e-IRG Workshop - Panel on Data Interoperability

Alberto Michelini
INGV
Italy

Main Questions

- What is in your view global research data interoperability?
- Who are the main stakeholders to achieving global research data interoperability?
- Who are the main implementers of global research data interoperability?
- How can we create data infrastructures that overcome fragmentation?
- What do we expect from the different organizations, in particular from RDA?
- How can we improve data re-use by improving interoperability at different layers?
- Do we need a bottom-up or top-down process or do we need both?

What is in your view global research data interoperability?

- Provide tools/services to analyze data *across* communities so that scientist can “cross-correlate” diverse physical processes (e.g., physical phenomena and human reactions)
- Data must be “close” to computational facilities
- Caution with the results obtained since in principle anything can be correlated (!)
- A new class of researchers specialist in the exploitation of large and diverse data volumes is to be formed

Who are the main stakeholders to achieving global research data interoperability?

- It depends very much on the type of services provided

  ➡ Higher level services will be of interest to researchers

  ➡ The education system (e.g., teachers, schools)

  ➡ Citizen scientists

  ➡ Industry

  ➡ ...

Who are the main implementers of global research data interoperability?

- The “data center communities” and the “big data centers” where replica of the community data are also stored through the adoption of the same, though general, **abstract model** of data organization which must necessarily include *persistent identification (PID)*, *metadata* and the creation of *central registries* for discovery.

How can we create data infrastructures that overcome fragmentation?

- Some level of fragmentation is inevitable. We must **engage** at the beginning the largest communities in adopting some common, standard “abstract model".

- This can be achieved through use cases, tutorials, training...

- The preparation of APIs that would allow the "qualification" of the data according to the abstract model, would give the community data centers not only to test but also to take a pro-active role in the data organization, the adoption of existing (core) services and the development of new ones.

What do we expect from the different organizations, in particular from RDA?

- RDA should be acting as a entity/forum where the “abstract model” is discussed and defined in its various forms considering the diversity of the various data, the requirements dictated by discovery of the data (e.g., metadata, ...), the needs of mining huge data volumes, the constraints posed by connectivity, the ever emerging technologies, ...

- The RDA should be aware of what goes on in the communities and provide guidance also through the definition of standards.

How can we improve data re-use by improving interoperability at different layers?

- The overall vision should be that of making each step of a scientific analysis completely transparent and reproducible. (e.g., ...Just think a paper where it is possible to click on the data, the analysis software, the intermediate results and play around with different parameter values to see how stable the results are or to compare the results with another work...).

- Data once properly archived and "qualified" will be necessarily re-used. This is why it is so important to agree on a standard, general enough abstract model of data organization in which all can find place to organize their data.

Do we need a bottom-up or top-down process or do we need both?

- We need both!
- The communities acquire the data and are the only ones that can perform the required quality control and metadata assignement.
- The ITs can provide modern, technical solutions and the big data centers space for storage besides facilitating access.
- The origin of the data must be always clear and referable univocally to facilitate re-use, citation and acknowledgement.
    - Note that the communities are sometime a bit uneasy to open their data because very often their acquisition, QA and archiving efforts are not properly acknowledged.