

ATLAS Needs that result in Computational Challenges

Michael Ernst

Information and Communication Technologies are the most recent transformational and enabling factors when extracting the science out of High Energy Physics (HEP) data. They enable close and almost instantaneous collaboration between scientists all over the world and they provide access to unprecedented volumes of scientific information that can in turn be processed on powerful computational platforms. With robust infrastructure for data transmission and data processing in place, we need now start to think about the next step: data itself. A high level goal is a scientific community that does not waste resources on recreating data that have already been produced. Researchers should be able to concentrate on the best ways to make *use* of data. Data become an infrastructure that scientists can use on their way to new frontiers.

Our stock of intangible knowledge, expanding at today's hyper-speeds, needs to be thought of as a new kind of asset in itself that serves all. As such, it requires professional analysis and engineering. Its contents are heterogeneous –different data formats, value and uses. There is tremendous value in having the data made seamlessly available, to use, reuse and recombine to support the creation of new knowledge. And the data must be available to whomever, whenever and wherever needed, yet still be protected if necessary by a range of constraints including licenses, time embargos, community or institutional affiliation.

To collect, curate, preserve and make available ever-increasing amounts of scientific data requires new types of infrastructures. The result will be a vital scientific asset: flexible, reliable, efficient, cross-disciplinary and cross-border.

The anticipated infrastructure is supposed to support seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance.

- All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.
- Researchers are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.
- Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. Federated repositories work to international standards, to ensure they are trustworthy.

The focus is on scientific data because, when the information is so abundant, the very nature of research starts to change. A feedback loop between researchers and research results changes the pace and direction of discovery. The "virtual lab" is already real, with the ability to undertake experiments on large instruments in other continents remotely in real time. The ATLAS experiment at the Large Hadron Collider at CERN is an example. Researchers with widely different backgrounds can collaborate on the same set of data from different perspectives. Just how will we train people to work in this environment? What tools do we have or will we need to move, store, preserve and mine these data? How to share them? How to understand them? How will researchers know the data they access remotely are accurate, uncorrupted and unbiased? These are just a few of the profound policy questions posed by this new age of data-intensive science.

Specific challenges associated with data infrastructure for science include

- Open deposit, allowing user-community centers to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the ATLAS/HEP community
- Persistent identification, allowing data centers to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored data
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorization and accounting
- A high degree of interoperability at format and semantic level