ARCHIVER Project

Technical Summary

THE BABAR EXPERIMENT

Problem Definition:
In 2020 the BaBar Experiment infrastructure at SLAC will be decommissioned. As a result, the 2 PBs of BaBar data can no longer be stored at the host laboratory and alternative solutions need to be found. Currently a copy of the data is being held by CERN IT. We want to ensure that a complete copy of Babar data will be retained in such a way that will allow researchers to re-use and re-analyse their data.

Lifecycle - Workflow Characteristics:
The end users are not allowed to self-deposit the data. The data is curated by the service team together with physicists and is being released together after checks.

The recall of the data from the archive is rare. The recall may happen on demand by the Service Manager in order to replace a defect file.The end users won't have access to the Archive. In case of data recall by the Service Manager, the full dataset unit will be downloaded from the archive, although it would be preferable to be able to access one particular file from the dataset.

In order to ensure data reusability and research reproducibility for this particular deployment scenario, preservation efforts should not only focus on the data, but also on the analysis level software and data format, the reconstruction/simulation software, as well as the additional relevant documentation.

Authentication and Management Functions:
The authentication needs of the Babar Experiment data are minimal, because they do not require any special access control or federated authentication for users. The only person allowed to deposit data to the Archiving Service is the Service Manager, which might use either a local account or a CERN account via federated authentication.

Due to the particular nature of research data material, the Archive does not have to process the deposited material in any special way beyond the usual conformance tests and fixity checks. It

would be useful to have periodic reports of any data management processes and activities happening in the Archive, such occasional fixity checks.

Data and Metadata Characteristics:
The data is packaged in the form of datasets. The full dataset is 2 PB in size. Their metadata is rather simple and define very basic attributes of the files such as source, location, type (e.g. raw) etc.

Interface Characteristics:
For the simple archiving of the BaBar dataset, there are no particular requirements on interfaces. The Service Manager should have a well-defined way to deposit data for ingestion by the Archive and to retrieve data from the Archive. For both ingestion and recall, the HTTP protocol with well-defined REST API services are preferred, but other protocols could also work.

For the service manager dashboard, a web access should be offered for interactive peruse. A REST API access to report logs for individual assets or for certain time periods would be a plus.

Reliability Requirements:
For the simple archiving of the BaBar dataset, the Archive should be able to guarantee file recall within a few hours. In both cases, the Archive is expected to guarantee bit preservation. This could be done via sufficient number of copies of data on disks or tapes or in any other suitable way. However, bit preservation is not enough to ensure the full data reusability and research reproducibility needed for this deployment scenario.

Compliance and Verification:
The Archive should validate the integrity of deposited material through schema validation upon ingestion. The Archive should also run periodic fixity checks to verify the checksums of the deposited files.

Cost Requirements:
We are looking at the most effective cost solution that would guarantee minimally reasonable ingestion and recall speeds (~ 10 Gbps). The focus is on providing excellent data integrity in the Archive in order to avoid data loss. The data recalls from the Archive are expected to be very rare. The service cost should include data preservation management tasks to guarantee the service over a long period (~ 5 years). For the 2 PBs of BaBar data, the cost is estimated to be below 100K € per year which equals to 50K € per PB per year.

Initial Data Management Plan:

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	BABAR is a particle physics experiment designed to study some of the most fundamental questions about the universe by exploring its basic constituents - elementary particles. The BABAR Collaboration's research topics include the nature of antimatter, the properties and interactions of the particles known as quarks and leptons, and searches for new physics.
Documentation and data quality	See the project website at http://www-public.slac.stanford.edu/babar/default.aspx. BaBar publications can be found at http://www-public.slac.stanford.edu/babar/Publications.aspx.
Storage and backup during the research process	During the research process the primary data was stored in IBM's HPSS system on robotic tape storage in the SLAC computer center. Copies were made to TierA sites, including several in Europe (RAL in the UK, IN2P3 in France, KIT in Germany etc.) Given the data volumes, the data itself was not "backed up" but HPSS was responsible for ensuring data integrity. BaBar (SLAC) was a member of the DPHEP Study Group and details of its data preservation objectives and benefits can be found in http://arxiv.org/pdf/1205.4667.
Legal and ethical requirements, codes of conduct	N/A. Information on the organisation of the BaBar collaboration, responsibilities etc can be found at https://www.slac.stanford.edu/BFROOT/www/Organization/index.html (including the Collaboration Governance, Management Plan and Membership).
Data sharing and long-term preservation	The BaBar experiment does not make "Open Data" releases. Its data may only be used by

	members of the BaBar Collaboration. However, in principle anyone can become a collaboration member by writing to the BaBar Spokesperson. Some of the BaBar data is unique and a double copy is already “preserved” at CERN. Additional copies also exist at sites that are members of the Collaboration and the intent is to store a further copy using ARCHIVER services to study feasibility, cost of entry, cost of ownership etc.
Data management responsibilities and resources	See https://arxiv.org/pdf/1205.4667.pdf.