Australian Synchrotron Development Plan
Project Submission Form

Section A: Summary and Proponent Details

Project Title

Petabyte Data store for beamline data

Spokesperson

Name	Richard Farnsworth
Institution	AS
Email	Richard.farnsworth@synchrotron.org.au
Phone	+613 8540 4118

Executive Summary (approx. 100 words)

The Australian Synchrotron has a need to retain and archive data generated on all the beamlines and the accelerator for a significant period of time. This data should be stored in a secure and safe fashion allowing the original researcher and subsequent authorised researcher to access the data for any purpose, including original research, verification, data mining, and further deeper data analysis. Original data is kept for a short time, and only to the limit of the particular technology of the beamlines. The onus for data storage and curation is left to the user A Petabyte storage array is the collective name for the various technologies used to such a safe and secure data storage.

Other proponents (add more rows if necessary)

Name	Institution	Email address

Section B: Detailed Description

Attach a document using the following headings (max 10 pages):

B1: Description of Proposed Beamline/Development Project

The Australian Synchrotron has a need to retain and archive data generated on all the beamlines and the accelerator for a lengthy period of time. This data should be stored in a secure and safe fashion allowing the original researcher and subsequent authorised researcher to access the data for any purpose, including original research, verification, data mining, and further deeper data analysis. Currently this function is not able to be performed by the facility. Original data is kept for a short time, and only to the limit of the particular technology of the beamlines. The onus for data storage and curation is left to the user.

A Petabyte storage array is the collective name for the various technologies used to such a safe and secure data storage. It will consist of an array of disk, tape and networking technologies combined to give the facility a better ability to retain the data for the long term.

It needs to constantly be evolving to take into account the fast moving technology – for example certain tapes and disks in common use a few years ago can no longer be read - and is likely to be several Petabytes in size.

This project will setup a set of disk, tape and networking technologies continuously refreshed in a series of ongoing upgrades that will allow the Australian Synchrotron to securely store initially around two terabytes per day. This figure is anticipated to grow as the beamlines become more mature and data collection becomes more efficient.

In order to store five years of data, it is anticipated that three to five petabytes is required, although this project should be structured so that it caters for initially around 1-1.5 petabytes with some constant upgrades built into the technology. In this way at the end of five years a five petabyte data store will exist based on the expected technology improvements and the constant refreshing of technology.

.

B2: Applications and Potential Outcomes to Australian Scientific Community

How does the project advance synchrotron-based research in Australia/NZ? What are the likely outcomes? Include specific examples where possible.

The physical retention of the data files in originally collected format allows a whole new field of research based on secondary data researchers. It also allows the original data collector (or researcher) to have confidence in the data removed from the experimental area. It means that the data itself can be cited, and independent data verification by third parties, including for example science historians can be made.

One example is the use of data from the Hubble space telescope. Once the Principle investigator (PI) has had a period of time with exclusive access to the data (For Hubble this is one year) the

data is made available to all. This has allowed an approximately three fold increase in data citations compared to regular telescopes, Ref [1]

Ref [1] Publication statistics for recent Papers from the Hubble space telescope

Helmut A. Abt

Kitt peak national observatory, TUSCON, Az 5-Feb-2001

B3: Match to Selection Criteria
These criteria can be found in the guidelines.

B4: Potential Users
Does the project address a clearly identified need in the community? The need may be actual or potential.

The retention of Beamlines data by the facility supplement a growing trend in the e-research community to provide tools and techniques to search and catalogues data. This include the development of web "portals" to access data and the use of grid technologies to move and process data. Significant other projects in both data cataloging and data transmission are supplemented by the basic data storage at the synchrotron.

Additionally, the retention of data locally, means that post data collection processing by high performance computing facilities, already planned for the AS can be performed. This will allow greater efficiencies and accuracy in the data collected and thus improve the data citation rates.

It is anticipated the users on the IM beamlines using compound tomography, users on Microspectoscopy using Fluoresce tomography and the macro protein crystallographers will benefit the most initially.