Perspective from the HISWG on SKA Data Processing and Science Analysis

Lourdes Verdes-Montenegro
Instituto de Astrofísica de Andalucía (CSIC)

Co-chair of the HISWG (together with Sarah Blyth)
On behalf of the HISWG

27th June 2019, AENEAS – European SKA Regional Centre Design Meeting (Lyon)

Perspective from the HISWG on SKA Data Processing and Science Analysis

- Strawman SKA1 HI KSPs
- Key steps/procedures of the pipelines/workflows (not delivered by the Observatory)
- Advanced Data Products: list, volumes, formats
- Special considerations on processing
- Required tools, or tools that users are developing and would like to integrate
- Network requirements
- Commensality
- Open science related feedback
- Others

Strawman SKA1 HI KSPs

- High priority HI Science Objectives for SKA1 include both observations in survey mode and pointed observations

4 fiducial KSPs:

Resolved kinematics out to z~0.8

IGM / cosmic web

ISM in Milky Way

z evolution of cold neutral gas

Survey	Area (deg^2)	Freq MHz	HI Resolution	<z> (z_{lim})	T (hrs)
Medium wide	400	950-1420	10"	0.1 (0.3)	2000
Medium deep	20	950-1420	5"	0.2 (0.5)	2000
Deep	1 pointing	600-1050	2"	0.5 (1)	3000
Targeted ISM	30 targets	1400-1420	3"-30"	0.002 (0.01)	3000
Targeted Accretion	30 targets	1400-1420	30"-1"	0.002 (0.01)	3000
Galaxy/MS	500	1418-1422	10"-1'	0 (0)	4.500
Galaxy Abs	(5000)	1418-1422	2"	0 (0)	(10.000)
Absorption	1000+	350-1050	2"	1 (3)	1,000+
	1000	200-350	10"	4 (6)	1.000

Meyer on behalf of HISWG (Town Hall meeting, 2017)
Updated from Staveley-Smith & Oosterloo, 2015, PoS, AASKA14, 167

Strawman SKA1 HI KSPs

- High priority HI Science Objectives for SKA1 include both observations in survey mode and pointed observations

Cosmology Environment Evolution with z

Survey	Area (deg^2)	Freq MHz	HI Resolution	<z> (z_{lim})	T (hrs)	N_{gal}	N(HI) at/cm2
Medium wide	400	950-1420	10"	0.1 (0.3)	2000	~30,000	2e20
Medium deep	20	950-1420	5"	0.2 (0.5)	2000	~25,000	0.6e20
Deep	1 pointing	600-1050	2"	0.5 (1)	3000	~3000	0.4e20
Targeted ISM	30 targets	1400-1420	3"-30"	0.002 (0.01)	3000	30	
Targeted Accretion	30 targets	1400-1420	30"-1"	0.002 (0.01)	3000	30	
Galaxy/MS	500	1418-1422	10"-1'	0 (0)	4,500	1	
Galaxy Abs	(5000)	1418-1422	2"	0 (0)	(10,000)	(~4,000)	
Absorption	1000+	350-1050	2"	1 (3)	1,000+	~5,000	
	1000	200-350	10"	4 (6)	1,000	Unknown	

Band 1 (0.35–1.05GHz)
Band 2 (0.95–4.6GHz)

Meyer on behalf of HISWG (Town Hall meeting, 2017)
Updated from Staveley-Smith & Oosterloo, 2015, PoS, AASKA14, 167

Key steps/procedures of the pipelines/workflows (not delivered by the Observatory)

- Data compression of the **visibilities** for long-term storage.
- Reprocessing of calibrated visibilities
- Daily cubes at multiple resolutions for each pointing
- **Combine cubes** (and beams) for individual pointings in the uv and image domain, e.g to create integrated deep cube
- Combination of spectral line data cubes from different observing runs
- (Deep Galactic and Magellanic HI Survey) = fully calibrated I, Q, U and V cubes at full spectral resolution

Key steps/procedures of the pipelines/workflows (not delivered by the Observatory)

- **Multiscale deconvolution** of sources
- **Spectral line / continuum separation:**
  - ASKAP: (1) subtract continuum sky model, then (2) subtract residual continuum in the image domain.
- Subtracting solar sidelobes
- **Mosaicking** of multiple fields, followed by cutouts in RA, Dec, Freq
- **Zero-spacing** correction
- Calibration techniques for **wide-field** imaging
- **Source-finding** and source parameterisation

Advanced Data Products: List

- Continuum source and spectral line catalogs
- Image cut-outs
- Spectra
- Minicubes
- Catalogues
- Moment maps
- Masks used to make moment maps
- Signal-to-noise maps

Special considerations on processing

- **Constraints on data placement (for both ODP and ADP)**
  - Data for same pointings should be at the same location
  - Global Sky Model could be stored in any node
  - Multiple fields to be mosaicked should be in the same location
  - ADPs (previous list) have small sizes so location is not critical
- **Can all/most processing be done in an unsupervised (batch) mode or is manual interaction/checking essential?**
  - All we can know now is from on-going work with pathfinders/ precursors
  - Supervised:
    - Quality assurance for the calibration and RFI mitigation.
    - Multiscale deconv. of strongest sources
    - Source-finding, masking may require human inspection in complicated cases (interacting galaxies)
  - Unsupervised: in general post-processing
    - Combination of cubes from individual pointings
    - Mosaicking
    - Generation of cutouts, minicubes, moment maps

Advanced Data Products: Volumes, Formats

Calculations courtesy of Erwin de Blok

- SKA1-MID: visibilities
  - 197 dishes, 32768 channels, 4 polarisations, uv sample dump rate of 0.5 Hz
  - —-> Data rate =10.6 GB/s.
  - —->10h track = 372.8 TB
  - —->1000h survey = 36.4 PB.
  - —-> Straw-man survey very roughly 20.000 h = 728 PB
- SKA1-MID: image cubes
  - 3" beam —> 1" pix, 15m dish (FWHM 0.83 deg x 2 =1.67 deg)
  - —-> 6k x 6k pix. + 32768 channels =1179.6 Gpix
  - 64 bit pixels —-> 9,2 Tbyte per data cube of a single pointing

- Extracted data products after removing undetected sources ~10 times smaller (moment maps, pos.vel cuts, spectra)

Required tools, or tools that users are developing and would like to integrate

- Analysis:
  - 3D source finders (SoFIA): Serra+ (2015) Now parallelised
  - Parameterization (BUSY function): Westmeier+ (2014)
  - Kinematic analysis (FAT, Kamphuis+ 2016; 2DBAT, Oh+ 2018; TiRiFiC, Jozsa+ 2007; GalAPAGOS, Wiegert 2011; GIPSY/GuiPSY, Sánchez- Exposito+ 2016; Barolo, Teodoro & Fraternalli 2015; GBKFit Bekiaris+ 2015)
  - Profile decomposition: Oh+ (2019)
  - Stacking: Delhaize+ (2013), Hu+ (2019)
  - CASA
- Visualization (indicate impact of data placement)
  - SlicerAstro (Punzo+ 2016), VISIONS (Task Gipsy 2017), X3D (Vogt+ 2016)
  - VR inspection tools (Marchetti, Jarrett, et al.)
  - New tools to be developed at the e.g. SRCs needed?
- Numerical simulations/models

Network requirements

- Remote visualisation of a volume of 1 sqdeg x 100 MHz or similar

Commensality

- Relevant since this can imply also either common tools or additional tools (use of pol for flagging)
- Existing HI surveys with commensal observing

HI Survey	Commensal with:	
WALLABY (shallow, large area, ASKAP)	continuum (EMU, POSSUM?)	
LADUMA (deep, single pointing, MeerKAT)	MIGHTEE deep continuum + magnetism / polarisation	ThunderKAT slow transients survey
MHONGOOSE (pointed deep individual galaxies, MeerKAT)	MeerQuittens (polarisation/ magnetism)	MIGHTEE
CHILES (deep, single pointing, JVLA)	CHILES con pol (polarisation)	CHILES Verdes (transient variable
FLASH (large area absorption line, ASKAP)	SEAFOG (collab with eROSITA)	
APERTIF shallow and medium deep surveys	many commensal projects including SHARP (absorption)	
MeerKAT Fornax Survey	MIGHTEE	

Compilation Blyth/Verdes-M. Manchester April 2019

Open science related feedback

- Archiving and tagging according to VO standards
- SKA data exploitation through a platform that facilitates:
  - Connection between VO tools and radio tools
  - Metadata to ensure provenance: E.g. for spectra, description of used parameters, like the size of the extraction region.
- Collaboration among international teams in order to extract the maximum scientific knowledge
- Data sharing + re-use & re-purposing of the analysis tools
- Accuracy and reproducibility of our scientific methods
- Open Access: enabling users to provide public links to SKA science data products in their research publications with DOIs

Others

- What science extraction will take place on laptops or non-SRC compute facilities?
- Are the teams developing pipelines/workflows for precursor/ pathfinders thinking on running them in the future at the SRCs?
- Keep this process open beyond AENEAS (SRCSC?): learnings from pathfinders processing feed into this process. Need to keep up with the latest algorithms, software, hardware, cloud and GPU processing, VR and other visualisation advances.
- Key (computing) staff from supercomputer centres and industry to be involved, incl. computing students/postdocs. Workshops (3-5 people per team + two experts) led to a speed-up of factor 15-20 for the ASKAPsoft pipeline.
- Can the SRCs have a role in that raw visibilities are kept?

My suggestion

- Go to representatives of pathfinders precursors as “users” could work better to answer the more detailed questions (or just more questions)…