Theme 6 - Data Management Software Systems

James Dempsey, Minh Huynh, Rodrigo Tobar
12 April 2017

Workshop goals

- Develop several use cases for an SKA Regional Centre data archive
- Start collaborations necessary for developing an SKA Regional Centre data archive

Workshop plans

- Work on a few detailed use cases - focus on enhanced data products (see section 5)
- Extract any requirements or design principles resulting from the use cases

Long term plans

Initially the longer term goal was to gather lessons from the pathfinder and precursor data archives. As discussions went further along on the second day the benefits of a longer term goal of setting up a data archive in China became clear.

Replication of existing radio-astronomical data from Australia to China

As a preparation step towards the establishment of a Chinese SRC, a proposal will be written for a project to replicate existing, public ASKAP data from Australia to China, similar to how data will be replicated from the Science Data processor (SDP) into the SRCs, and to use existing CSIRO ASKAP Science Data Archive (CASDA) tools to present this data to users. This new archive will be actively refreshed with new incoming data.

This work presents the following benefits for the development of SRCs:
- It will give a first overview of how data replication will occur from the SDP to the SRCs (and between SRCs) in general, and from Australia to China in particular.
- It will be an opportunity to explore different technologies for data transfer, and assess the strengths and shortcomings of each of them.
- It will provide an example of how federated queries and VO will work on such a system of data archives.
- Additionally, it will enable the Chinese community to easily access publicly available ASKAP data.

To perform the data transfer we propose to initially use the NGAS system, but an alternative based on swarm technologies (e.g., the BitTorrent protocol or similar) should also be investigated to better understand the implications of such a system in the overall architecture of the data management software, and its benefits in potentially reducing the bandwidth usage. Once the data has been transferred, it will be made available using the currently existing CASDA software, including both its web front-end and its VO interface.

There exists at the moment a similar effort to have an MWA archive in Shanghai, coordinating these efforts is probably desirable.

The following assumptions have been made, and need to be ensured:

- ASKAP Data is publicly available.
- Data can be transferred from Australia to China using a reliable link.

SHAO, ICRAR, and CSIRO are interested in this work, but resources would be required for the project. We plan this to be a work package or sub-project within any funding proposal to the "Australia-China Science and Research Fund (ACSRF)". Before that, during the next ERIDANUS workshop we will focus on producing a detailed design of this work.

Assumptions About the SKA Regional Centres

Our theme had discussions about the roles of the SDP vs SKA Regional Centres, and how the SKA Regional Centres may operate. These areas are still undefined but we used the following assumptions to guide our use cases and project proposal.

- SRCs will have specific areas of interest, not all SRCs will have all data
- SDP and SRCs need to be able to discover which SRCs have which data products
- Middleware used for data distribution has to be the same for all SRCs
- SDP will use the SRCs to provide an active backup
- SDP has a mapping of projects to SRCs to use for backups
- SDP will push the data products out to the appropriate SRCs based on the project mapping
- Middleware will service requests for data not in the local SRC by knowing which sites have the data and use swarm style transfer from SDP/SRCs

Use Cases for Enhanced Data Products

User-initiated Deposit of Enhanced Data Products

User steps:

1. Authenticate/login
2. Use a user interface to:
   - Define data type
   - Describe data product
   - Provenance e.g. Telescope, project id
   - Contributors
   - etc
3. Set embargo period/access restrictions
4. Provide data:
   - Upload from local laptop/desktop (small files only); or
   - Upload from SKA RC computer
5. Is there an approval process?

System Steps:

1. Automated data checks
   - Metadata checks

- Data format checks
2. Index data
- Spectral, spatial and temporal coverage
- File list
- Provenance, project ID
3. Copy and physically put data in archive
4. Send data product listing back to SDP for dissemination

Automated generation and deposit of enhanced data products

This process aims to support repeated actions such as source finding catalogues generated in the SRC from images.

User script actions:

1. Authenticate/login
2. Use an API to generate the product (e.g. catalogue)
- Choose an image
- Run source extractor (e.g. Blobcat, Aegean)
- Produces a catalogue (best case in VO Table format)
3. Script or program to initiate deposit
- Need to supply (e.g. generate) metadata as per manual process
- Upload/point to data
- Provide embargo/access restrictions

System Steps are the same as in use case 5.1