



Norwegian Infrastructure for the Curation of Scientific Data

Jacko Koster
UNINETT Sigma

www.norstore.no

Trends in Norwegian e-Infrastructure:

Need for support of research that is based on access to collections of primary research data and information incl. research that does not have a need for HPC

Need to separate long-term storage from HPC facilities

Need for strategy, policy and practice regarding the creation, management, and long-term care of data:
data curation

Data repositories that are actively curated have become a reality. Data is not merely archived anymore

Need for tools to assist in discovery, re-exploitation and presentation of data

Trends in Norwegian e-Infrastructure (ctd):

Sizes of scientific data collections have increased to the Terabyte scale

Information technology tools evolve rapidly and the flexibility in using these tools put the very data they create and transform at risk

Survival of digital scientific information depends on a hierarchy of constantly shifting technologies – hardware, storage media, operating systems, applications software and middleware. Overcome the technical obsolescence problems

There are many reasons to keep data:

- retention of unique observational information which is impossible to recreate
- retention of expensively generated data which is cheaper to maintain than to recreate
- reuse data for new or future research purpose
- validate and account for publicly funded research
- for compliance with legal requirements
- for educational and teaching purposes

Survey 2007: several user groups in Norway expressed a need for long-term storage of (large) data:

- International Polar Year - collection of 30 projects
- Meteorological Institute
- Bjerknes Centre for Climate Research (>1 PB/year)
- Nansen Environmental and Remote Sensing Centre (NERSC) – data assimilation
- CARBOOCEAN – biogeochemistry and CO₂
- PGP – study of violent processes in (geo)physics
- CMBN – molecular biology and neuroscience

The main objective:

Establish and maintain a broad and sustainable infrastructure for the curation, archiving and preservation of data from computational science and the natural sciences.

NorStore will

- operate storage resources and peripheral equipment
- provide support to researchers need storage capacity, digital repositories and curation services
- promote a set of standard services and establish best practices and policies that aim to improve the reuse and the reusability of scientific data
- provide easy, secure and transparent access to distributed storage resources, provide large aggregate capacities for storage and data transfer, and optimize the utilization of the overall capacity

The infrastructure will be an integrated part of the national e-Infrastructure (and co-exists with HPC, national network and national grid)

The infrastructure must proven to be sustainable, cost-efficient, allow efficient utilization of the available resources, services and competencies and shall be attractive to a wide range of sciences.

The infrastructure must

- enhance the ability of researchers to extract further meaning from masses of data stored in institutional, national, international or community repositories,
- contribute to the standardization and interoperability of repositories and software interfaces for storage in general
- increase the pooling of resources and competencies across the participating centres

The data revolution raises non-technical issues wrt:

- security
- confidentiality and continued privacy
- ownership
- assured provenance
- authenticity and integrity

How to guarantee the quality of the primary data and associated metadata? **Trust** in data can be enhanced by the existence of qualified domain specialists who curate the data

Open access: immediate, free, unrestricted on-line access

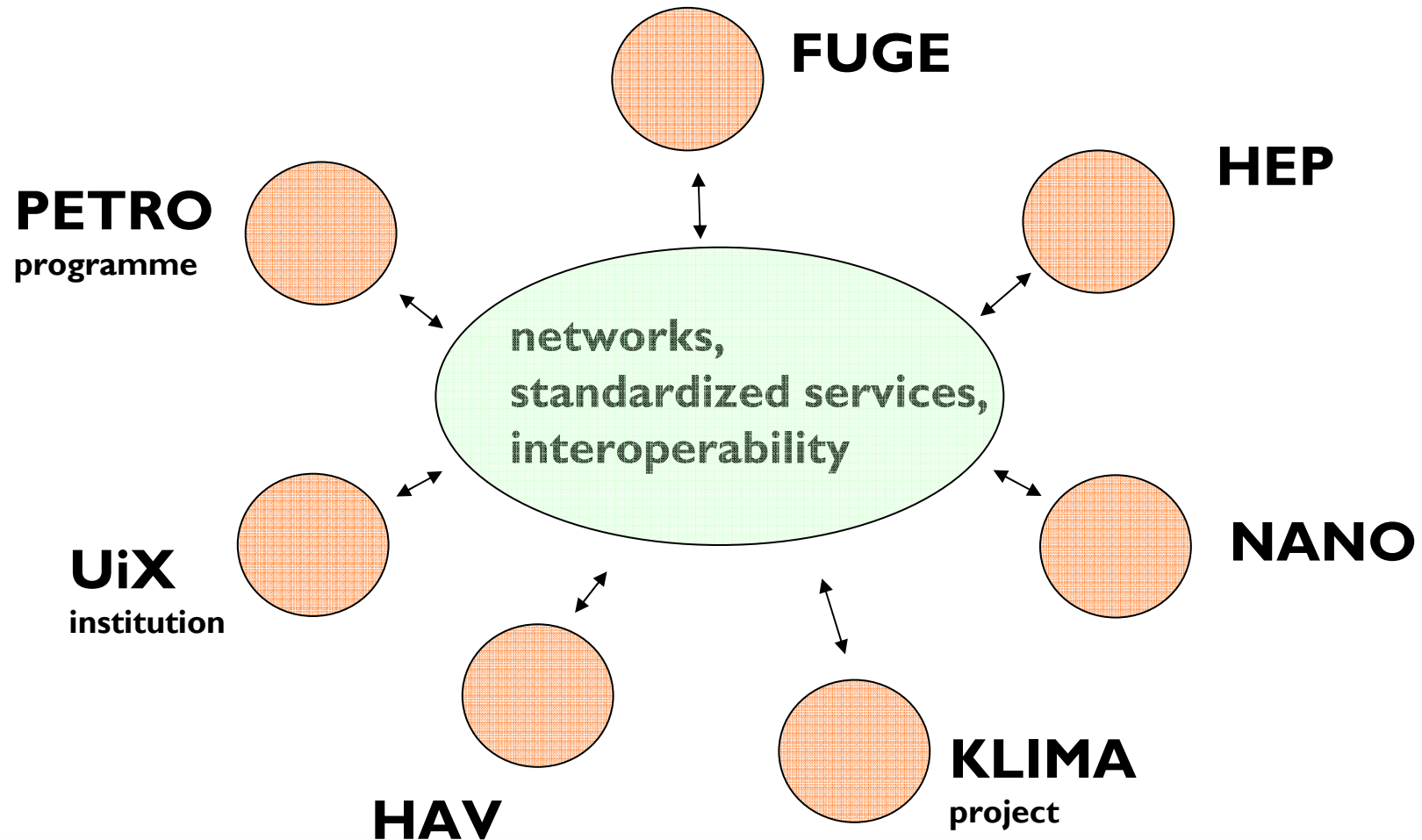
Concrete tasks for the project:

- coordination of investments in the infrastructure, in particular in large-scale storage resources and recruitment of skilled personnel
- operation of the resources, tools and services in the infrastructure
- provide expert advice and assistance to users of the infrastructure and more generally, to individual and groups that create and maintain data collections
- ensure cost-efficient and high utilization of the overall infrastructure

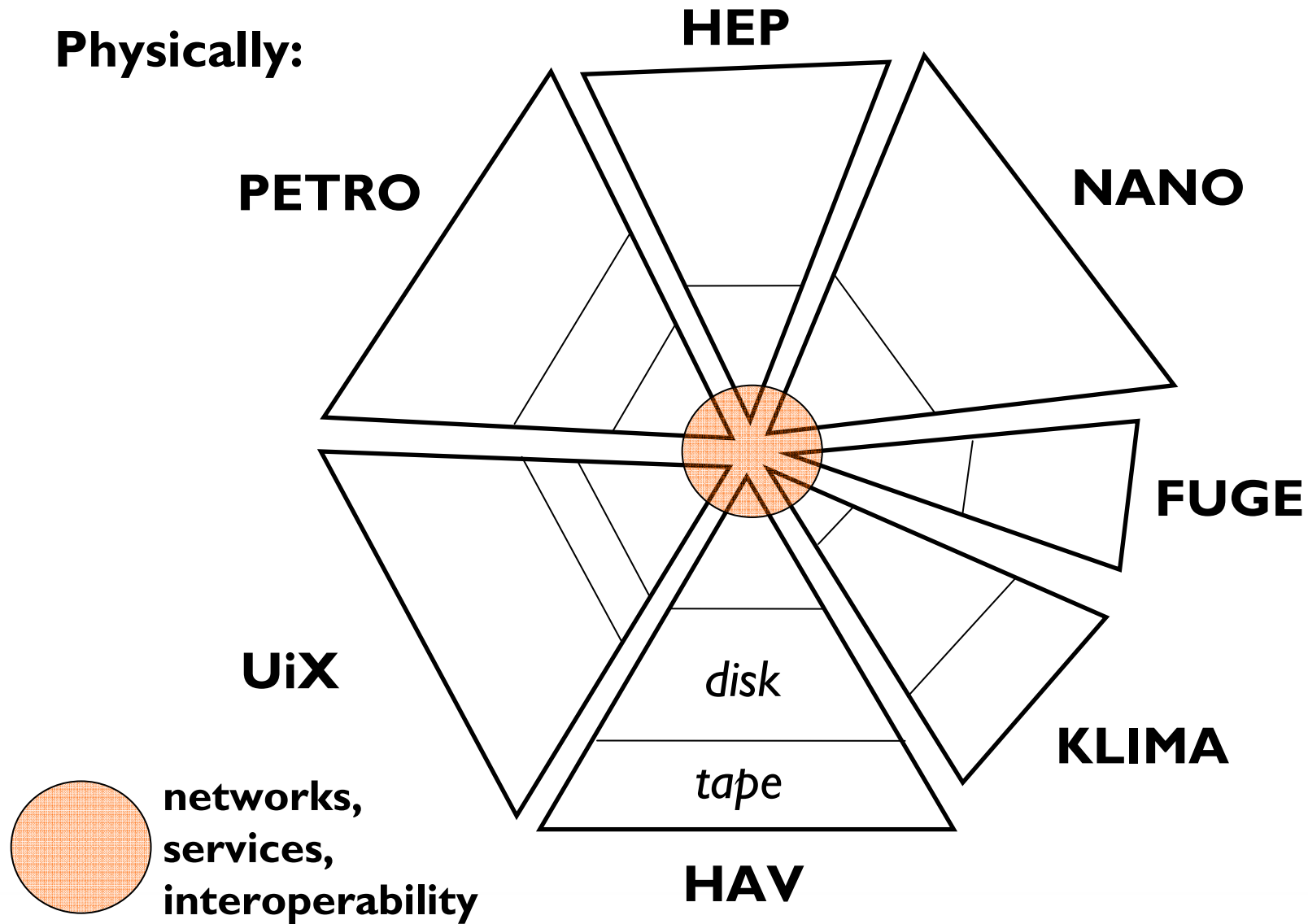
Concrete tasks for the project (ctd):

- maintain a core set of services, tools and protocols, as required by user communities, (international) collaborative partnerships
- coordination and administration of access to the infrastructure
- contribute to national policy and establishment of best practices for using the infrastructure and for the curation, archiving, and preservation of data in general
- increase awareness of the importance of data curation to all stakeholders

Logically: (repositories)



Physically:



User access:

The infrastructure is open to all sciences that are within the responsibility of the Norwegian e-Science programme (computational sciences, natural sciences, ...)

Access to the infrastructure will be by application and will be governed by a Committee appointed by the Research Council of Norway.

Application criteria shall include

- scientific merit of the research
- feasibility of the usage of the infrastructure
- duration of usage, known/expected future usage
- type of usage (e.g., on-line repository, archive, ...)
- proper content management of the data collections

Also aspects of security, confidentiality, privacy, ownership, provenance and formats of the data, restrictions for third parties to access the data shall be considered.

Initial consortium: Universities in Bergen, Trondheim (NTNU), Oslo, Tromsø, UNINETT (NREN), UNINETT Sigma

Collaborations with organizations and international efforts that have an interest in infrastructure for scientific data. E.g.,

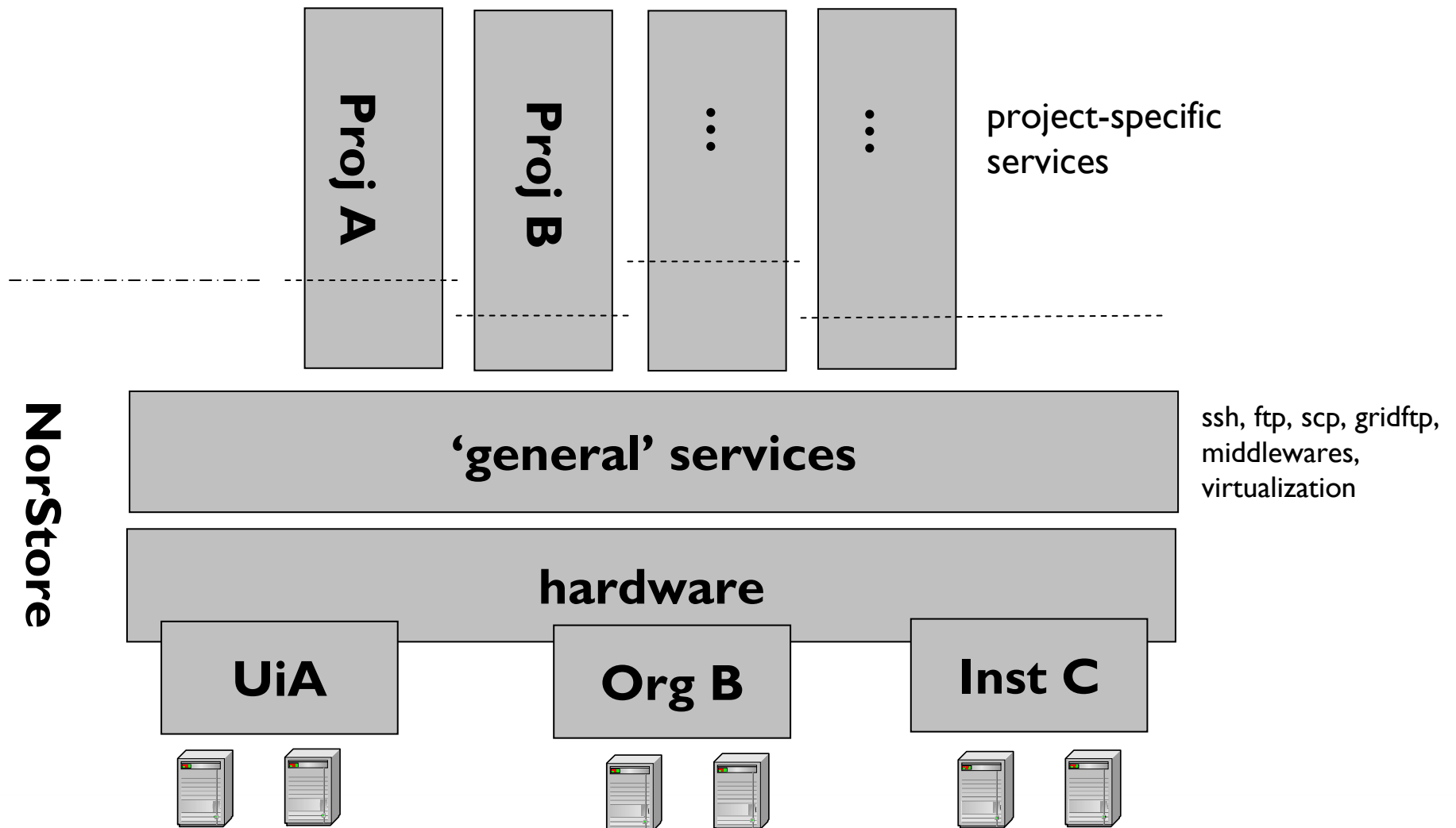
- National Library, Norwegian Social Science Data Services
- Similar Nordic initiatives
- Coupling to European initiatives: EGI, PRACE, ...

In 2007, focus is on the initial specification of the infrastructure, choice of technologies, levels of curation, roadmap for 2008, accumulation of experience.

A main activity late 2007 and early 2008 is the investment in hardware for the initial infrastructure. The investments must be such that the initial infrastructure can be expanded and upgraded in a cost-efficient manner in the coming years.

Characteristics and challenges:

- Distributed infrastructure (multiple sites, long distances)
- Heterogeneous infrastructure (multiple types of resources and storage media)
- Heterogeneous data, large data sets
- Heterogeneous usage: active repositories, archiving, back-up, scratch areas, ...



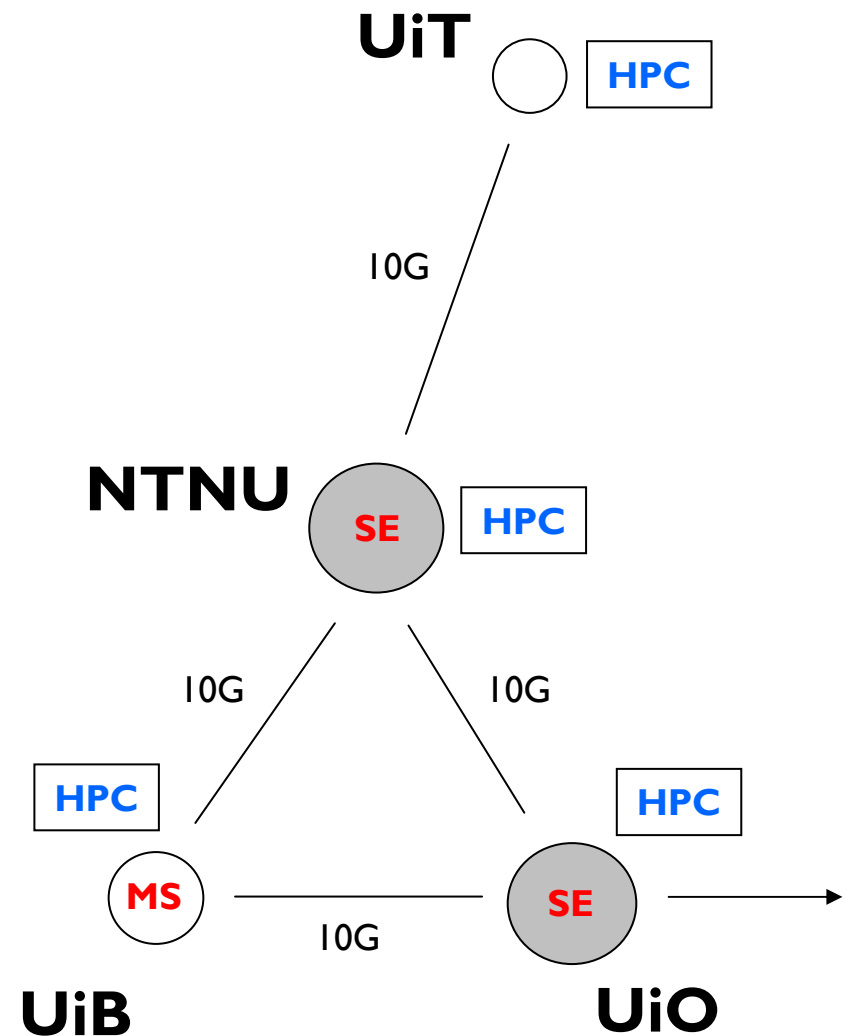
NorStore

Initially: Two storage elements
(ca. 600 TB each). Installation
January 2008.

Tape-robot expansion at UiB.

Built bottom-up from identical
homogeneous independent
resources. Core set of services
(back-up, mirrors, archives, ...)

Technology project to investigate
software solutions





Further information

www.norstore.no