DOI: 10.7250/eunis.2013.045

Data storage architecture to support various research needs and data lifecycle management

Ville Tenhunen[1], Minna Harjuniemi[2]

[1]University of Helsinki, P.O. Box 28, FI-00014 University of Helsinki, ville.tenhunen@helsinki.fi
[2]University of Helsinki, P.O. Box 28, FI-00014 University of Helsinki, minna.harjuniemi@helsinki.fi

Keywords
Research data, Data storage, Data strategy, Data life cycle management.

1. ABSTRACT

IT architecture and life cycle management is essential when cost-efficient solutions are developed for research data management. The IT center of the University of Helsinki (UH) has an ongoing project to find suitable services for varying needs to handle research data and offer services to researchers with storage needs. Several findings and suggestions have been made during the project.

2. INTRODUCTION

The amount of research data increases and new areas of science have become more data intensive.

Storage size is not only the challenge which one has to solve. Questions about life cycle of the data or future estimation of the storage size are needed for architectural considerations. Backups, access methods, user rights, used applications, needs to process data in storage or willingness to share data are conventional points of view for building storage solutions. IT administration is also interested in system maintenance, data security and privacy issues.

Helsinki University Library and UN's IT Center has couple of development projects concerning research data management, processes and services. In Finland The Ministry of Education and Culture has launched The National Research Data Project (TTA) which focus is to set up a service solution for research data storage (IDA), to support the production of metadata and to work towards a long-term preservation solution together with the National Digital Library.

3. VARIOUS NEEDS FOR STORAGES IN RESEARCH

Planned and to at least some extent governed data life cycle management is essential. The need of a service varies depending on the phase of the data life cycle. For example, a typical life science data may go through the following phases: raw data (created by, for example, a DNA sequencer), processed data (rough analysis, maybe compressed) waiting to be fully analyzed, active data that is used in analysis, data to be published and archived data. Not all data goes through all the phases, but different technical solutions may be needed in each phase. The data that needs storage may be created as a result of a research project or as a raw material for research itself.

Another aspect is local services vs. centralized ones. In many cases some capacity is needed close to the scientific instruments in the collecting phase. Services for the next phases may be offered centrally by the university IT or in some cases nationally by NRENs. For IT this means various types of architectures and number of actors within one service framework.

UH's IT Center has some productized services to handle different size needs, mainly solutions based on NAS- or SAN-technologies or local storage servers. Also customized solutions and services from Finnish NREN (CSC) have been used.

4. ARCHITECTURE AND LIFECYCLE MANAGEMENT

Unless the published data is totally open, various mechanisms to authorize users are needed. User federations are a solution to some extent, but fine-grained access control may be a difficult thing to achieve on the international level.

DOI: 10.7250/eunis.2013.045

Open data and open access are preferred goals, these do not eliminate the need for metadata governance which has to be done to ensure the usability and availability the data to public users. Adding metadata to the data sets is usually not perceived as a part of the research process. It is important to encourage researchers to work with metadata. Situation will be better if administration of data sets will be taken into account as researchers merits.

Data privacy and security have to be handled differently depending on the data nature. Not all data consists of human or commercially valuable information, but the data that does, needs to be fully secured with specialized and sometimes costly solutions. Differentiation is needed to do things cost-efficiently and to avoid overkilling.

Data formats and software are important life cycle topics. There is not any guarantee that all data will be readable with future applications. This is one major challenge of long term preservation.

The storage solutions in UH were built based on the needs of business critical systems like email, SAP and learning environments. This means that these solutions are far too high-end and thus costly for basic research needs. Sometimes manual solutions are good enough for research needs and costs are reasonable. Standardization of hardware solutions reduces the costs.

5. DISCUSSION

Administrative environment of research data management has become more challenging. Research administrations of the state (like ministries) or university have set some rules and regulations concerning publicity, intellectual property rights etc. Sponsors could have their own restrictions (e.g. patents) and ownership demands. Research communities have their principles. Libraries and long term preservation set some demands for metadata, formats and interfaces. Data security rules and regulations are more complex and so on. Therefore it is important to make technological solutions as simple as possible to use.

Architecture of research data management is heterogeneous because researchers' needs vary a lot. It is not possible to build research data architecture based on only one technological solution. For example cloud services could be a part of the solutions, but not a whole story.

The researcher knows exactly the needs she/he has, but not always knows what the best solution is. Tight interaction between research and IT experts is needed, and probably in many universities the IT's problem is the same as in UH: the IT Centers are used to serve administrative needs and are now currently finding the correct ways to switch focus to the needs of IT intensive research. It is not a question of fine storage systems, it is the question of great science.

6. AUTHORS' BIOGRAPHIES

Ville Tenhunen, M.Sc.

Works as a Project Manager in the Technology Services Unit of the IT center in the University of Helsinki. In the current position he has lead research data projects where research data services for researchers have been designed, developed and implemented. He is a member of data infrastructure working group of The National Research Data Project (TTA) funded by the Ministry of Education and Culture.

Minna Harjuniemi

Works as an IT manager and is the head of the Technology Services Unit of the IT center in the University of Helsinki and continues studying IT governance. Before the current position she was working as an IT expert in several areas including identity management, servers and storage and internet services. She is also a member of steering committee of The National Research Data Project (TTA) funded by the Ministry of Education and Culture.