence, B.; Jones, C.; Matthews, B.; Pepler, S.; Callaghan, S. Citation and Peer Review of Data: Moving Towards Formal Data Publication. *Int. J. Digit. Curation* **2011**, *6*. [CrossRef]
19. The Royal Society Science Center. *Science as an Open Enterprise*; The Royal Society Science Center: London, UK, 2012.
20. Costello, M. Motivating online publication of data. *Bioscience* **2009**, *59*, 418–427. [CrossRef]
21. Parsons, M.A.; Fox, P.A. Is data Publication the Right Metaphor? *Data Sci. J.* **2013**, *12*, WDS32–WDS46. [CrossRef]
22. Willinsky, J. *The Access Principle: The Case for Open Access to Research and Scholarship*; MIT Press: Cambridge, MA, USA, 2006.
23. Cragin, M.; Heidorn, P.B.; Palmer, C.; Smith, L. An Educational Program on Data Curation. In Proceedings of the American Library Association Conference, Science and Technology Section, Washington, DC, USA, 25 June 2007.
24. Kobielus, J. Big Data and the Power of Positive Curation. 2014. Available online: http://www.ibmbigdatahub.com/blog/big-data-and-power-positive-curation (accessed on 6 January 2017).
25. Gray, J.; Szalay, A.S.; Thakar, A.R.; Stoughton, C.; Vandenberg, J. *Online Scientific Data Curation, Publication and Archiving*; Technical Report MSR-TR-2002-74; Microsoft Research: Redmond, WA, USA, 2002.
26. Ikeda, R.; Widom, J. Panda: A System for Provenance and Data. *IEEE Data Eng. Bull.* **2010**, *33*, 42–49.
27. Moreau, L.; Freire, J.; Futrelle, J.; Mcgrath, R.E.; Myers, J.; Paulson, P. The Open Provenance Model: An Overview. In *IPAW 2008: Provenance and Annotation of Data and Processes*; Springer: Berlin, Germany, 2008; Volume 5272.
28. Strang, T.; Linnhoff-Poppien, C. A Context Modeling Survey. In Proceedings of the First International Workshop on Advanced Context Modeling, Reasoning and Management Associated with the Sixth International Conference on Ubiquitous Computing, Nottingham, UK, 7 September 2004.
29. Batini, C.; Scannapieco, M. *Data Quality: Concepts, Methodologies, and Techniques*; Springer: New York, NY, USA, 2006.
30. Gray, J.; Liu, D.T.; Nieto-Santisteban, M.; Szalay, A.; Dewitt, D.J.; Heber, G. Scientific Data Management in the Coming Decade. *SIGMOD Rec.* **2005**, *34*, 34–41. [CrossRef]
31. Chavan, V.; Penev, L. The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science. *BMC Bioinform.* **2011**, *12*, 2399–2405. [CrossRef] [PubMed]

32. Gruber, T. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In *Formal Ontology in Conceptual Analysis and Knowledge Representation*; Technical Report KSL 93-04; Knowledge Systems Laboratory, Stanford University: Palo Alto, CA, US, 1995.
33. Calvanese, D.; Giacomo, G.D.; Lembo, D.; Lenzerini, M.; Poggi, A.; Rosati, R. Ontology-based Database Access. In Proceedings of the Fifteenth Italian Symposium on Advanced Database Systems, SEBD 2007, Torre Canne, Fasano, Italy, 17–20 June 2007; pp. 324–331.
34. Poggi, A.; Lembo, D.; Calvanese, D.; Giacomo, G.D.; Lenzerini, M.; Rosati, R. Linking Data to Ontologies. *J. Data Semant.* **2008**, *10*, 133–173.
35. Thanos, C. The Future of Digital Scholarship. *Procedia Comput. Sci.* **2014**, *38*, 22–28. [CrossRef]
36. Paskin, N. Digital object identifier for scientific data. In Presented at the 19th International CODATA Conference, Berlin, Germany, 7–10 November 2004.
37. Altman, M.; King, G. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Mag.* **2007**, *13*, 11–26.
38. Thanos, C. Mediation: The Technological Foundation of the Modern Science. *Data Sci. J.* **2014**, *13*, 88–105. [CrossRef]
39. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. *Int. J. Semant. Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]
40. Bizer, C. Interlinking Scientific Data on a Global Scale. *Data Sci. J.* **2013**, *12*, GRDI6–GRDI12. [CrossRef]
41. Zimmerman, A.S. New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. *Sci. Technol. Hum. Values* **2008**, *33*, 631–652. [CrossRef]
42. Thanos, C. Scientific Data (Re)Usability: Concepts, Impediments, and Enabling Technologies. In Proceedings of the International Conference on Digital Presentation and Preservation of Cultural and Scientific Heritage, Veliko Tarnovo, Bulgaria, 28–30 September 1015.
43. JISK, Data Centers: their use, value and impact. A research Information Network Report, September 2011. Available online: http://www.rin.ac.uk/system/files/attachments/Data_Centres_Report.pdf (accessed on 6 January 2017).

© 2017 by the author; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).