

Towards a self-healing archive Stephan Witz, NRAO



History & status

- NRAO Archive spans 4 decades, includes Legacy VLA, \bullet VLBA, Jansky VLA and GBT: different instruments & formats, data from commissioning, testing & production, raw data files with embedded or associated metadata + searchable metadata index
- 1.6 PB + 400 TB/year, ~20k projects, ~12m scans
- Generations of systems:
 - 1st: metadata index in text file, call the data analysts, they make a tape and send it to you
 - 2nd: metadata index database in Ingres, then Oracle, very simple search engine, download via http or ftp, limited reprocessing abilities
 - 3rd: (imminent): metadata index database in PostgreSQL &
 - SOLR, 2 browser interfaces, scriptable, extensive reprocessing

capabilities ADASS XXVI – Towards a self-healing archive



2

Scope of the problem

- Developing next generation, noticed inconsistencies in metadata index during migration to PostgreSQL
 - Projects with stop dates = start dates, missing or incorrect fields (array configurations), unit conversion errors
 ~10% are slightly problematic, ~2% highly problematic
- Problems are mostly with extracting metadata & persisting to index database, not with file storage or bad files
- Discovery led to a dose of humility & two questions:
 - How do we fix the old problems we've discovered?
 - Given we are about to replace the archive ingestion system & we will make some mistakes no matter how well we test, how will we fix new problems we might introduce?



- Solution: self-healing, continuous re-ingestion of metadata
- metadata Use scriptable workflow w/cluster to retrieve projects from the archive and run re-ingestion using latest version of software
 - Flag projects in metadata database as presumed good, suspect or known bad, keep track of date of last re-ingestion & version number of re-ingestion software
 - Focus on known bad and suspect first, but re-ingest all the projects with each new release of re-ingestion software
- Compare (diff) old metadata with new

Reject or add

- If difference is additive replace old with new
- If difference matches an expected pattern, replace old with new
- Otherwise defer decision until human can investigate & intervene

pattern to expected patte

Other uses of self-healing

- Not just for fixing problems, can use self-healing to deploy new metadata or index features to previously observed projects
- Example: scan level retrieval of Jansky VLA projects (desired for Very Large Array Sky Survey)
 - Extend metadata model & database schema to support mapping scan or sub-scan information to BDFs (binary data files) instead of filesets so we know which BDFs to extract
 - 2. Modify data-fetcher to support new model
 - 3. Extend re-ingestion software to support new model
 - 4. Test & release new re-ingestion software
 - 5. Flag existing Jansky VLA projects as suspect
 - 6. Sit back and let system take care of it



Timetable & Risks

- November 2016: testing new metadata models, database schema and search index (SOLR)
- April 2017: testing workflow for re-ingestion, tracking reingestion state in schema
- June 2017: testing difference engine, deferred decisions & expected patterns
- August 2017: throw the switch
- Risk we mess up the metadata database, mitigate with backups and reporting
- Risk we won't have manpower to resolve deferred decisions, but many of these have been bad for a while, so any progress is good progress



Towards a self-healing archive Design of new archive







www.nrao.edu science.nrao.edu public.nrao.edu

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.



ADASS XXVI – Towards a self-healing archive