Helmholtz Data Federation

Large-scale experiments and simulations in science generate an increasing amount of data. The transformation of data and information to findings and knowledge, however, also needs a new quality of storage and analysis capability. The Helmholtz Association now takes an architectural role in the permanent, secure, and usable storage of data. For managing big data in science, it has established the Helmholtz Data Federation (HDF). Within the next five years, about EUR 49.5 million will be invested into multi-disciplinary data centers and modern data management. The HDF will establish a data federation comprising three elements: innovative software technologies, excellent user support and leading-edge storage and analysis hardware.

The HDF as a national research data infrastructure constitutes the long-term federation of powerful, multi-disciplinary data centers. Combining these federated data storages with the existing expertise and knowledge of the six partners in research data management and user support provides a unique research infrastructure, which will promote and foster the transformation of data into knowledge and thereby support excellent science in Germany and beyond.

The federation is built on efficient software methods and tools of distributed data management and secure network links among each other and within Helmholtz, to university partners and further research organizations in Germany and internationally via DFN. The HDF represents the nucleus of a national research data infrastructure across science organizations, which is open to users in the whole German science community. International connections will make it compatible with the future European Open Science Cloud (EOSC).

The data centers at Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Deutsches Elektronen-Synchrotron (DESY), GSI Helmholtz Centre of Heavy Ion Research, German Cancer Research Centre, Forschungszentrum Jülich, and Karlsruhe Institute of Technology with strong topical profiles (Figure 1) are enhanced with leading-edge storage and analysis resources and technologies. This will ensure that the ever increasing volume of valuable research data in various scientific disciplines is stored and archived, long term access is guaranteed, data ownership is preserved and new perspectives can arise for intra- and interdisciplinary transformation of data into knowledge with high relevance for science, industry and society.

Federating resources and knowledge is a common principle in science and well adopted in various research domains. The guiding principle behind the HDF is the open federation of leading-edge storage and analysis hardware through innovative software as well as excellent user support for the preservation of data and metadata itself, their integrity, provenance, moral and legal ownership as well as their original access rights. Initially this federation is build up across several Helmholtz Centers connecting to a majority of science disciplines in Helmholtz. The HDF can be used by scientists from

Fig. 1: Centres and Use Cases.

Helmholtz, universities and other research organizations and institutes across Germany ultimately leading to a nation-wide, federated infrastructure for research data of the entire German science system. Conceptually additional data and computing centers from Helmholtz, universities and other research organizations in Germany can be added in an efficient, secure and transparent way. The federation is established through structural elements of innovative data management software, security and identity mechanisms, broadband network connections as well as a competence network of human experts. Besides the federation and sharing of research data, the federation and sharing of knowledge, expertise and software among the HDF partners and with the scientific communities using the HDF will provide unprecedented advancements beyond state of the art. The federated approach of the HDF will foster existing and new scientific collaborations inside and across scientific domains and communities. New collaborations will arise between biology, life science and photon science for the understanding of biological structures and processes or between energy, climate and marine research to optimize energy systems of the future based on renewable energies. The HDF will allow for mutual use of data by cross-linking and annotating. The development and deployment of methods to enhance sharing and re-use by applying common standards will bring science one step closer to universal data access, where researchers from different disciplines will have the chance not only to search but also to find answers from data collected in other scientific domains. For example, information from metagenomics (an approach to reveal the full diversity of life in each specific sampling location) can be combined with environmental parameters (physical, chemical, and other biological) in these locations. This will enable a much more detailed understanding of ecosystems, which in turn will help to predict changes in biological productivity (e.g. fisheries, agriculture) under conditions of climate change.

Written by Daniel Mallmann
Jülich Supercomputing Centre (JSC), Germany

and Prof. Dr. Achim Streit
Karlsruhe Institute of Technology, Steinbuch Centre for Computing

Contact: Daniel Mallmann, d.mallmann@fz-juelich.de
Prof. Dr. Achim Streit, achim.streit@kit.edu