SCIENCE DATA CHALLENGE 2: DATA DESCRIPTION

Document Number	SKA-TEL-SKO-0000000
Document Type	DTE
Revision	A
Author	SKAO Science Team
Date	2020-11-26
Document Classification	UNRESTRICTED
Status	Draft

Name	Designation	Affiliation	Signature	
Authored by:				
			Date:	
Owned by:				
			Date:	
Approved by:				
			Date:	
Released by:				
			Date:	

DOCUMENT HISTORY

Revision	Date Of Issue	Engineering Change Number	Comments
A	2017-01-01	-	First draft release for internal review

DOCUMENT SOFTWARE

	Package	Version	Filename
Word processor	MS Word	Word 2007	Document2
Block diagrams			
Other			

ORGANISATION DETAILS

Name	SKA Organisation
Registered Address	Jodrell Bank Observatory Lower Withington Macclesfield Cheshire SK11 9DL United Kingdom Registered in England & Wales Company Number: 07881918
Fax.	+44 (0)161 306 9600
Website	www.skatelescope.org

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

RFI	Radio Frequency Interference
SDC2	Science Data Challenge 2
SKA	Square Kilometre Array
SKAO	SKA Project Office
SRC	SKA Regional Centre

1. Introduction

1.1. Purpose of the document

The purpose of this document is to provide information on how the SKA science data challenge 2 has been produced and to set the challenge for the community.

1.2. Scope of the document

In this document, we describe how we produced the dataset for the SKA science data challenge 2, we set the challenge for the community and describe how submissions will be scored. For rules of participation and other information, see sdc2.astronomers.skatelescope.org.

2. References

2.1. Reference documents

The following documents are referenced in this document. In the event of conflict between the contents of the referenced documents and this document, **this document** shall take precedence.

[RD1] Jones, M. et al. , 2018, MNRAS, 477, 2
[RD2] Duffy et al, 2012, MNRAS 426, 3385
[RD3] Wang, J et al. 2016, MNRAS, 460, 2143
[RD4] Bonaldi, A, et al. 2019, MNRAS, "The Tiered Radio Extragalactic Continuum Simulation, 482, 2
[RD5] Baugh et al., 2019, MNRAS, 483, 4922
[RD6] Leahy, J.P., Bridle, A.H., Strom, R.G. "An Atlas of DRAGNs", http://www.jb.man.ac.uk/atlas/frame.htm
[RD7] Sault R.J., et al, 1995. In *Astronomical Data Analysis Software and Systems IV*, ed. R. Shaw, H.E. Payne, J.J.E. Hayes, ASP Conference Series, 77, 433
[RD8] Anticipated SKA1 Science Performance, SKA-TEL-SKO-0000818
[RD9] Braun 2012, ApJ 749, 87
[RD10] Broeils & Rhee 1997, A&A 324, 877
[RD11] Walter et al., 2008, AJ 136, 2563
[RD12] Heald et al., 2011, A&A 526, 118
[RD13] Guide for Reproducible Research,

https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html
[RD14] https://www.software.ac.uk/

3. SKA Science Data Challenge 2 dataset description

3.1. Introduction

SKA Science Data Challenges will be regularly issued to the community as part of the science preparatory activities. The purpose of these challenges is to inform the development of the data reduction workflows, to allow the science community to get familiar with the standard products the SKA will deliver, and optimise their analyses to extract science from them. These challenges may consist of real data from currently operating radio facilities or of simulated SKA data. Data at different stages along a data reduction workflow have been broadly categorised into four main Data Layers (DL)

- DL1: Raw Data. These are typically observations of a few hour duration, consistent with a single SKA scheduling block. Data are typically uncalibrated and the main focus of a challenge might be to inform the calibration strategy, its implementation, efficiency and scalability.
- DL2: Observatory Data Products. These are calibrated products of different types. The focus of a challenge at this DL is to carry out the kind of processing that will typically be performed at SKA Regional Centres (SRCs) to produce Advanced Data Products.
- DL3: Advanced Data Products. These are the standard SRC outputs, and typically generated by the users. The objective of data challenges at this layer is to extract science from the data, with a focus on algorithm development.
- DL4: Scientific Results. This is a proposal-specific product, that is ultimately the goal of the whole observation and analysis.

Data for SKA Science Data Challenges can be made available at any of the four stages. The objective for each data challenge is defined; however, usage of the data for other purposes is encouraged. Usage of the data beyond the defined challenge should be suitably acknowledged. This dataset can be referenced as SKAO data challenges, Science Data Challenge 2; for other references see Sec. 2.2.

3.1.1. SDC2 dataset

The SDC2 dataset is a DL2 product simulating an HI imaging data cube and the necessary ancillary data. All codes used to generate the dataset will be made publicly available upon the completion of the challenge.

The full-size dataset consists of:

- HI data cube;
- Companion radio continuum data cube;

Details of the simulated HI data product are the following:

- 20 square degrees area;
- 7 arcsec beam size, sampled with 2.8 x 2.8 arcsec pixels;
- 950–1150 MHz bandwidth, sampled with a 30 kHz resolution. This corresponds to a redshift interval z = 0.235–0.495;
- noise consistent with a 2000 hour total observation;
- systematics include imperfect continuum subtraction, simulated RFI flagging and excess noise due to RFI.

The radio continuum product covers the same field of view of the HI data cube, with the same spatial resolution and a 950-1400 MHz frequency range with a 50 MHz frequency resolution.

Together with the full-size challenge dataset, two additional datasets will be made available for development and evaluation purposes. Both are generated using the same procedure of the full-size dataset, but each with a different statistical realization, and covering a smaller portion of the sky. The 'development' data set will be provided along with a truth catalogue of HI sources. The 'evaluation' dataset, provided without a truth catalogue, can be used by teams with a challenge scoring service to evaluate pipelines against the official challenge scoring procedure (Section 7.1).

The development and evaluation datasets will be available for download while the full-size dataset will only be accessible through a network of computational facilities, where participating teams will deploy their software pipelines. The challenge results are scored on the full-size dataset.

4. Sky model description

The sky model was generated primarily using the *Python* scripting language, making use of *astropy*, *scipy* and *skimage* libraries for image and cube generation and *FITSIO* for writing to file. *FORTRAN* was used for the generation of source catalogues.

4.1. HI and Continuum catalogues

A catalogue of HI emission sources has been generated by sampling from an HI redshift-dependent mass function derived from [RD1] for the considered redshift interval (z= 0.235 - 0.495). Conversion from HI mass to flux (Jy/Hz) follows the relation in [RD2] and source sizes are modelled using [RD3].

A catalogue of radio-continuum sources (star-forming galaxies, SFGs and Active Galactic Nuclei, AGN) has been generated with the T-RECS code [RD4] for the frequency interval 950-1400 MHz and with a flux limit of 2 10^{-7} Jy at 1150 MHz. In addition to the model described in [RD4], an HI mass has been associated with each radio continuum galaxy using the [RD5] relation between dark mass and HI mass for AGN, and the correlation between HI mass and star-formation rate for SFGs.

The HI catalogue and the portion of the radio continuum catalogue sharing the same redshift interval have been further processed to identify those that would constitute a counterpart, i.e. be hosted by the same galaxy. To this end, HI galaxies were matched to available radio continuum galaxies having the closest modelled HI mass.

Finally, in order to introduce a realistic clustering signal to the sources, the galaxies have been associated with dark matter (DM) haloes from a high-resolution cosmological simulation [RD5]. Two factors have been considered:

1. the mass of the host: galaxies have been associated with available DM haloes having the closest mass in the same redshift interval;
2. the environment: for HI galaxies, eligible DM haloes in volumes of the cone having local density higher than 50 objects per cubic Mpc were preferentially selected.

4.2. HI cube

HI sources were injected into the field using an atlas of high quality HI source observations. The atlas was collated using samples available from the HALOGAS [RD11] survey, available online, and the THINGS [RD12] survey, made available after the application of multi-scale beam deconvolution. The preparation of atlas sources involved the following steps:

1. Blanking of pixels in order to produce a positive definite noiseless model.
2. Measurement of HI major axis diameter at 1 $M_\odot$ pc^{-2}.
3. Rotation to a common position angle of 0 degrees.
4. Preliminary spatial resampling after application of a low pass filter, such that the physical pixel size of the resampled data would be no lower than required for the lowest redshift simulated sources.
5. Preliminary velocity resampling after application of a low pass filter.

For each source from the simulation catalogue, a source from the prepared atlas of sources was chosen from those nearby in normalised HI mass-inclination angle parameter space. In order to exploit the diversity of the limited atlas sample, matches were selected at random from those atlas sources located within a suitable radius. This radius in parameter space was chosen to be large enough to allow a wide spread of matched altas sources, but small enough not to produce matches which would deviate too far from the desired catalogue HI mass and inclination angle.

Once matched with a catalogue source, atlas sources underwent transformations in size in the three cube dimensions of RA, Dec and frequency, ν. An appropriate low pass filter was applied prior to all scalings, in order to preserve sufficient sampling. Transformation scalings along HI major axis D_{HI}, HI minor axis b, and line width w_{20} were determined using the catalogue source properties of HI mass M_{HI}, inclination angle i, and redshift z, and making use of the following relations:

$D_{\mathrm{HI}} = 0.51 \log M_{\mathrm{HI}} - 3.32$,

from [RD 10], in order to determine spatial scalings for mass;

$V_{\mathrm{rot}} = \mathrm{G} M_{\mathrm{dyn}} / r$,

where V_{rot} is the rotational velocity at radius r and M_{dyn} is the dynamical mass, in order to determine frequency scalings for HI mass;

$\cos^2(i) = (b / D_{\mathrm{HI}})^2 - \alpha^2 / 1 - (\alpha)^2$,

where $\alpha = 0.2$, in order to determine spatial scalings for inclination;

$V_{\mathrm{rot}} = V_{\mathrm{rad}} \sin(i)$,

where V_{rad} is the observed radial velocity, in order to determine frequency scalings for inclination.

Spatial scalings for redshift were determined by calculating the angular diameter distance D_A, assuming a standard flat cosmology with $H_0 = 67$ km s^{-1} Mpc^{-1}, $\Omega_M = 0.32$ and $\Omega_\Lambda = 0.68$. The optical velocity V_{opt} definition,

$V_{opt} = cz = c\ (\nu - \nu_0) / \nu$,

with rest frequency ν_0, has been used throughout and its derivative,

$dZ = (c\nu_0)/(\nu^2))*d\nu$,

used to determine the scalings from velocity sampling dZ to frequency sampling $d\nu$, at a given redshift.

Finally, each transformed object was rotated to its catalogued position angle, scaled according to total integrated HI flux and convolved with a circular Gaussian of 7" FWHM, before being placed in the full HI emission field at its designated position in RA, Dec and central frequency.

4.3. Continuum cube

Continuum sources from the catalogue were injected into the simulated fields either as:

- an extended source, if its major axis size is larger than 3 pixels;
- a compact source, if its size is smaller than 3 pixels.

In the catalogue, the "size" of all AGN populations is defined as the Largest Angular Size (LAS) of the source and that of the SFGs as the exponential scale-length of the disk. For the purpose of determining whether a source is extended or compact, however, a FWHM was considered in both cases.

All compact sources were modelled as unresolved, and added as Gaussians of the same size as the beam size, 7'' FWHM. Images of all extended sources were generated according to their morphological parameters and then added as "postage stamps" to an image of the full field, after applying a Gaussian convolving kernel corresponding to the 7'' beam.

The morphological model for the extended SFGs is an exponential Sersic profile, projected into an ellipsoid with a given axis ratio and position angle. All AGN populations are treated as the same object type viewed from a different angle, so that steep-spectrum AGN exhibit the typical double-lobe FRI/FRII morphology, while flat-spectrum ones exhibit a compact core component together with a single lobe viewed end-on. For the steep-spectrum AGN we used a library of real, high-resolution AGN images [RD6], scaled in total intensity and size, and randomly rotated and reflected, to generate the postage stamps. All flat-spectrum AGN were added as a pair of Gaussian components; one of 7'' scale with a given "core fraction" of the total flux density and another of a specified larger size.

4.4. Calculation of HI Absorption Signatures

Potential sources of HI absorption are determined by calculating the neutral hydrogen column density associated with every pixel in the HI model cube from:

$N_{HI} = 49.8\ S_L(\nu)\ \Delta\nu\ M_\odot\ (1+z)^4 / (N_p\ m_H\ \Delta\theta^2\ C_M^2)$

where S_L is the HI brightness in the pixel (in Jy/beam), $\Delta\nu$ the channel spacing (in Hz), $M_\odot$ a solar mass, z the redshift of the HI 21cm line that applies to this pixel, N_p the number of pixels per spatial beam, m_H the H atom mass, $\Delta\theta$ the spatial pixel size (in radians) and C_M a Mpc (expressed in cm). The preceding constant in the equation follows the flux density to HI mass conversion of [RD2]. The apparent HI column density (when observed with 100 pc or better physical resolution) can be related to an associated HI opacity [RD9], $\tau\,\Delta V$, as:

$N_{HI} = N_0\, e^{-\tau\,\Delta V} + N_\infty\,(1 - e^{-\tau\,\Delta V})$ yielding:

$\tau\,\Delta V = \log[\,(N_\infty - N_0)\,/\,(N_\infty - N_{HI})\,]$

where $N_0 = 1.25 \times 10^{20}$ cm^{-2}, $N_\infty = 7.5 \times 10^{21}$ cm^{-2} and a nominal $\Delta V = 15$ km/s provide a good description to the best observational data in hand. In the current case, the physical resolution is much too coarse (some 10 kpc per pixel) to resolve the individual cold atomic clouds that give rise to significant HI absorption opacity. The apparent column densities per pixel have therefore been subjected to an arbitrary power law rescaling as:

$N_{HI}' = \text{antilog}_{10}\{19 + [\log_{10}(N_{HI}) - 19]\,\beta\}$, if $N_{HI} > 10^{19}$,

with power law index $\beta = 1.9$. This is followed by a hyperbolic tangent asymptotic filtering:

$N_{HI}'' = N_\infty\,[\exp(2\,N_{HI}'/N_\infty) - 1]\,/\,[\exp(2\,N_{HI}'/N_\infty) + 1]$

with N_∞ as above, to avoid numerical problems when solving for the opacity. With this potential opacity in hand, two further tests are considered before its application. First, is a check that the red-shift of any continuum emission source along this line of sight is greater than the red-shift of the HI for the frequency pixel under consideration. To enable this test, an image of the intensity-weighted emission red-shift of the continuum sky model was generated as per §4.3 above. Second, is a check that the brightness temperature of the continuum emission source along this line-of-sight exceeds a threshold, $T_{min} = 100$ K. The corresponding flux density is:

$S_{min} = 7.35 \times 10^{-4}\; T_{min}\;\Delta\phi^2\,/\,\lambda^2$

with $\Delta\phi$ the beam size (in arcsec) and λ the observing wavelength (in cm) yielding S_{min} in Jy/beam. At the angular resolution of this data product this test restricts the occurrence of absorption toward continuum sources brighter than about 4 mJy/beam. This test is introduced to limit the occurrence of HI absorption to only those lines-of-sight where the continuum brightness exceeds a plausible maximum brightness temperature of the HI emission itself.

If both tests are passed, then the signature is calculated as:

$S_{HIA}(\nu) = S_C\,[1 - \exp(-\tau\,\Delta V/dV)]$

where S_C is the continuum model flux density at this frequency and dV is the actual channel sampling in units of km/s.

5. Telescope simulation description

The simulation of telescope sampling effects has been implemented with a sequence of command line calls to tasks of the *miriad* package [RD7]. These calls are embedded within several *Python* scripts that bundle generation of the complete data product. Each plane of the output data product is generated independently of all others, allowing for straightforward multi-processor parallelisation over frequency channels.

5.1. Preprocessing of Continuum Cube

The simulated continuum sky model, $S_C(\nu)$, is provided with an initial frequency sampling of 50 MHz and spatial sampling of 2.8 by 2.8 arcsec. As a first step, the frequency sampling is increased to 10 MHz by interpolation.

A deliberately imperfect version of the continuum sky model, $S_{NC}(\nu)$, is then generated for each plane at 10 MHz frequency sampling by generating a two dimensional Gaussian noise image (with unit dispersion) that is uncorrelated from pixel to pixel at a spatial sampling of 512.2 by 512.2 arcsec. A unique random number seed is used for each image plane to ensure that the noise field of each is uncorrelated with the others. This coarsely sampled noise field is interpolated up to the 2.8 by 2.8 arcsec sampling of the sky model. The imperfect continuum model at each plane is formed by multiplying the perfect model with $(1+\sigma N)$ where σ is an assumed RMS gain calibration error, here chosen to be $\sigma = 1\text{x}10^{-3}$ and N is the finely sampled noise field image.

When combination of the continuum with the HI emission and absorption signatures is undertaken (as described below) a further interpolation in frequency is done, of both the perfect and imperfect continuum models , to the final required sampling, of 30 kHz in this case. The imperfect continuum model then has errors that are correlated over 10 MHz in frequency and 512 arcsec spatially. This is intended to represent the residual bandpass calibration errors that might result from the typical spectral standing wave pattern of an SKA dish at these frequencies together with the angular scale over which direction dependent gain differences might be apparent.

5.2. Net Emission/Absorption Cube

With all signatures in hand the net continuum subtracted input cube is calculated from the sum:

$$S(\nu) = S_L(\nu) + S_C(\nu) - S_{NC}(\nu) - S_{HIA}(\nu)$$

where the explicit frequency dependence is included to stress that all quantities are evaluated at the final required frequency sampling.

5.3. Calculation of Effective PSF and Noise Level

The synthesized telescope beam is based on a nominal 8 hour duration tracking observation of the complete SKA1-Mid configuration that is sampled at one minute intervals (to make beam calculation sufficiently realistic yet minimising calculation overheads), while the thermal noise level is based on nominal system performance [RD8] and an effective on-sky integration time of 2000 hours distributed uniformly over the 20 deg^2 survey field. The effective integration time per unit area of the survey field increases toward lower frequencies in proportion to wavelength squared, due to the variation in the primary beam size in conjunction with an assumed survey sampling pattern that is fine enough to provide a uniform noise level even at the highest frequency channel in the data

product, so that the nominal RMS noise, σ_N, level declines linearly with frequency between 950 and 1150 MHz.

The nominal noise level and 8 hour sampling are then modified by consideration of an actual total power spectrum obtained with the MeerKAT telescope while observing the South Celestial Pole. This total power spectrum provides an estimate of the system noise temperature floor of the MeerKAT receiver system as a function of frequency together with any excess average power (during the four hour observation) due to detected Radio Frequency Interference (RFI) at the SKA1-Mid site. The ratio of excess RFI to system noise temperature, γ_{RFI}, is used to scale the nominal SKA1 noise in each frequency channel and determine the degree of simulated RFI flagging to apply to the nominal visibility sampling.

Both the duration of RFI flagging and the maximum baseline to which flagging extends (starting in all cases from B_{MIN} = 0) are derived from the RFI noise ratio, γ_{RFI}. The maximum baseline (in units of wavelengths) to be flagged is given by,

$B_{Max} = 71 \text{ antilog}_{10}[(\gamma_{RFI} - 1)^{1/3}]$

which yields B_{MAX} = <15m to about 10km for the range of RFI noise ratios actually encountered at the relevant observing frequencies. The duration of RFI flagging, ΔHA, in hours is determined from,

$$\begin{aligned} \Delta HA &= 0 && \gamma_{RFI} < \gamma_{MIN} \\ &= 8\,(\gamma_{RFI} - \gamma_{MIN})/(\gamma_{MAX} - \gamma_{MIN}) && \gamma_{MIN} > \gamma_{RFI} > \gamma_{MAX} \\ &= 8 && \gamma_{RFI} > \gamma_{MAX} \end{aligned}$$

where we take γ_{MIN} = 1.1 and γ_{MAX} = 2, to define the range of RFI ratios over which the flagging is intermittent, rather than either absent or continuous. This flagging interval is placed randomly within the tracking window. After application of any flagging to the simulated visibilities within the nominal HA = -4h to +4h tracking window at each observing frequency, the synthesized beam and corresponding "dirty" noise image are generated. A unique random number seed is provided for each frequency channel to ensure that the resulting noise fields are not correlated across frequency. During imaging, a super-uniform visibility weighting algorithm is employed that makes use of a 64x64 pixel FWHM Gaussian convolution of the gridded natural visibilities in order to estimate the local density of visibility sampling. The super-uniform re-weighting is followed by a Gaussian tapering of the visibilities to achieve the final target dirty PSF properties, namely the most Gaussian possible "dirty" beam with 7x7 arcsec FWHM. The effective PSF is then modified to account for the fact that the survey area will be built up via the linear combination of multiple, finely spaced, telescope pointings on the sky. The effective PSF in this case is formed from the product of the calculated dirty PSF with a model of the telescope primary beam at this frequency, as documented in RD8.

The "dirty" noise image produced as described above is rescaled to have an RMS fluctuation level, σ_i, corresponding to the nominal sensitivity level for this channel degraded by the RFI noise ratio, γ_{RFI},

$\sigma_i = \sigma_N \gamma_{RFI}$

5.4. Simulated Sampling and Deconvolution

The “net emission/absorption cube” described in §5.3 is then subjected to simulated deconvolution and residual degradation by the relevant synthesized “dirty” beam. All features, both positive and negative, that deviate from zero by more than three times the local noise level, $3\sigma_i$, are extracted as a “clean” image and replaced by that threshold to form a residual sky image. The residual sky image is subjected to a linear deconvolution (via FFT division) with a 7x7 arcsec Gaussian, truncated at 10% of the peak and then convolved with the “dirty” beam. The final data product cube is formed from the sum of these “dirty” residuals, the previously extracted “clean” feature image and the “dirty” noise image for this channel.

6. Limitations of the simulated data products

While significant effort has been expended to make a realistic data product for the SDC2 analysis, there are many limitations to the degree of realism that could be achieved. Some of the most apparent are outlined below.

6.1. Catalogue Limitations

The HI catalogue and the continuum catalogue are modelled and generated independently, and subsequently associated with a DM host (which in some cases can have both an HI and a continuum luminous counterpart) by means of relatively simple recipes. The full complexity of the sky for what concerns the environmental dependencies and detailed astrophysics of galaxies is therefore not captured.

6.2. Continuum Emission Model Limitations

The morphological model for continuum sources is simpler than the real sky. SFGs and flat-spectrum AGN sources adopt a simple profile; steep-spectrum sources are based on real images however the morphological diversity is limited by the size of the atlas we adopted. Future improvements could involve the use of source generation via machine learning techniques, in order to broaden the diversity of sources.

6.3. HI Emission Model Limitations

The HI emission model has been generated using a relatively small sample of real HI observations. Efforts to maximise the diversity of objects have ensured as wide a spread as possible in object selection from the limited sample, without compromising object realism during image cube scalings for mass and inclination. Future improvements could involve the use of source generation via machine learning techniques, in order to broaden the diversity of sources.

We note that our treatment of the published HIMF and the HI atlas cubes that we introduce into the simulation are based on the assumption of negligible HI self-opacity. Although this is a widespread assumption in the current literature it is unlikely to be the case [eg RD9]. The integral of HI emission gives a lower limit to the actual HI mass. What little data is available at present suggests the magnitude of underestimate may be 40 to 50% in the local universe. While it is unclear how that

might change with redshift, the enhanced occurrence of molecular gas in high redshift galaxies suggests systematically higher pressures in galaxy disks, which would also promote a larger proportion of self-opacity within the HI.

6.4. HI Absorption Model Limitations

The mechanism for introducing HI absorption signatures into the data product is known to be quite crude. The major limitation to realism is a consequence of undertaking the assessment of the physical properties along each line of sight at an extremely coarse physical sampling, namely the data product spatial pixel size, of about 10 kpc, and to a lesser degree the frequency sampling that corresponds to about 9 km/s. A more appropriate sampling to match the relevant physical scales where the HI absorption arises would be better than about 100 pc and 1 km/s. A real HI object atlas with these attributes does not exist, so it would be necessary to rely on very high resolution simulations of HI galaxies, that include realistic tracking of both the locations and proportions of warm HI (with T_k ~ 10,000 K) which gives rise to negligible absorption and the cool HI (with T_k ~ 100 K) that is necessary to yield absorption. Such simulations are only beginning to be undertaken. This resolution mismatch has necessitated the artificial non-linear enhancement of peaks in the apparent column density distribution in order to generate any detectable absorption features at all. For this reason, the rate of occurrence of such features in the data product can not be considered as being self-consistent with the modeled distribution of HI emission nor have any real predictive power for what might be encountered in an actual observation of this type.

A related limitation arises from the low resolution (7 arcsec FWHM) being applied to the background continuum radiation toward which a detectable HI absorption signature might occur. At this relatively low resolution the apparent brightness temperature of typical continuum sources (of mJy brightness) is so low (T_B < 100 K) that it may not exceed the brightness temperature of the HI emission signature if the emission comes close to filling the beam. Absorption and emission signatures can then not be distinguished. We have therefore artificially restricted the occurrence of absorption signatures to only those continuum sources that exceed this brightness limit (and satisfy the other requirements noted in §4.4). Increasing the spatial resolution at which the analysis is undertaken would allow much fainter continuum sources to be considered as absorption candidates. This phenomenon applies to actual observations of HI absorption signatures and argues for utilising the highest possible angular resolution for such work to circumvent mixing of emission and absorption signatures.

6.5. Telescope Sampling Limitations

The most significant limitation of the adopted approach to simulate telescope sampling has been the use of an image plane convolution of the sky model to approximate the actual imaging process. Injection of the sky model into two thousand hours of suitably sampled visibility data (7.4 PBytes with 1s time sampling), followed by simulated calibration and imaging is well beyond our current capabilities. One implication of this approach is that the scope for introduction of calibration errors and other forms of systematic errors into the data product has been quite limited.

Even the image plane convolution is undertaken with a synthesized beam model that has been sampled only every 60s, rather than the 1s that would be better matched to eliminate significant smearing effects (at 7 arcsec resolution) for sources in the sidelobes of the telescope primary beam for actual visibility data. This coarser time sampling for the image plane convolution approach was adopted after extensive testing to determine that the artificial radial PSF sidelobes that are its direct

consequence have a negligible impact on the residual “dirty” sky model that is generated with this PSF.

Moreover, the process of sampling the survey region with a dense grid of individual telescope pointings, as would be undertaken for an actual observation, has also not been simulated directly. Instead a spatially uniform noise floor and a single effective Point Spread Function (PSF) were used to approximate the outcome of such a data acquisition, imaging and combination strategy. The implication is that some classes of errors, such as telescope pointing errors could not be directly simulated. On the other hand, account has been taken of the systematic variation of effective integration time with frequency that accompanies mosaicking as well as the effective PSF that applies to the survey region, at least under the assumption that a similar PSF and sensitivity are achieved within each pointing direction of such a survey.

The simulation of RFI effects in the data product is also of limited realism. While the frequency dependence of such effects has been guided by actual site specific data, and the magnitude of the incoherent receiver noise enhancement is likely to be plausible, the treatment of RFI impact on the visibility data is very crude. In reality the visibilities are affected by a correlated noise component that pertains to the detection of the individual RFI sources that are typically at very large angular distances from the pointing direction. The degree of correlation diminishes with both time and frequency smearing for these extreme off-axis emitters. Rather than modeling these signatures in detail we have merely introduced the arbitrary data flagging prescription outlined in §5.3, and not included any residual unflagged RFI signatures into the data product. There is certainly scope for improving the simulations in this respect.

7. The challenge defined

Participating teams are invited to access the full-size data cube on dedicated facilities provided by our computational resource partners. Teams will then undertake:

1. **Source finding**, defined as the location in RA (degrees), Dec (degrees) and central frequency (Hz) of the dynamical centre of each source.

2. **Source property characterisation**, defined as the recovery of the following properties:
    a. Integrated line flux (Jy Hz): the total line flux integrated over the signal $\int F \, \mathrm{d}\nu$
    b. HI size (arcsec): the HI major axis diameter at 1 $M_\odot$ pc^{-2}.
    c. Line width (km s^{-1}): the observed line width at 20 percent of its peak.
    d. Position angle (degrees): the angle of the major axis of the receding side of the galaxy, measured anticlockwise from North.
    e. Inclination angle (degrees): the angle between line-of-sight and a line normal to the plane of the galaxy.

Submissions must strictly adhere to a specific format, which will be detailed in a set of instructions and demonstrated by an example file.

7.1. Scoring

Results will be submitted via a dedicated scoring service, which compares each submission with the catalogue of truth values and returns a score. For the duration of the challenge, scores can be updated at any time; the outcome of the challenge will be based on the final scores submitted by each team.

The overall score is determined as follows:

1. Each entry in the submitted catalogue is recorded as a detection.
2. All detections are cross-matched with a truth catalogue; those detections with positions in RA, Dec and central frequency within range of a source in the truth catalogue are recorded as matches. For each truth source, the range in the spatial and frequency dimensions within which a match is defined is determined by the beam-convolved HI size and the line width, respectively.
3. Detections that lie outside the range of a truth catalogue source are recorded as false positives.
4. The properties of all matched sources are scored for accuracy.
5. The accuracy scores for each property are weighted by the number of properties and are summed, such that the maximum total score for a single matched source is 1.
6. Scores for matched sources belonging to the same truth source are further summed and divided by the number of duplicate matches to generate a unique matched source score per truth source.
7. The final score is determined by subtracting the number of false positives from the summed scores of all unique matched sources.

7.2. Reproducibility awards

Alongside the main challenge, teams will be eligible for 'reproducibility awards', which will be granted to all teams whose processing pipelines demonstrate best practice in the provision of reproducible methods and Open Science. An essential part of the scientific method, reproducibility leads to better, more efficient science [RD13]. Open Science generalises the principle of reproducibility, allowing previous work to be built upon for the future.

Reproducibility awards will be run in parallel and independently from the SDC2 score, and there is no cap on the number of teams to whom the awards can be given.

A checklist developed in partnership with the Software Sustainability Institute (SSI) [RD14] will be provided to teams for the purposes of self-assessment during the challenge, and will be used periodically to evaluate all teams and award those who meet bronze, silver or gold standards of reproducibility best practice.

Acknowledgements

We would like to thank members of the SKA HI SWG for useful feedback. The simulations make use of data from WSRT HALOGAS-DR1. The Westerbork Synthesis Radio Telescope is operated by

ASTRON (Netherlands Institute for Radio Astronomy) with support from the Netherlands Foundation for Scientific Research NWO. The work also made use of 'THINGS', the HI Nearby Galaxy Survey (Walter et al. 2008), data products from which were kindly provided to us by Erwin de Blok after multi-scale beam deconvolution performed by Elias Brinks.