Scientific Data Reusability: Concepts, Impediments and Enabling Technologies

Costantino Thanos

Institute of Information Science and Technologies of the Italian National Research Council
thanos@isti.cnr.it

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, as it allows reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others. The paper addresses the technological dimension of data reusability: the scientific data universe, the impediments of data (re)reuse; the data publication process as a bridge between data author and user and the relevant technologies enabling this process.

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. 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 to solve and also lead to serendipitous discoveries. 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 usability we intend the ease of using data for legitimate scientific research by experts in the field for which the data was collected. We use the term data reusability to mean the ease of use of data collected for one purpose to study a new problem [1]. This term denotes the reutilization of existing datasets in significantly different scientific contexts. Data reusability is becoming a distinct characteristic of modern scientific practice, as it allows reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others.

Data reusability has four dimensions: policy, legal, economic and technological. The policy dimension regards actions, at an institutional level, aimed at favouring the open availability of scientific data; the legal dimension regards laws and rules aimed at allowing legal jurisdictional boundaries to be overcome; the economic dimension regards the identification of the costs involved in making data shareable and usable as well as how these costs are distributed between data authors and users; the technological dimension regards technical solutions that 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 scientific data universe. Section 3 discusses the impediments to data (re)reuse while Section 4 identifies and describes the technologies that contribute to span the distance between the data author and user. Section 5 lists the main actions to be undertaken when addressing the pressing need of making scientific data reusable.

2 The Scientific Data Universe

The scientific data universe is complex, involving many actors using many types of data for many different scientific purposes.

Scientific Data

By scientific data we mean scientific or technical measurements, values calculated, and observations or facts that can be represented by numbers, tables, graphs, models, text, or symbol and that are used as a basis for reasoning or further calculation [2].

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.

Data can be referred to **as raw**, **derivative**, or **verified** [3]:

- **Raw data** consist of original observations or generated by an instrument or sensor or collected by conducting an experiment.
- **Derivative data** are generated by processing activities.
- **Verified data** are generated by curatorial activities.

Data Collections/Databases

Scientific data are stored into managed data collections/databases. Data collections fall into one of three functional categories [4]:

- **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.
- **Resource or Community Data Collections** serve a single science or engineering community. They contain data that conform to standards often established at a community-level.

- **Reference Data Collections** are intended to serve large segments of the scientific community. Therefore, conformance to robust, well-established, and comprehensive standards is essential.

Data Actors

The main actors in the scientific data universe are [4]:

- **Data Authors**: are individuals or teams involved in research activities that generate 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.
- **Data Users**: are representatives of the scientific communities. Their interests lie in having ready access to datasets 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. They have the responsibility for (i) data archiving and preservation; (ii) providing for the integrity, reliability, and preservation of the collection and all aspects of change control; (iii) implementing community standards; (iiv) providing for the security of the collection; and (v) providing mechanisms for protecting property rights, confidentiality, privacy.
- **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 [4]:

- **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 entirely different purpose with respect to the original dataset(s).

Diversity in the Scientific Data Universe

In conclusion, we can affirm that scientific 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 making them reusable.

3 Impediments to 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 heterogeneity of representations [5].

- *Heterogeneous Data Representations:* there are 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 systems because we lack an agreed-on protocol specifying how 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 database management systems.
- *Vocabulary Mismatching:* Another barrier to data reuse is when a common vocabulary and domain terminology is lacking.

Discovering data

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.

Understanding data

The next problem regards the capacity of the data user to understand the information/knowledge embodied in it. To make data understandable, they must be endowed with auxiliary information, including metadata, community-specific ontologies or taxonomies, and terminologies.

Moving data

Data actors and data collections inhabit multiple contexts. There is the risk, when data are moving across contexts, of interpreting data representations in different ways. The intended meaning becomes distorted when the data move across semantic boundaries. This is due to the loss of the interpretative context and can lead to a phenomenon called “ontological drift”. 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 dataset does not meet the quality expectations of the user of this dataset.
- *Data-incomplete mismatching* occurs when a dataset 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 user.

4 Data Publishing: A Process for Bridging the Gap Between Data Author and Data User

An emerging “best practice” in the scientific method is the process of Publishing scientific data. By *Data Publication* it is intended a process that allows a data user 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 [6, 7]. The ultimate aim of Data Publication is to make scientific data available for reuse both within the original disciplines and the wider community.

The Data Publication process is composed of a number of procedures that altogether implement the overall functionality of this process. In particular, they should support the following functionality relevant for achieving data reusability: (i) data-peer reviewing; (ii) data discoverability; (iii) data understandability and (iv) making data assessable.

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. We have identified seven main enabling technologies:

4.1 Scientific (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.

Metadadata Modeling

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, provenance, data layout and ideally a great deal more.

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 [8].

Data Provenance Modeling: In its most general form, provenance (also sometimes called lineage) captures where data came from, how it has been updated over time. Provenance can serve a number of important functions [9]: explanation, verification, and re-computation.

There has been a large body of very interesting work in data provenance modelling over the past two decades. There has been considerably less work on querying provenance. On the long-term, a standard open representation and query model is needed. A promising example is the "Open Provenance Model" [10], 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 which 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. 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 [11]: Key-value Models, Mark-up Scheme Models, Object Oriented Models, Logic Based Models, and Ontology Based Models.

Data Uncertainty Modeling: As models of real world, scientific datasets are often permeated with forms of uncertainty. Uncertainty is the quantitative estimation of 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.

Data Quality Modeling: The quality of data is a complex concept, difficult to define. There is no common or agreed definition or measure for data quality, apart from such 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 [12]. 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*.

Data Paper. Recently, a mechanism, the *data paper*, able to improve data understandability has emerged. A data paper is a journal publication whose primary purpose is to describe data, rather than to report a research investigation. Its purpose is threefold [13]: (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.

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.

4.2 Domain-specific Ontologies

Ontologies constitute a key enabling technology enabling a wide range of data services. They provide the semantic underpinning that enables reuse of research data. Current research is exploring the use of formal ontologies for specifying content-specific agreements for a variety of data/knowledge reuse activities.

Ontologies were initially developed by the Artificial Intelligence community to facilitate knowledge sharing and reuse. An ontology consists of a set of concepts, axioms, and relationships. It constitutes an explicit specification of a conceptualization of a domain of interest, that is, an abstract view of the objects, concepts, and other entities that are assumed to exist in a domain of interest and the relationships that hold among them [14]. The abstract view of a domain of interest can be represented by a declarative formalism and the set of objects that can be represented is called the domain of discourse.

A community of practice has to establish an appropriate level of abstraction at which to view objects, concepts and relationships and choose a formalism, i.e., a knowledge representation language, in order to create its own domain-specific ontology.

The main reason to build an ontology–based conceptual view of a data collection is to improve data intelligibility and, thus, usability.

4.3 Data Discoverability

The ability to determine where datasets are located, what is in those datasets, and who can access them is a critical but necessary step in order to enable researchers to access all the data stored in data collections distributed in a networked scientific environment that are relevant to their research activities.

The process of discovering data is supported by *search* and *query* functionality which exploits data registration and citation capabilities, 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 [15] 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.

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 [16]: the author of the dataset, the date the dataset was published, the dataset title, a Unique Global Identifier system (LSID, DOI, URN, etc.) and a Universal Numeric Fingerprint (UNF).

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.

The data dictionary also includes the physical location, such as a server DNS (domain name system) name or the IP address. The data collection name, the instance, the table, and the field name are particularly important for the scientist seeking relevant data. This information is even more important if the scientist has to cross multiple systems to gather the necessary pieces of information for her/his research.

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.

4.4 Linking Data to Publications

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. Linking scientific data to publications will produce significant benefits as publications: (i) facilitate data discoverability; (ii) facilitate data interpretability; and (iii) provide the data author better credits for the data. As a consequence, accessing a dataset through a scientific publication will increase the usability of this dataset.

In order to address the new requirements of modern scientific communication, a new concept of publication has been proposed. A publication is intended as a set of "information units", including text, data sets, images, etc. meaningfully connected by relationships. These models enable integration of data and publications. An example of these new publication models is the "enhanced publication". An *enhanced publication* is intended as an existing publication, e.g., a peer-reviewed textual article, enhanced with relationships to a number of existing objects, such as further publications (cited, similar, etc.) or datasets (used in experiments, resulting from experiments, etc.) [17].

4.5 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 authors to interconnect their data with those produced in other scientific disciplines and for data users 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 [18]. These principles are the following: (i) use URIs as names for things; (ii) use HTTP URIs so that people can look up those names: (iii) provide useful information using recommended standards (RDF, SPARQL); and (iv) include links to other URIs so that they can discover more things.

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

To exploit the potential of Linked Data scientific work environments should have Linked Data import and export features and should provide for publishing scientific data directly to the Web of Linked Data. To enable researchers to navigate between data sources and explore the Web data space a number of Linked Data Browsers have been developed: OpenLinlk RDF Browser, Zitgist RDF Browser, Tabulator Browswer, Disco Hyperdata Browser, etc. In additions, a number of search engines for crawling the web data space and providing best-effort query answers over crawled data have been developed: Falcons, VisiNav, Sig.ma, etc.

Recently, a grassroots effort, the Linked Open Data, is aiming at publishing and interlinking open license data sets from different data sources as Linked Data on the Web.

4.6 Standards

The role of standards in increasing data understandability and reusability is crucial. Standardization activities characterize the different phases of the scientific data lifecycle. 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.

Standardization is particularly important for the reuse of data *across distance* [1], where the use of data outside their original context implies distance.

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.

5 Concluding Remarks

Scientific data reuse is the quintessence of the open science and open data principles on which modern science is based. In this paper, we have only discussed the technological dimension of data reusability problem; this does not mean that we underestimate the importance of the policy, legal and economics dimensions and their power to hinder data reuse.

We suggest, that in order to overcome the impediments, a Data Publication process must be put in action. This process should support the:

- definition of appropriate levels of data abstraction to be communicated to the potential users ;

- definition/adoption of an appropriate metadata model as well as ontologies/taxonomies/vocabularies for a comprehensive documentation of the data to be published;
- adoption of appropriate standards for data registration and citation.

References

1. A. Zimmerman, “Data Sharing and Secondary Use of Scientific Data: Experiences of Ecologists”, Unpublished Dissertation, Information and Library Studies, University of Michigan, Ann Arbor, 2003
2. National Research Council, “Bits of power: Issues in global access to scientific data”, Washington, DC: National Academy Press, 1997
3. National Research Council – Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest 1999, “A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases”
4. National Science Board, “Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century”, September 2005
5. R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, W. Swartout, “Enabling Technology for Knowledge Sharing”, in AI Magazine Vol. 12, Number 3 (1991)
6. B. Lawrence et al “Citation and Peer Review of Data: Moving Towards Formal Data Publication”, in International Journal of Digital Curation, 6 (2)
7. The Royal Society Science Center Report 2012 “Science as an Enterprise”
8. J. Gray, et al “Omline Scientific data Curation, Publication and Archiving” , Technical Report MSR-TR-2002-74, Redmond, WA: Microsoft Research
9. R. Ikeda, and J. Widom 2010 “Panda: A System for Provenance and Data”, in IEEE Data Engineering Bulletin, Special Issue on Data provenance, 33
10. L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, P. Paulson, “The Open Provenance Model: An Overview”, in IPAW 2008, LINCS 5272, Springer-Verlag Berlin, 2008
11. T. Strang and C. Linnhoff-Poppien 2004 “A Context Modeling Survey” in Workshop on “Advanced Context Modeling, Reasoning and Management” associated with the Sixth International Conference on “Ubiquitous Computing”
12. C. Batini and M. Scannapieco 2006 “Data Quality: Concepts, methodologies, and Techniques”, Springer, New York
13. V. Chavan, L. Penev, “The Data Paper: A Mechanism to Incentivize Data Publishing in Biodiversity Science” in BMC Bioinformatics 2011, 12 (Suppl 15)
14. T. Gruber, “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
15. N. Paskin 2004, “Digital Object Identifier for Scientific Data”, 19th International CODATA Conference, Berlin
16. M. Altman and G. King, “A Proposed Standard for the Scholarly Citation of Quantitative Data” in D-Lib Magazine March/April 2007
17. S. Woutersen-Windhouwer et al. 2009 “Enhanced Publications”, edited by M. VernooyGerritsen, SURF Foundation, Amsterdam University Press
18. C. Bizer, T. Heath, T. Berners-Lee, “Linked Data – The Story So Far” in International Journal on Semantic Web & Information Systems, 5 (3), 1-22, 2009