Introduction to the Special Theme

Scientific Data Sharing and Re-use

by Costantino Thanos and Andreas Rauber

Research data are essential to all scientific endeavours. Openness in the sharing of research results is one of the norms of modern science. The assumption behind this openness is that scientific progress requires results to be shared within the scientific community as early as possible in the discovery process.

The emerging cultures of data sharing and publication, open access to, and reuse of data are the positive signs of an evolving research environment. Data sharing and (re)usability are becoming distinct characteristics of modern scientific practice, as they allow reanalysis of evidence, reproduction and verification of results, minimizing duplication of effort, and building on the work of others. However, several challenges still prevent the research community from realizing the full benefits of these practices.

Data sharing/reusability has four main dimensions: policy, legal, technological and economic. A legal and policy framework should favour the open availability of scientific data and allow legal jurisdictional boundaries to be overcome, while technology should render physical and semantic barriers irrelevant. Finally, data sharing/reuse involves economic support: who will pay for public access to research data?

To make scientific data shareable and usable it should be discoverable, i.e. scholars must be able to quickly and accurately find data that supports scientific research; understandable to those scrutinizing them; and assessable enabling potential users to evaluate them. An emerging 'best practice' in the scientific method is the process of publishing scientific data. Data Publication refers to 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 data creators to receive academic credit for their work.

The main technological impediments to data sharing/reuse are:

- *Heterogeneity of 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.
- *Discoverability of data:* In a networked scientific multidisciplinary environment pinpointing the location of relevant data is a big challenge for researchers.
- *Understandability of data:* The next problem regards the capacity of the data user to understand the information/knowledge embodied in it.
- *Movement of data:* Data users and data collections inhabit multiple contexts. 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:* Several data mismatching problems 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.

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
- query data collections/databases - i.e., standard query languages
- model domain-specific metadata information - i.e., metadata standards
- identify data - i.e., data identification standards
- create a common understanding of a domain-specific data collection - i.e., standard domain-specific ontologies/ taxonomies and lexicons
- transfer data between domains - i.e., standard transportation protocols, etc.

A big effort has been devoted to creating metadata standards for different research communities. Given the plethora of standards that now exist, some attention should be directed to creating crosswalks or maps between the different standards.

This special issue features a keynote paper from an EU funding organization, an invited paper from a global organization that aims to accelerate and facilitate research data sharing and exchange, an invited paper from a prominent US scientist and an invited paper from a large Australian data organization. The core part of this issue presents several contributions of European researchers that address the different aspects of the data sharing and (re)use problem.

Please contact:
Costantino Thanos, ISTI-CNR, Italy,
E-mail: thanos@isti.cnr.it

Andreas Rauber, TU Vienna, Austria
E-mail: rauber@ifs.tuwien.ac.at