AENEAS- WG5
Access and Knowledge creation

Marcella Massardi
(INAF- IRA / Italian node of the European ARC)

*This work package (WP5) is focused on the **interface between a distributed European SKA Data Centre (ESDC) and a distributed body of end users** whose goal is the exploitation of SKA data for knowledge creation. WP5 will therefore study the design of "user interaction models" that could be implemented for the ESDC*

Task 5.1 Survey of existing user interaction models for large-scale radio astronomy facilities and **integration of WP5 outputs into consolidated ESDC design study** (responsible M. Massardi)

Task 5.2 Recommendations for the design of user interfaces for data discovery, access, and retrieval (responsible R. Smareglia)

Task 5.3 Recommendations for the design of user interfaces for data processing, reprocessing, analysis, and visualization (responsible A. Costa)

Task 5.4 Integration with VO Interoperability Framework (responsible C. Knapic)

Task 5.5 Recommendations for the resourcing of an ESDC user interaction model (responsible J. Brand)

Task 5.6 Recommendations for a plan of user community formation and knowledge distribution (responsible M. Massardi)

Deliverable (number)	Deliverable name	Work package number	Short name of lead participant	Type	Dissemination level	Delivery date (in months)
D5.1	Survey report	5	INAF	R	PU	18
D5.2	Gap analysis	5	INAF	R	PU	18
D5.3	Design recommendations #1	5	INAF	R	PU	24
D5.4	Design recommendations #2	5	INAF	R	PU	24
D5.5	Applicability of VO framework	5	INAF	R	PU	28
D5.6	User interaction model resourcing	5	INAF	R	PU	28
D5.7	Growing the ESDC community	5	INAF	R	PU	28
D5.8	Final integration of WP5 materials	5	INAF	R	PU	34

Questions for Astronomical facilities

The Square Kilometre Array will enable transformational science across a wide range of research areas. By the same token, the large scale, rate, and complexity of data the SKA will generate present challenges in data management and computing that are similarly world-leading. Based on current projections, the SKA Observatory, once operational, is expected to produce an archive of standard data products with a growth rate on the order of 300 petabytes per year. Although the challenges associated with populating and maintaining the SKA science archive are already impressive, these data products actually represent only the first part of the full science extraction chain. Any further processing and subsequent science extraction by users will require significant, additional scientific, computing and storage resources in the form of a federated, global network of SKA Regional Centres.

Sent to the responsibles of LOFAR, MWA, ATCA, JVLA, PdBI, VLBA ...
European radiotelescopes, VLBI Networks

(Only 2 replies so far)

Services provided and interaction policies

1. How do you define a user of your facility and what are the requirements to be a user?

Your answer

2. Please provide a summary of how users interact with your facility and the staff at your facility to generate science outputs

Your answer

3. Please list the services your facility provides (including observing, online tools, archive, user support...)?

Your answer

4. Can you provide links to facility policies that are relevant for users (e.g. proposal submission, time allocation policies, data access policies, policies describing the level of support from the facility for users to extract science from their data, procedures for publications and outreach)

Quantification of user community and products

6. How many users do your facility have?

- ◯ <100
- ◯ 100-500
- ◯ 500-1000
- ◯ 1000-2500
- ◯ >2500

7. How many proposals per call did you receive, on average in the last 3 calls for proposals?

	<50	50-100	100-500	500-1000	>1000
Proposals	◯	◯	◯	◯	◯

8. How many of these are accepted (including fillers) and, how many of the accepted are observed (including fillers) ?

	<30%	30-50%	50-70%	70-90%	>90%
Accepted	◯	◯	◯	◯	◯
Observed	◯	◯	◯	◯	◯

11. What do you estimate is the most resource-consuming aspect of supporting users of your facility, and why?

Your answer

12. Rate the driver for the cost of your user interaction model (1=strong, 5=weak)?

	1	2	3	4	5
data storage	○	○	○	○	○
computing	○	○	○	○	○
number of projects	○	○	○	○	○
number of staff required	○	○	○	○	○
number of users	○	○	○	○	○
software and analysis tools	○	○	○	○	○
Other(specify)....	○	○	○	○	○

13. How many FTE staff does your facility employ (or utilize) to implement your facility's user interaction model (include FTEs associated with computing centres that support your facility in an in-kind fashion, if appropriate) divided in the following categories?

<10	10-20	20-50	50-100	>100

Analysis of cost drivers

Staff skills and activities

16. Rate the following skills that the people involved in your facility's user interaction need (1=necessary, 5=less important)

	1	2	3	4	5
Knowledge of the facility itself (available instruments, observing procedures, calibration requirements)	☐	☐	☐	☐	☐
Knowledge of the science carried out with the facility	☐	☐	☐	☐	☐
Knowledge of the software required to observe, calibrate, image	☐	☐	☐	☐	☐
Knowledge of the use of large-scale computing facilities	☐	☐	☐	☐	☐
Knowledge of managing big dataset	☐	☐	☐	☐	☐

17. What fraction of the total FTE is used for each of the following activity?

	<10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%
Archive maintenance	○	○	○	○	○	○	○	○
Data processing and quality assessment	○	○	○	○	○	○	○	○

9. What volume of raw data does your facility collect per year?

	<10 GB	10GB-1TB	1TB-100TB	100TB-500TB	>500TB
Raw data	○	○	◉	○	○

VLBA

10. Which fraction of it is made accessible through the archive?

	<10%	10-50%	50-70%	70-90%	>90%
In raw data	○	○	○	○	◉
In fully calibrated data	◉	○	○	○	○
In advanced products	◉	○	○	○	○

11. What do you estimate is the most resource-consuming aspect of supporting users of your facility, and why?

Training new users. This is very time consuming of experts. New users continue to make mistakes that are usually caught by data analysts.

Questions for Users of Astronomical facilities

The Square Kilometre Array will be one of the world's most powerful radio telescopes and enable transformational science across a wide range of research areas. By the same token, the large scale, rate, and complexity of data the SKA will generate present challenges in data management, computing, and networking that are similarly world-leading. Based on current projections, the SKA Observatory, once operational, is expected to produce an archive of standard data products with a growth rate on the order of 300 petabytes per year. Although the challenges associated with populating and maintaining the SKA science archive are already impressive, these data products actually represent only the first part of the full science extraction chain. Any further processing and subsequent science extraction by users will require significant, additional computing and storage resources.

Sent to the user communities of the astronomical facilities and to members of SKSP

76 replies in the first 4 hours ...122 in 5 days (including we)

Please indicate whether you work at a European or non-European institute *

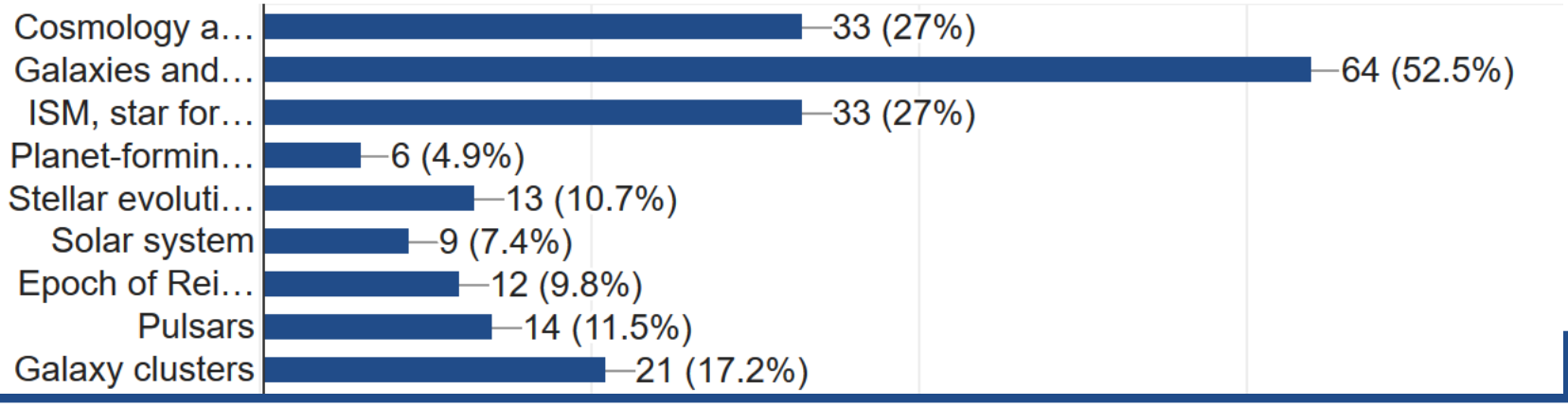
- ○ European (including UK and non EC Countries)
- ○ Non-European
- ○ Both

Contact details (optional): If you are happy for us to get in touch with you about this survey, please give contact details here. These will not be shared outside the AENEAS project.

Your answer

Select your main areas of scientific interest *

- ☐ Cosmology and the high redshift universe
- ☐ Galaxies and galactic nuclei
- ☐ ISM, star formation and astrochemistry



Part 1: user interests

80% in European institutes

Mostly radioastronomers

95% interested in SKA

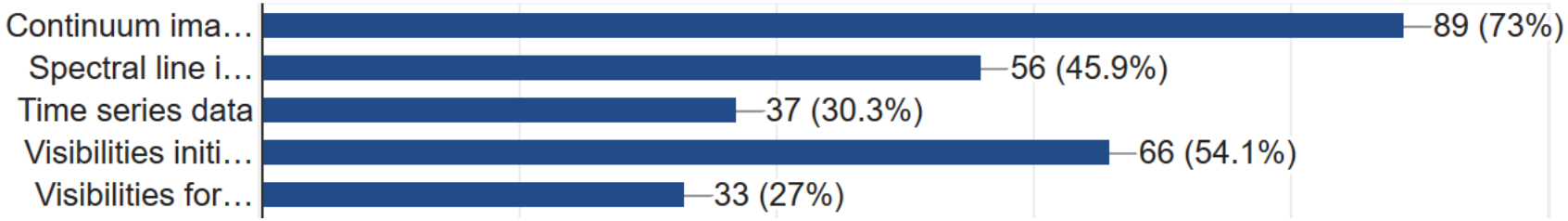
Mostly extragalactic science

Part 2: user habits in data handling

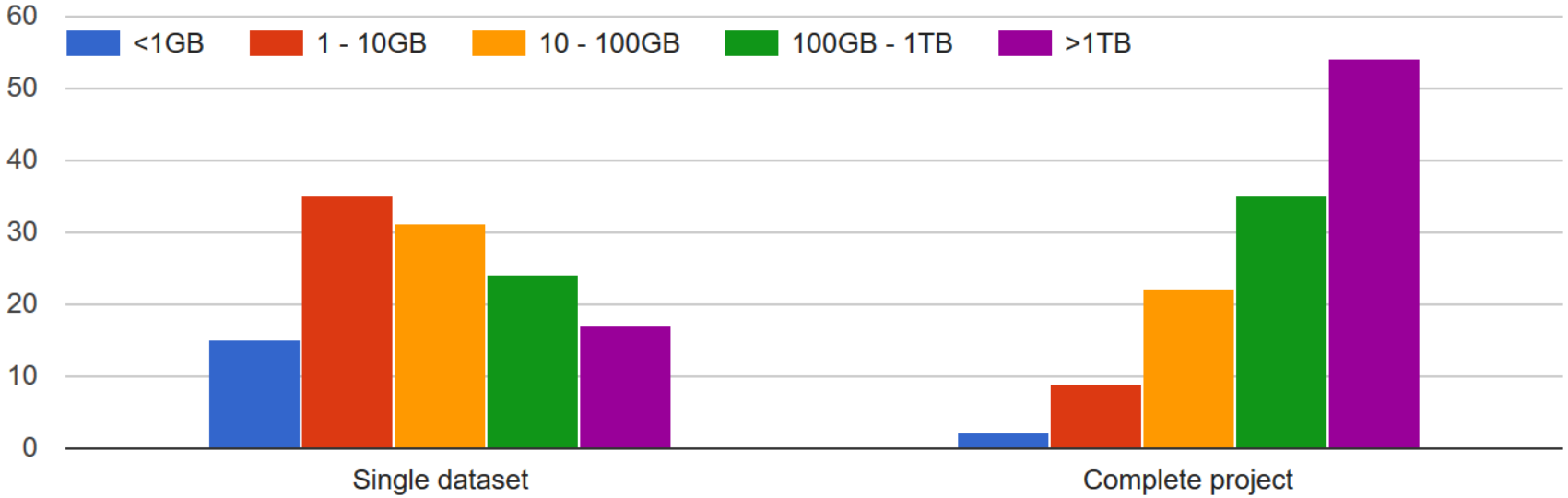
Mostly aiming at images

Which kind of data do you typically handle to extract the information you need? (you may select more than one) *

- [ ] Continuum images
- [ ] Spectral line images
- [ ] Time series data
- [ ] Visibilities initially, but I make images from them
- [ ] Visibilities for uv-plane analysis only

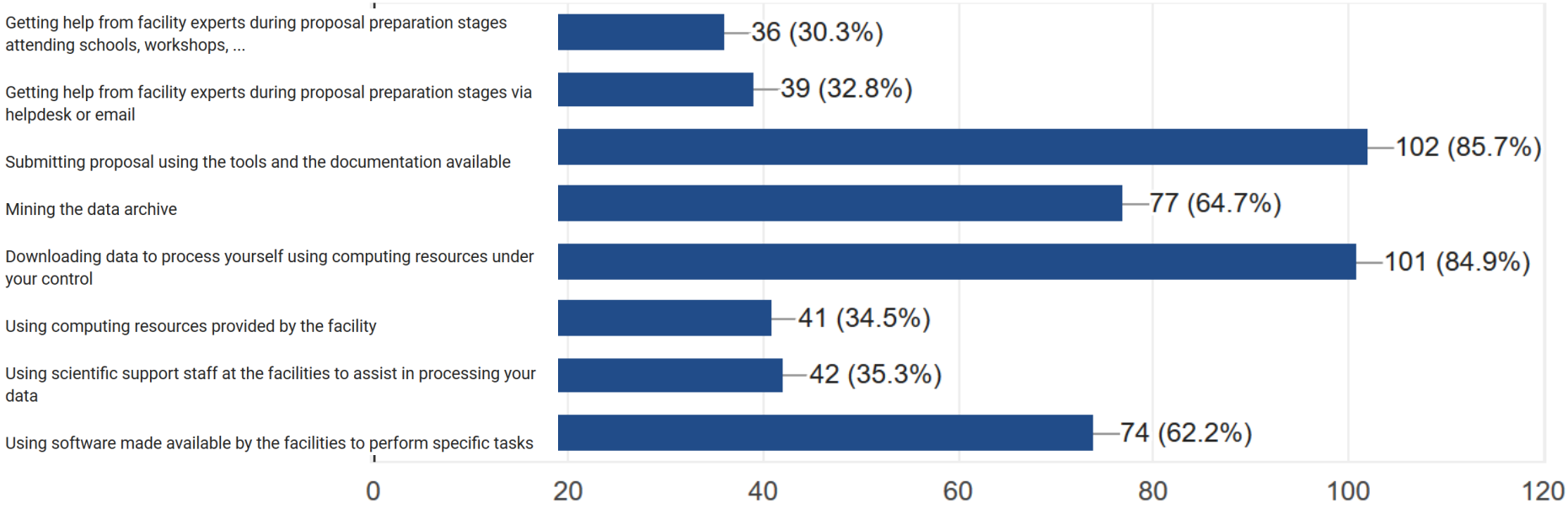


What is the typical size of the data you handle? *



How do you typically interact with the facilities listed above?

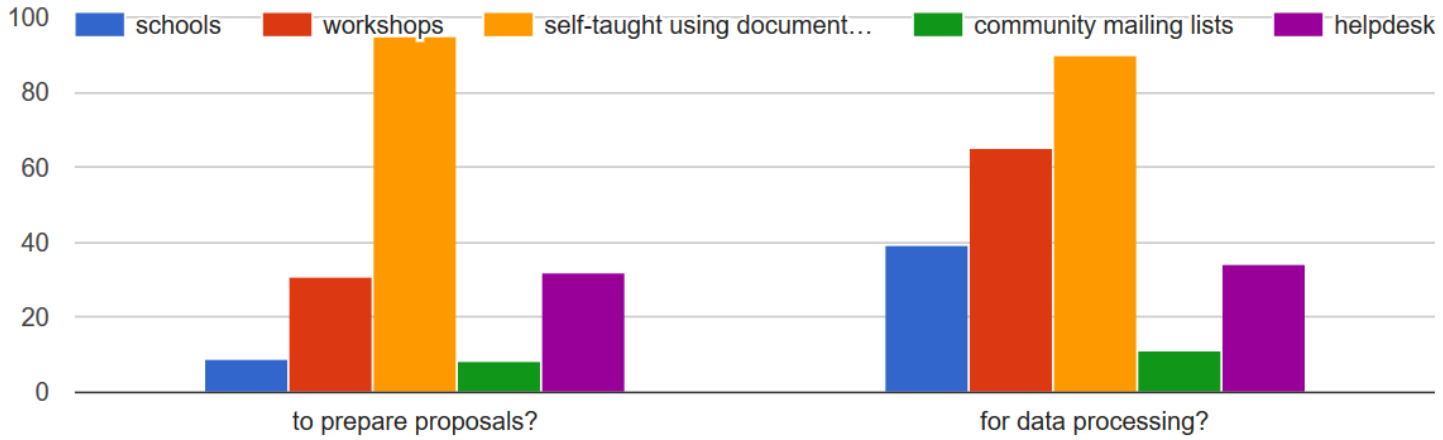
119 responses



Mostly submitting proposals and downloading data to process using computing resources under user's control

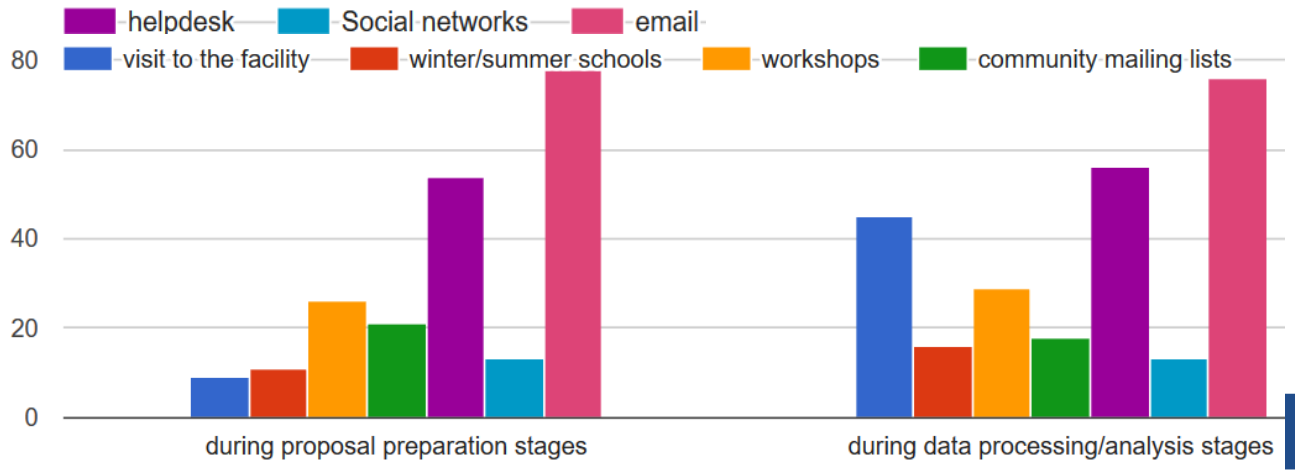
What is your preferred way of training...

Self-taught...

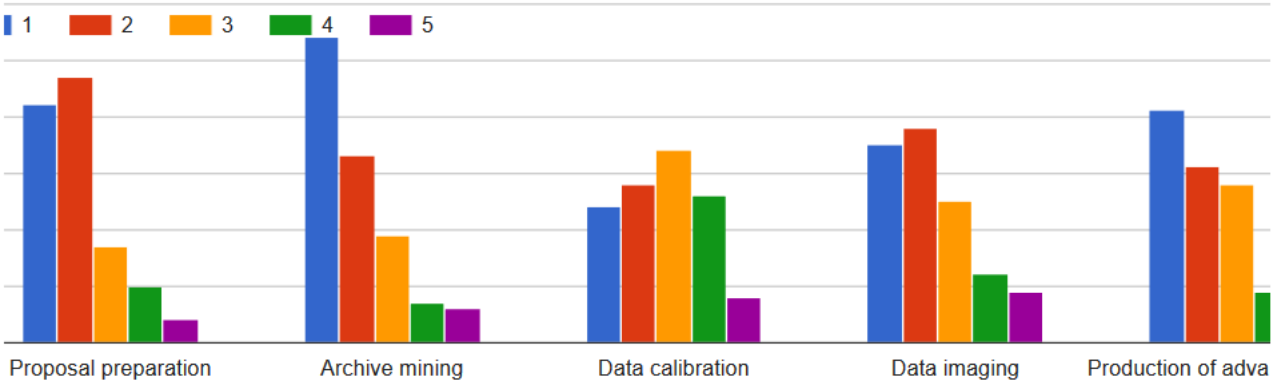


How do you typically get help from facility experts

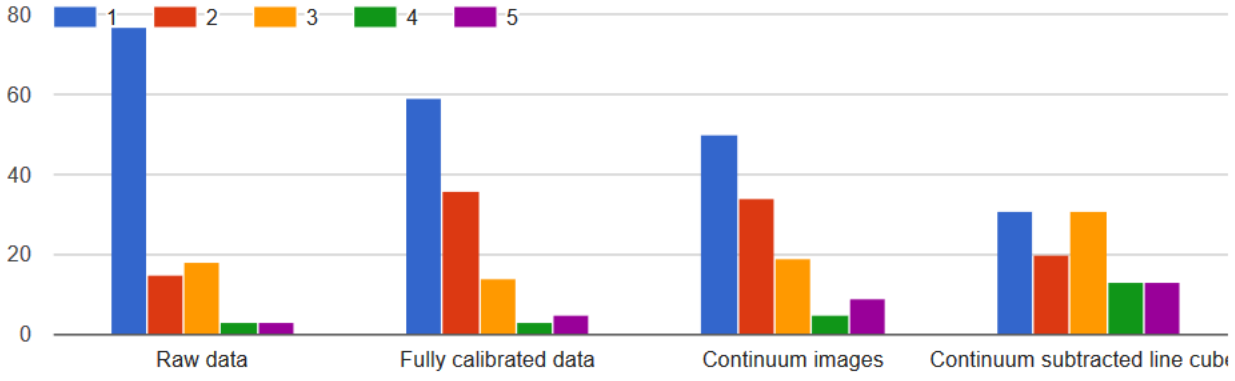
If needed I write an email to an expert



Evaluate the level of support from a facility expert you usually need for each stage of a project (1=none, 5=fully performed by the expert)



Rate what would you like to find in an archive (1=necessary, 5=useless)



Give me the raw data and I can get my own science!!!

Please feel free to add any other comments based on your experiences that are relevant to future users of SKA (optional).

The large datasets from SKA precursors cause a qualitative change in data processing issues which breaks many previous solutions, specifically the parts of the process which traditionally rely on active work by the astronomer (flagging, tweaking calibration parameters, diagnosing problems with the calibration). ***The SKA plan is to fully automate these components which would be excellent if it works but automated pipelines have been promised for each new radio facility since the early 1970s and have yet to prove completely satisfactory****.*

*It seems inevitable that the concept of "****bringing your code to the data****" rather than the other way around needs to spread throughout astronomy, but with a few exceptions,* ***most of astronomy is not ready for this and it is much more common to bring the data to individual computers for analysis using self-built tools****. To change that will require it to be possible to keep the data on* ***some supercomputer*** *somewhere, but have* ***a great flexibility*** *in user ability to write their own code and scripts to interact with it if it is impossible to do so on a personal laptop.*

The survey does not seem to meaningfully reflect the potential or present interaction of users with experts, such as is found in the European ALMA regional center node network (which is a very useful distribution and access to expert support!). ***A distributed node network appears to be a very powerful way to reach out to and to support the community (both expert users and novice users).***

Which is the real GAP?

Data size?

New tools to be built?

Network velocity?

User mentality???

Let’s discuss tomorrow in the

Scalability discussion

Next steps

Task 5.1 Analysis of survey results and gaps

Task 5.2 Test interfaces for user data access and retrieval
(tests on-going for ALMA Re Imaging ESO Study)

Task 5.3 Summary of existing models for data processing and visualization
(document being drafted)

Task 5.4 Investigation of VO applicability to various interfaces

Task 5.5 Analysis of facility survey responses for staffing

Task 5.6 Plan knowledge distribution that could start already with informing about AENEAS efforts (and reducing the community stress when talking about SKA data handling and reduction:
astronomers are really terrified they have to learn how to reduce SKA data!!!)