LTA BF pulsar data

Old pre-Cycle0 data

- **List of all beamformed pre-Cycle0 observations** (mostly pulsars, but also Sun, cosmic rays, etc.):
    - https://www.astron.nl/lofarpwg/psr-precycle0-time.html (user: pwg passwd: pulpme)
    - I'd expect this list to be complete, but no guarantees as there could be missing parsets or other problems before. For most pulsar observations there are also diagnostic plots given from PulP. In some cases there are no plots, which does not necessarily mean that data are bad, could just be that my crawler missed the time window to collect the diagnostic plots.
    - Spreadsheet with observations/directories information for Pulsars2 and Pulsars project (two different sheets) - LTA_crawler_pulsars_20200818.xlsx
    - Listing of all files in the Pulsars projects **(excluding LOTAAS files)** together with the file names, sizes, and creation dates - lofar_pulsar_tape_no_lotaas-20200825.csv

- **Pulsars2** project, have run from 19 October 2011 till 30 November 2012, i.e. for slightly more than 1 year and ended right before the start of Cycle0. This means very high quality data when stations' elements have not aged yet.

- **Publications based on pulsar/BF pre-Cycle0 data:**
    1. Stappers et al. 2011 + van Haarlem et al. 2013
    2. Hassall et al. 2012 → L2009_16100, L2009_16102, L2009_16104, L2009_16116
    3. Asgekar et al., 2013, A&A → RRL (L31848, L25937)
    4. Hassall et al. 2013 → data from 2009-2012
    5. Coenen et al. 2014 + Coenen PhD thesis 2013 → LPPS, LOTAS data (Pulsars project)
    6. Sotomajor-Beltran et al. 2013 → (data from 2011, 2012) L25152 – L25158, L32350 – L32369, L53966 – L53977, L53942 – L53953, L53990 – L54001, L61473 – L61483, L61532 – L61542, L61520 – L61530
    7. Bilous et al. 2014 → pre-Cycle0 (L77924, L78450) + early LC0 data (L99010, L102418, L169237)
    8. Sobey et al. → L34789, L45754, L119505
    9. Pilia et al. 2016 → Pulsars2 and earlier data

Tentative plan to process pulsar BF data on the LTA for size reduction

---

Deletion

In general most of the data in both *Pulsars & Pulsars2* should be kept, namely:

- used in publications (see above)
- long 1+ hrs campaigns (magnetars, XDINSs)
- LOTAAS data (8-bit PSRFITS files) - 1.4 PB (out of 2 PB in both *Pulsars and Pulsars2*)
- pilot surveys (LPPS & LOTAS) - some cleaning could still be done (together with Jason/Joeri)

ACTIONS:

- look over web-summary page (together with the file listings for Pulsars and info in the Excel spreadsheet) for pre-Cycle0 BF data and manually check short (~5-min) observations whether they have scientific value
- *Pulsars2* data can be staged through LTA web-interface **(to be confirmed!)**, diagnostic plots on the web-summary page can be used as well if available
- for *Pulsars* data, need to use grid tools written by Joeri. Hanno Holties can help with that as well (Joeri agreed to share his grid credentials with him to download these data). This is of course only needed if we want to take a look what specific data actually are before making decision (note: if re-processing is needed it can be tricky and take times as data format has changed since then).

Re-processing to reduce size

Given:

1. LOTAAS data by far takes most of the data volume among pulsar BF data (but **only** at SARA). It is at least 7.2+ PB for 'projects' data + 1.4 PB in *Pulsars.*
2. At the moment the most straightforward way to reduce the size is to re-quantize the data to 4/2 bits.
3. Reducing the frequency resolution (number of channels) is possible but should only be considered separately on a project basis. Also, the same data with original (higher) frequency resolution could be used for other science than originally was proposed for.
    1. lesser priority than conversion to lower number of bits
    2. tools are needed to be written to repack files (PSRFITS, Filterbank, raw HDF5) to have smaller number of channels.
4. Items 4 and 5 below seem to be the most straightforward, easy-to-implement and provide significant reduction in the data volume right from the start. And not to mention that project data will be cleaner for an end user, and faster to download.
5. **Important note!** All re-processing should finish with new data ingest, so it's not direct replacement of old data. This ingest will be not through MoM and, thus, will require feedback

files in some different formats, or there will be some other way to provide necessary metadata for the ingest to LTA (file name, size, content of tarballs, etc.). Currently these feedback files are produced by PulP in the format needed by MoM. For simple deletion of the duplicate tarballs we don't need to run PulP, so must be another way to provide this updated metadata for the LTA.

1. raw 32-bit -> raw 8-bit (HDF5)

- identify raw 32-bit BF data on the LTA
- likely not many in Cycle operations. Some number in *Pulsars/Pulsars2* projects
- if in *Pulsars2/CycleX* observations, run 'digitize.py' to convert to 8 bits
- if in *Pulsars*, then modified tool has to be written first as raw data format have changed 2 (?) times before

2. raw 8-bit -> raw 4/2 bits (HDF5)

- rawTo8bit == True
- not sure if HDF5 format allows this at all, or how difficult it will be
- even if possible new tool is needed
- if possible there are quite number of raw 8-bit data already on the LTA, so could reduce size there as well. But comparison need to be done whether data are still good for the same science with reduced dynamic range (see steps below for LOTAAS data)

3. 8-bit -> 4/2 bits (PSRFITS) (mostly LOTAAS, but also for other Stokes I/IQUV observations)

- write the tool to convert 8-bit PSRFITS data to 4 or 2 bits
  → could use code for GBNCC data??
- use few datasets (LOTAAS and others with different levels of RFI during the obs) to convert PSRFITS data to 4/2 bits and then run same processing (search pipeline in case of LOTAAS) to compare the results. Currently it's not clear if we would suffer a lot by going to 4 or even 2 bits.
- make the decision of whether we can convert and to 4 or 2 bits
- identify LOTAAS data on the LTA
- run the conversion tool
- identify other Stokes I/IQUV data
- run the conversion tool
- make sure that these actions are kept in LTA metadata?
- make sure that converted PSRFITS files can be read by other processing tools, or provide description how the processing should change

4. remove tarball duplicates

- some of the tarballs were ingested twice or more times especially in the early days if MoM reported that ingest has failed, etc.
- look for the tarballs with the same name (checksums in the name will be different) for a given ObsID
- make listing of such observations/tarballs **[probably the latest version is the best, but we should see by the file size as well]**

- keep one latest version of the same tarball

5. remove 2nd copy of the same tarball for some observations

- some tarballs can have another tarball inside with the same data
- this was the case from when RO started to ingest data through MoM (i.e. from Cycle 3?) and continued until ???
- investigate when exactly it's started and finished:
  → by downloading a few observation in each Cycle
  → when such an observation found, check the file listing of the tarball in the metadata in the LTA.
  → if another tarball is also in the listing, then this could be used to identify all such observations in the LTA
- remove extra tarballs from within existing tarballs in the LTA, modify file listing, tarball file sizes in the metadata

6. For single-pulse pulsar BF data only: 8-bit Filterbank files (.fil) → 4/2 bits

- singlePulse == True OR rrats == True
- should be Filterbank file (.fil) for every TAB/part in the tarballs → convert to 4/2 bits but only if tests on LOTAAS and other data show that lesser dynamic range won't be a big issue

7. For XXYY data with 1 TAB only without raw 8-bit data (aka Pulsar Timing project): keeping summary tarballs only

- only smaller reduction of volume in comparison with other processing
- complex-voltage data == True
- single TAB data == True
- rawTo8bit == False
- check what number of subbands in the observation is and compare with the number of channels in the summary *.ar file. If the same than individual tarballs (not summary) can be removed. **[to be confirmed!]**
- If number of channels does not match, then we can try again to run *psradd* to add individual parts together. if successful we will re-run the whole processing again on the new *.ar file (PulP), and then individuaal tarballs can be removed.
- If two above bullets failed, then individual tarballs should be kept

8. For all XXYY data

- *.paz.FSCR.AR can be additionally removed if they are the same as *.paz.ar **[to be confirmed!]**