FP7 infra-2007 1.2.1 : Scientific Digital Repositories

Technical considerations which we may or may not be disclosed in the official proposal. The following collection of thoughts should should only be seen as a starting point of the technical discussion. I would expect a lot input here.

These are the areas identified as basic building blocks for the Scientific Repository Approach.

- Storage Implementations
- Storage Control Protocols
- Data access protocols (local and wide area)
- Data discovery protocols
- Storage Location capability and discovery protocols
- Meta Data Storage and discovery
- High level data replication protocols
- Application interface layers
- Applications

Details

Storage Element Implementation

These are the Storage Elements, we have under our control. They would need to be tuned to comply with the requirements of this project. At the time being this is dCache and StoRM, which would be covered by DESY and INFN. To our current understanding CERN is not willing to let DPM or CASTOR be part of this project for various reasons. In addition to that, there is a technical reason why the two Storage Elements, we picked, are more suitable than others.
In order to integrate non HEP communities, it will sooner or later turn out that a posix access to the data is required. The so called *Posix Like approach* won't be sufficient for the future if arbitrary communities are involved.

Storage Control Protocols

In order to shuffle data around, some storage control protocol might become necessary in order to negotiate on the transfer protocol, or to reserve space on the destination site. I would prefer not to step into the SRM discussion again but some SRM subset would certainly a good candidate. But any other (simple) solution would be welcome as well.

Wide area data transfer protocol.

The only standard candidates here are certainly *http(s)* and *(gsi)ftp*. Due to my experience with gsiFtp and firewall setting, I personally would prefer http(s).

Local area transfer protocols

In order to allow arbitrary applications to access data seamlessly, direct posix access should an option. The StoRM SE provides this by definition and dCache will to for nfs4.1 which we hope will be finally full specified very soon. If there is no other way we may still try to follow the traditional HEP approach of posix like access svia special libraries. (rfio,dcap,xroot).

Data Discovery Protocols & Meta Data

It's obvious that we have to provide a way of finding data. It is not clear to me how far we should go with this. For sure we need at least simple File Catalogues providing *Logical Filename* to Storage Location mappings. Higher level replication services need to be based on this. There could be the approach of semantic meta data repositories and searches. This always sounds very attractive to reviewers but is already difficult in single community projects and might turn out to impossible for our approach. (keyword : common ontology)

Storage Capabilities and Discovery Protocols

This is certainly needed. There is a lot effort going into the GLUE schema, so it would be wise to at least try to adopt this as close as possible. The transport protocol is by no means clear to me at this point.

High level replication protocols

Although those services are available (FTS and globus), it's not yet obvious whether or not we need this.

Application and Application layer interface.

Here I depend on the input of the non HEP communities.