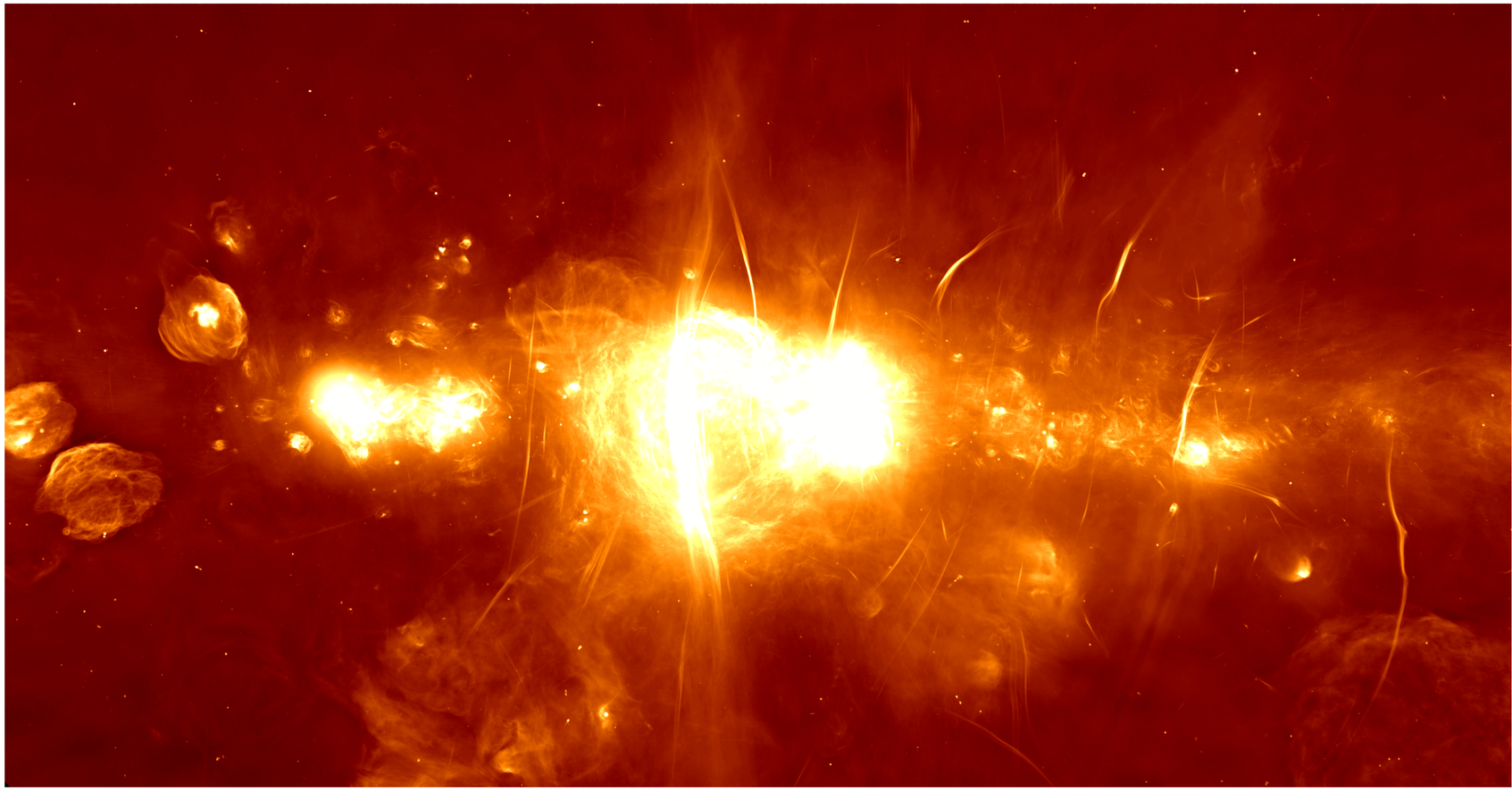
Evaluating Rucio for an SKA Regional Centre

AENEAS : An SKA Regional Centre for Europe

Develop a concept and design for a distributed, federated European Science Data Centre (ESDC) to support the astronomical community in achieving the scientific goals of the Square Kilometre Array

Rohini Joshi

University of Manchester

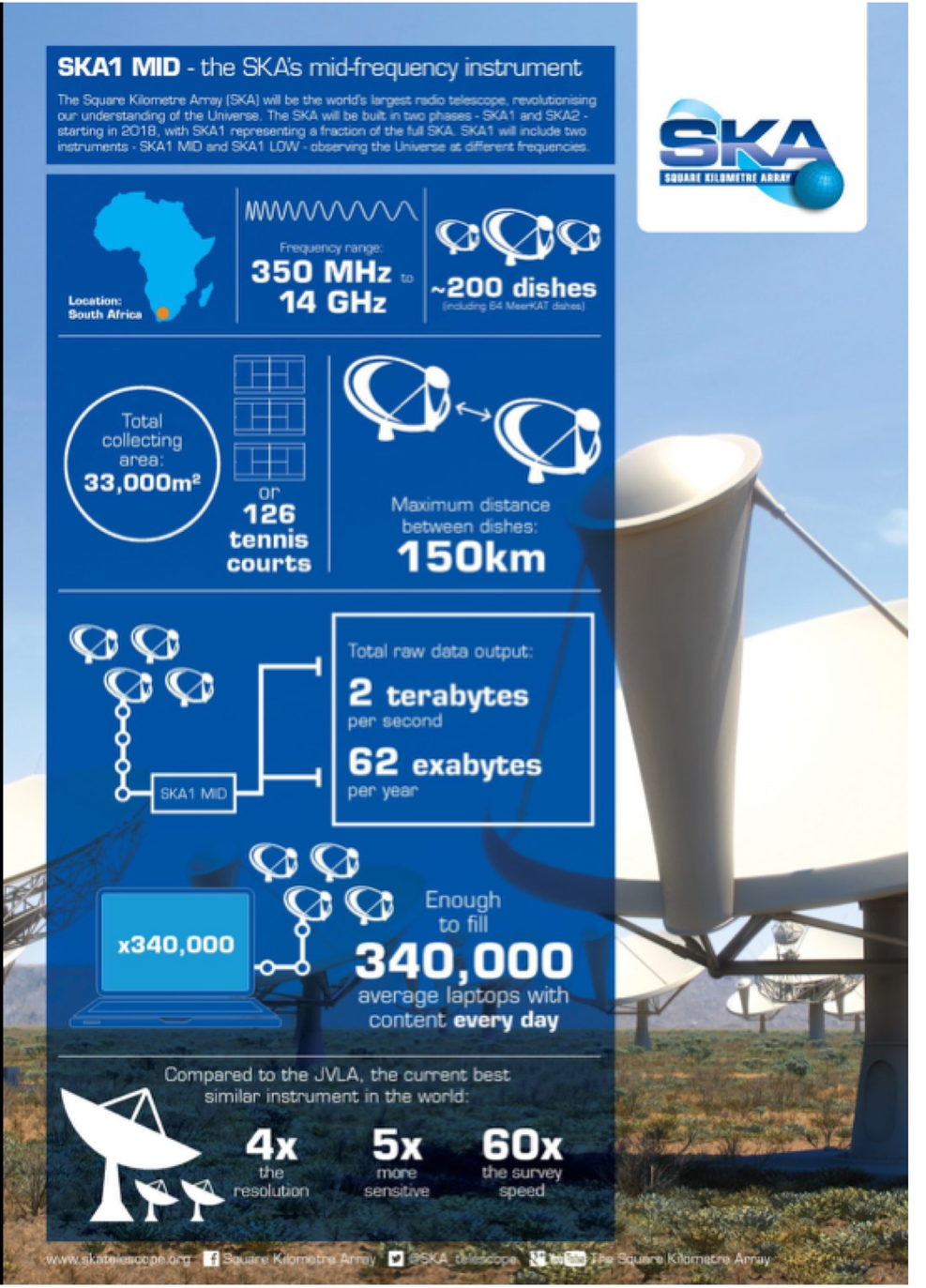


Credit: SARAO

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

DISH ARRAY

SKA1 MID

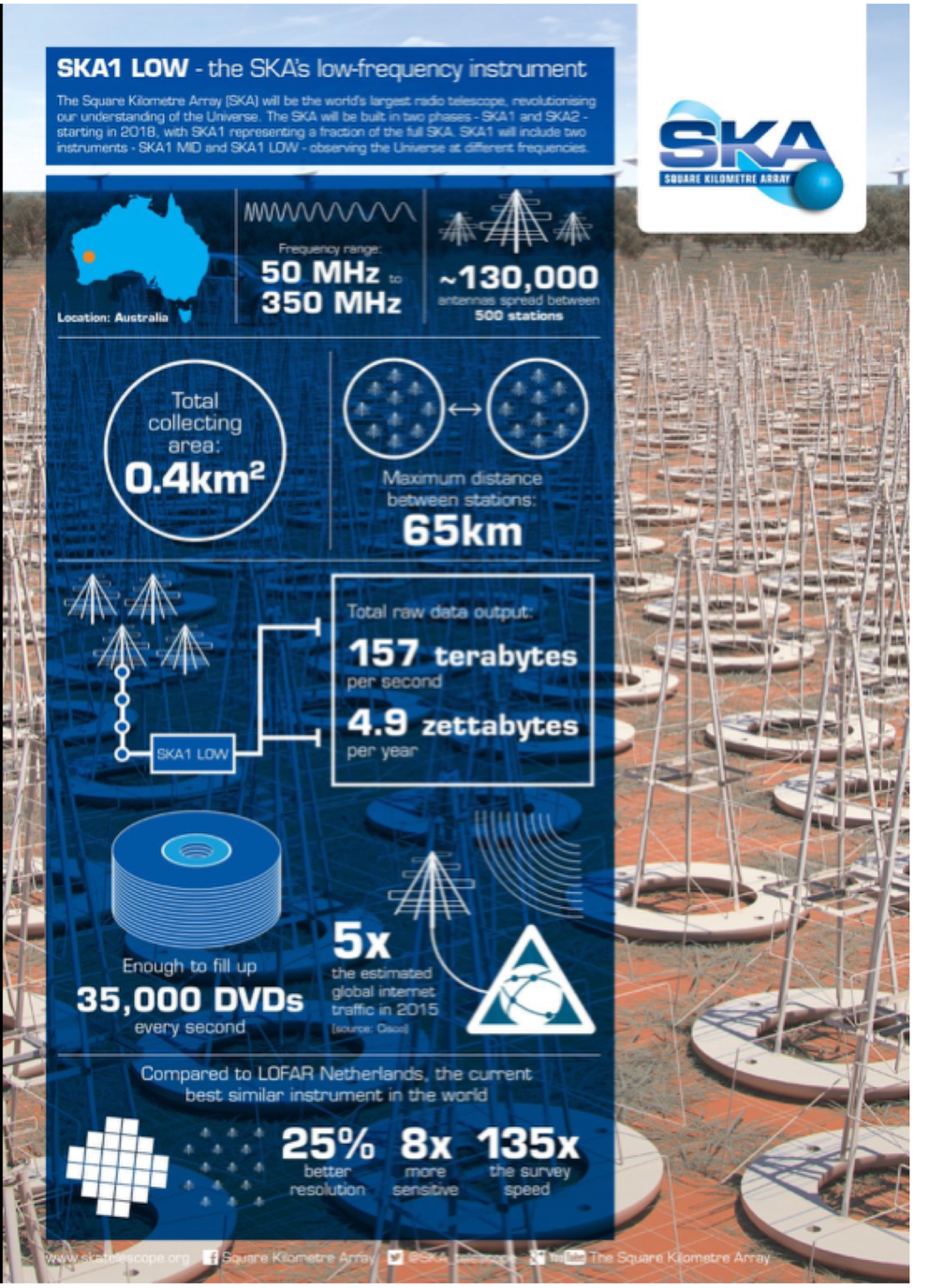


APERTURE ARRAY

SKA1 LOW

TERABYTE = 10^{12} BYTES

ZETTABYTE = 10^{21} BYTES



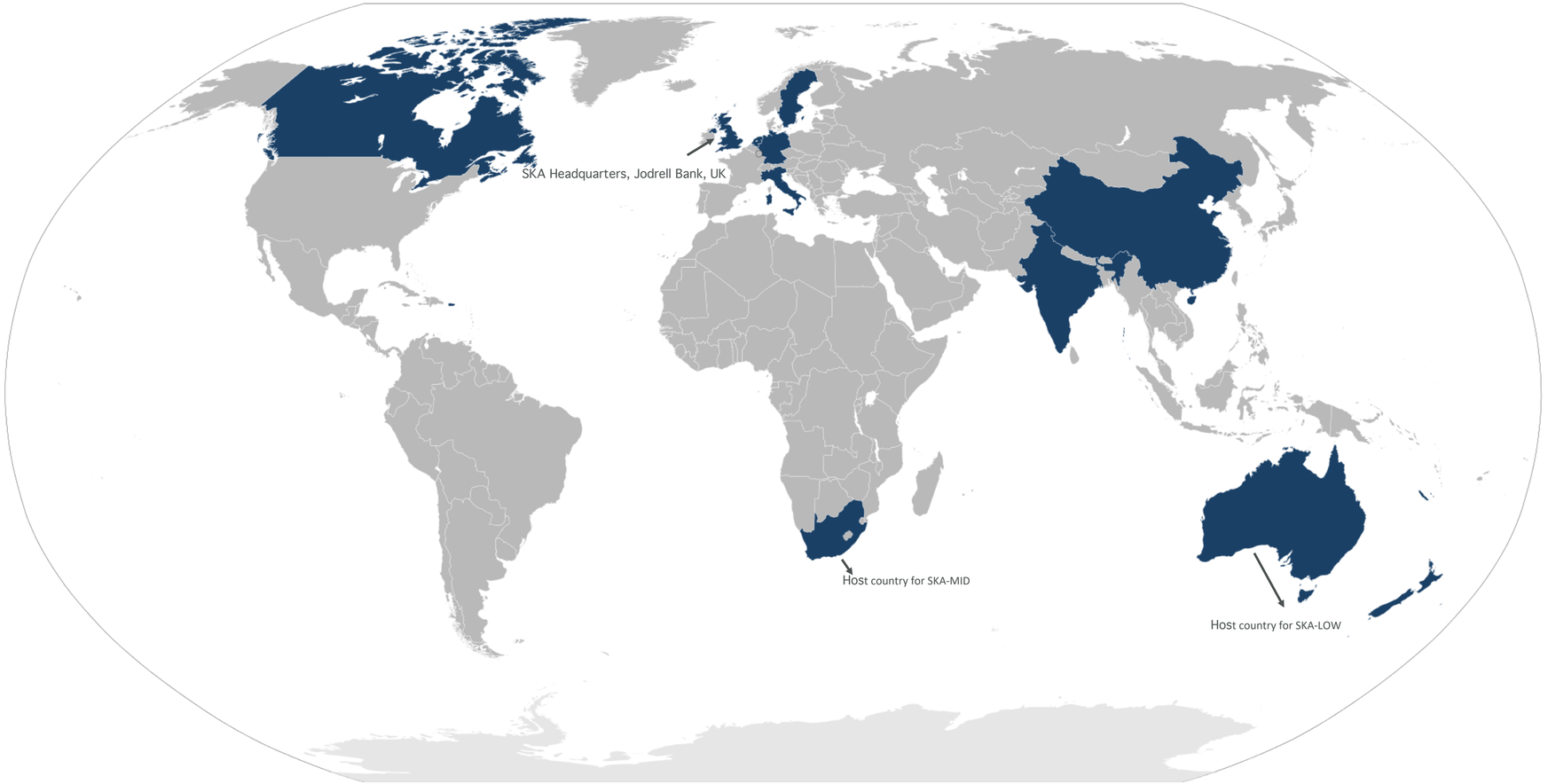
THE SQUARE KILOMETRE ARRAY

Science Data Processors at both the SKA sites will collect, process and churn out science data products that will be pushed out to one or more regional centres around the world

MEMBER COUNTRIES

- Australia
- Canada
- China
- India
- Italy
- Netherlands
- New Zealand
- South Africa
- Sweden
- UK

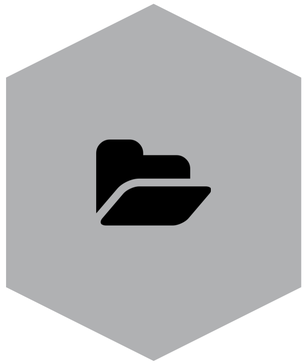
Potential new members: Spain, Portugal, Germany, France, others



Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

SKA REGIONAL CENTRES

SKA Regional centres will provide a platform for data access, data distribution, post-processing, archival storage, and software development.

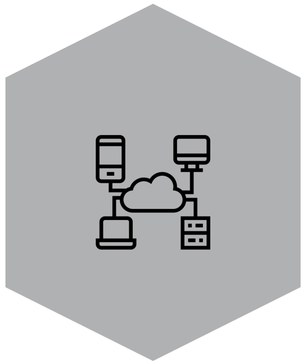


ARCHIVE

Archival of the raw data but also once scientific results are published, but outputs of analysis are made available.

DATA DISCOVERY

SDP has pushed the data to the regional centres, now what? How will users find/peruse their data. How will published results be easily found?

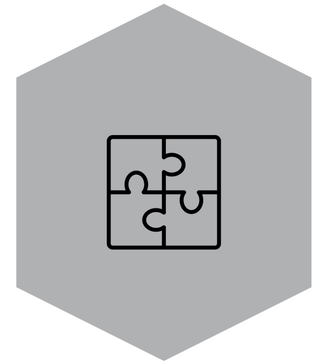


DISTRIBUTED DATA PROCESSING

Use cases are made to be reproducible. Compute comes to the data (high data volume).

USER SUPPORT

SRCs must support the key science project teams as well as general users. This will mean user ability with be varied.



INTEROPERIBILITY

Multiple regional SRCs, locally resourced but interoperable. SRCs may be heterogeneous in nature but with common core functionality

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

CHALLENGES

(k n o w n)

DIVERSE DATA/NON-HOMOGENEITY OF DATA

Diversity in data volumes, data formats, number of files, necessary structuring of data, access, what sort of secondary data products are generated. Regional centers must adhere to the data policies as defined by the SKA.

DISTRIBUTED RESOURCES, DISTRIBUTED USERS

Data and compute resources will be distributed, with a software stack such that a unified portal is shown to the user. Users will likely by few thousand.

DATA VOLUME

SKA Phase 1 Science archive is likely to be 300 PB per year. The regional centers would need to host advanced data products in addition to this.

TYPICAL USE(R) CASES

MY DATA IS WHERE? ON THE CLOUD?

Beginner users, good documentation.

HERE'S WHAT I WANT TO DO WITH MY DATA, HELP ME

Intermediate user, knows the best way to structure the data. User is equipped with some reproducible way to reduce their data (containers, Jupyter notebooks).

I KNOW WHAT I WANT, GET OUT OF MY WAY

Advanced user who will likely have complicated data and compute model, experienced with distributed data and compute resources. Also likely to need high data volume and compute resources.

DATA LIFECYCLE

Initial flurry of activity when observations come in, steady state as observations are monitored

Second stage of activity as data is being analysed and secondary data products are being created

Scientific results published, no more analysis of raw data, but outputs of analysis are openly available and raw data moved into long term storage

DATA PRODUCTS

These will vary from few GB per observation to several hundred. And data archives over time vary from few hundred GB to few hundred PB (from experiment to experiment). Users consuming data will also be generating secondary data which may not be smaller than raw data

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

WHAT WE HAVE TESTED SO FAR

Using the instance at RAL, more feature verification because even though data being used is astronomy data the volumes are not analogous to SKA data volumes

UPLOAD DATA

- Upload data using Rucio CLI
- Registering data already uploaded to Grid storage
- Generic metadata (ran into issues)

DATA REPLICATION

- Replicate data using Rucio CLI
- Parameterised replication rules
- Moved data around sites in the UK, SURFsara, and IDIA.

DATA DELETION

- Parametrised deletion rules
- Data deletion due to rule expiration

MISC

- Alternative methods of authentication when using the CLI
- Adding new RSEs

TO BE DONE

- Stress tests with large volumes of data
- Data integrity tests
- Register data in anticipation of SDP push
- Australian node coming soon

Thank you to Ian Johnson, Alastair Dewhurst !

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

HURDLES

(s o f a r)

ASTRONOMY DATA IS DIFFERENT THAN HEP DATA

Data models and compute models are simply not all known at the moment, leaving us to consider for all worst-case scenarios. Large variations in the size of data, number files, metadata, volume of secondary data generated, required proximity of the data. User community specific defaults might help.

CHALLENGES OF A FLAT NAMESPACE

A lot more thought and planning is required when working in a flat namespace (especially since DIDs must remain unique within a scope for all time). Users need to be convinced of the benefits as well.

X509 CERTIFICATES (NOT SPECIFIC TO RUCIO)

We are looking into a credential translation solution as an alternative to x509 certificates. Would be interesting to see how well it fits in with Rucio.

LOOKING AHEAD

ELASTIC SEARCH

We need plots!

LIGHTWEIGHT RUCIO CLI

Easy way to upload large amounts of data that doesn't sit on Grid storage/Grid compatible storage/isolated storage. SRC users may chose to download raw data, take it a suitable computing platform and upload the results.

PERMISSIONS

Astronomers like to keep their data protected. Traditionally, once data is available it is accessible only to the authorized people in the project. Once the proprietary period has passed, data is made public.

INTEGRATION WITH A WMS

DIRAC – Rucio collaboration in the form of a Rucio-mode for DIRAC sounds very promising. At the moment, we are unable to replicate a use case end to end

DTN – DTN TRANSFER

For long distance data transfers, there will be dedicated high-bandwidth network links that would be better suited. Routing the data via a DTN would provide a well-defined high speed data transfer environment.

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016

CONCLUSIONS

- Big data challenges coming with SKA, we want to use existing infrastructure, knowledge, solutions
- Rucio has worked well so far, and most of the issues found seem to be on the roadmap for this year (metadata, permissions)
- Integration with a WMS will be essential to demonstrate an end to end use case

Thank you!

Advanced European Network of E-infrastructures for Astronomy with the SKA AENEAS - 731016