

NorStore

Infrastructure for the curation of scientific data

eVITA, the research programme on e-Science from the Research Council of Norway, has recently established a new project - NorStore - that will deploy a sustainable infrastructure for the curation, archiving and preservation of scientific data.

Data curation

Modern collaborative and interdisciplinary science relies on increased sharing of expertise, instruments and computing resources, and, crucially, increasing access to collections of primary research data and information. The sharing of scientific data provides a knowledge base that creates new opportunities and horizons for research and discovery. Researchers rely on the availability of modern information technology tools that assist in the creation, transformation, discovery, re-exploitation and presentation of data. However, these tools evolve rapidly and the flexibility in using these tools puts the data itself at risk. The survival of digital scientific information depends on a hierarchy of constantly shifting technologies - hardware, storage media, operating systems, data formats, applications software and middleware. It also relies on tacit knowledge that is external to the data. In practice, much remains to be done at all levels to keep data usable and valid to future researchers.

At the same time, the sizes of scientific data collections have increased to the Terabyte scale. Large-scale, standardized and quality-controlled data infrastructures are emerging in areas like biology, physics and earth sciences. A well-known example is the Large Hadron Collider (LHC) at CERN that will enter production in 2008. Norway participates in the world-wide collaboration for the analysis of the data that will be generated by LHC and it is expected that only Norway will already store about 1.5 PetaByte (1 500 TeraByte) of data before the end of 2010. Also in other disciplines, high-resolution data is collected from real-time instruments (e.g., sensors) and large complex distributed databases are used. Observational data often concerns unique events and the measured data needs to be stored for a long period (or even forever) as it cannot be recreated.

In today's digital environments, it is not obvious to have trust in data which has been passed on. Trust in data can be enhanced by the existence of qualified domain specialists who curate the data and deal with issues of security, confidentiality and privacy, ownership, provenance, authenticity, integrity, as well as the quality of the primary data and associated metadata. Besides trust, aspects that impact quality include discoverability (e.g., how to find data in foreign domains or out-dated archives), access

management, heterogeneity of data formats, and complexity of composite data (possible including links to external objects and external dependencies). The long-term validity of data collections also crucially depends on the existence of policies and awareness on what to keep and how.

As a consequence, modern science demands increasingly advanced levels of data curation, i.e., strategy, policy and practice regarding the creation, management, and long-term care of data. Curation is about maintaining and adding value to digital information for contemporary and future use. Scientific data collections are not merely stored or archived anymore, but are subject to frequent revision and enhancement.

NorStore

The eVITA programme on e-Science from the Research Council of Norway has recently established a new project - NorStore - which shall meet the needs for data storage and data management from several fields within the natural sciences. The primary objective of the project is to establish and maintain a nationally coordinated infrastructure to support the curation, archiving and preservation of digital scientific data. The aim is that the infrastructure shall make scientific processes that rely on access to foreign data sets more efficient and ultimately, support multidisciplinary communities and improve the cross-fertilisation of scientific results. The infrastructure will facilitate the creation and use of digital scientific repositories that satisfy internationally accepted standards and protocols. The sciences that are targeted include, but are not limited to, earth sciences, biosciences, chemistry, physics, material sciences, fluid dynamics, and the medical sciences.

The project will operate large scale data storage resources, provide support for individuals and groups that have a need for storage capacity, digital repositories and curation services, and promote a set of standard services and best practices that aim to improve the reuse and reusability of scientific data. The infrastructure will support easy, secure and transparent access to geographically distributed databases and repositories, provide large aggregate capacities for storage and data transfer, and optimize the utilization of the overall resource capacity that is available in the infrastructure.

The infrastructure will be an integrated part in the national e-Infrastructure and will be connected to resources that are located at several major research centers in Norway. The project shall engage into collaborations with parties that have similar objectives, interests and needs for services. National and international co-operations will be established with organizations that have an interest in infrastructure for scientific data and a need for standardization and interoperability of services.

The strategic responsibility for the envisaged infrastructure as well as the prioritization of the user communities that will be granted access lies within the Research Council of Norway. The project will be a broad and nationally coordinated effort. UNINETT Sigma has the operational responsibility. The initial project consortium includes UNINETT and the four universities UiO, UiB, UiT and NTNU. The actual operation of the production systems will be the responsibility of centres that have the expertise, competencies, and the required infrastructure to provide a comprehensive service to meet the challenging demands of academic user groups.

International infrastructure for data

Also internationally there is increasing awareness that there is a need to establish infrastructure for scientific databases and repositories that is deployed according to strategies, standards, policies, and community needs. The European Strategy Forum for Research Infrastructures (ESFRI) has identified a number of strategic infrastructures for European scientists and engineers to remain competitive internationally and to maintain or regain leadership [1]. Several of these infrastructures crucially depend on the availability of large scale storage resources and curation services. A high-performance distributed data infrastructure is an indispensable tool to support the solution of challenging large-scale problems in emerging European infrastructures for large scale supercomputing (e.g., Partnership for Advanced Computing in Europe - PRACE) and federated computing environments (e.g., Enabling Grids for E-Science - EGEE).

Recently, the Nordic Council of Ministers recognized the necessity to establish a common Nordic e-Science strategy. A working group that was given the task to draft a first strategy recommended amongst others the establishment of a Nordic infrastructure for digital databases and repositories [2].

2008

The NorStore project envisages the creation of a permanent infrastructure that needs to be designed with care and in collaboration with many parties. The start-up of the project must address models for management, collaboration, operation, support, usage and financing. The project must also define policies and best practices for establishing and maintaining data repositories, define a core set of services, standards and interfaces that shall be maintained across the infrastructure and establish a peer review process to prioritize and support leading-edge science and optimal use of the infrastructure.

Another important activity in the start-up of the project is the establishment of the initial physical infrastructure and in particular the storage resources and services. The storage resources from the first procurement will be installed early 2008. It is envisaged that the infrastructure will be upgraded regularly and at least once a year. The upgrade will consist of expanding the storage capacity and adding new resources.

URL NorStore: <http://www.norstore.no/>

Reference:

- [1] European Roadmap for Research Infrastructures – Report 2006. ESFRI. ISBN-92-79-02694-1. <http://cordis.europa.eu/esfri/>
- [2] Nordic eScience. Research, Education, and Sustainable Infrastructure Services. A strategy document for the Nordic Council of Ministers 2007-07-17.