SKAO REGIONAL CENTRE NETWORK

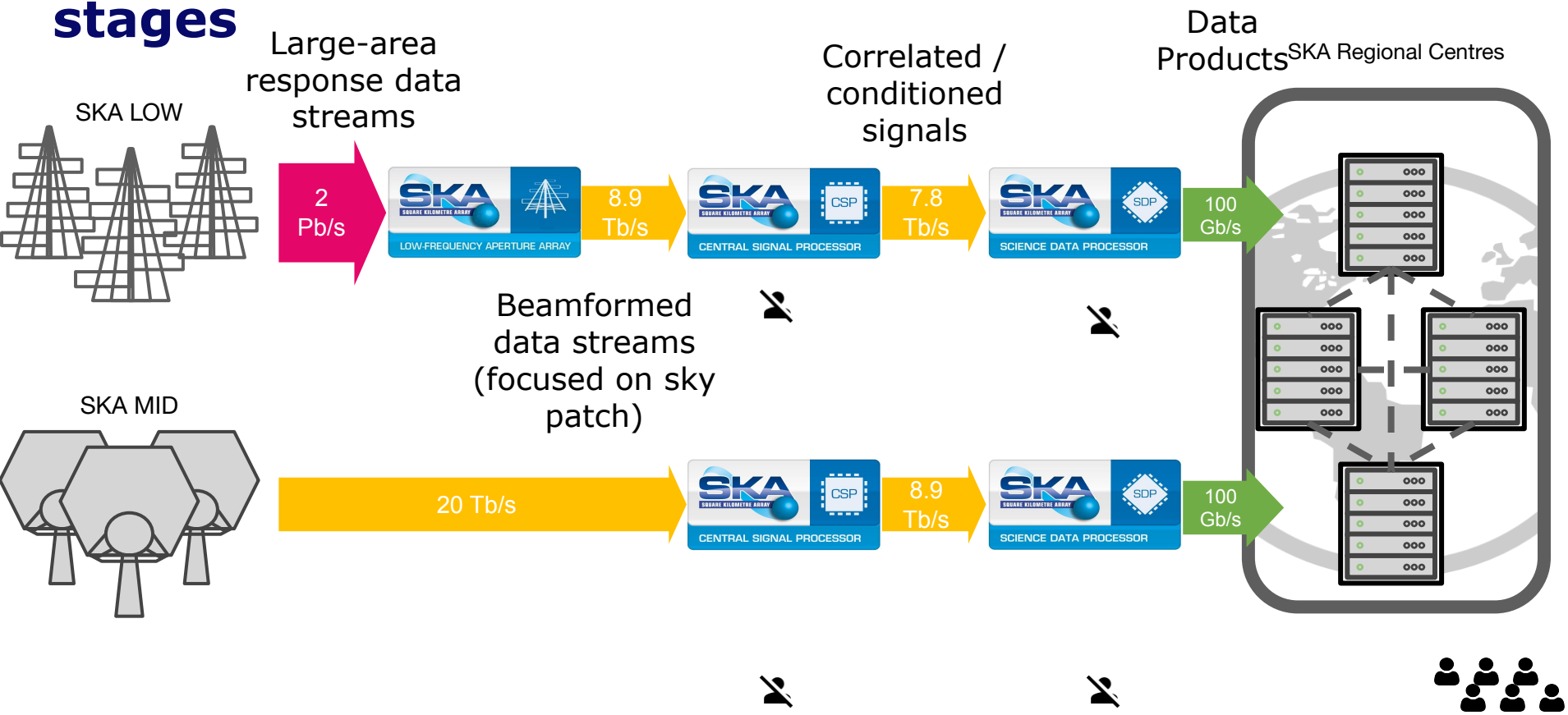
SKA Regional Centres

Dr Rosie Bolton

Head of Data Operations group, SKAO

27/09/2022

SKA Regional Centres: SKAO data processing stages



The SRC Vision

"To ensure that scientists can access SKA data products and use them to make discoveries"

What are we trying to achieve?

Globally distributed science teams
Diverse skill base - not all expert radio astronomers

Know who user is
Understand group membership
Respect proprietary periods
Support public access

"To ensure that scientists can access SKA data products and use them to make discoveries"

get data products from Observatory
allow inclusion of new data products
protect data products - ensure long term reliability
know where copies are
users and data products will be globally distributed
access depends on credentials not geography
meet SKA Construction timeline

data products analysed on SRC compute resources
allocation and tracking of resources to user groups
interactive sessions
user-defined software
batch workflows
user-defined data collections
create and save new data products and software
…and many more uses we need to keep talking with community about

What functions do SRCs need to provide?

The Role of SRCs: Data Intensity vs. User Flexibility

SRCs will bridge the gap between the highly data intensive **pre-defined workflows generating SKA data products** in the SDP, and the **iterative flexible, user-led data analysis** required to produce scientific results

"Size" or Data intensity

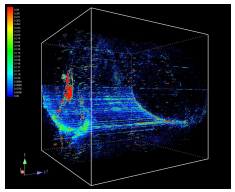


Image cubes

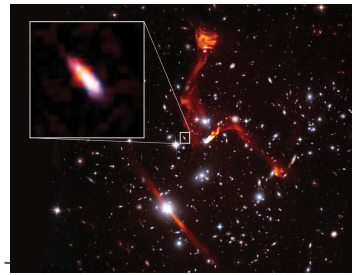
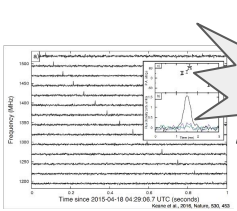


Image cut-outs

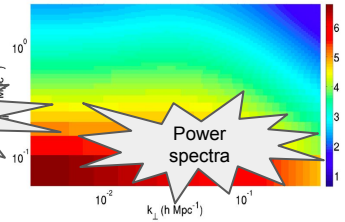


Time series data

Catalogues / Source List

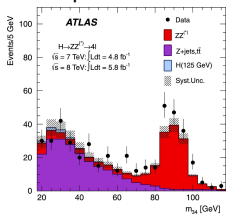
Credit: Heywood et al.; Sophia Dagnello, NRAO/AUI/NSF; STScI.

Paul, Sourabh et al. (2016). ApJ 833. 10.3847/1538-4357/833/2/213.



Power spectra

Plots and scripts for publication

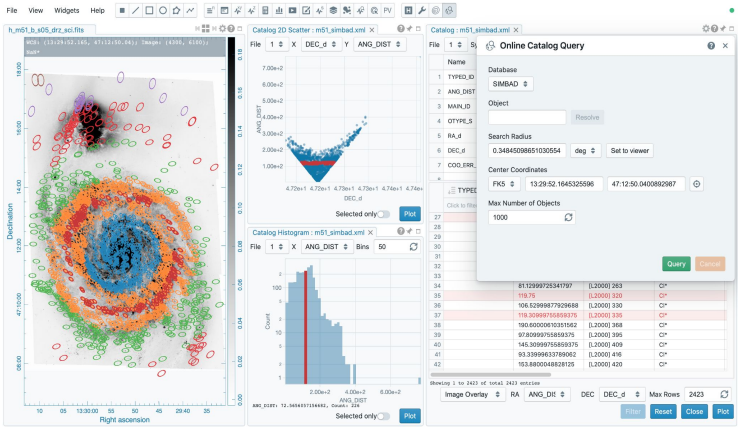


"Pre-defined" (Observatory Data Products)

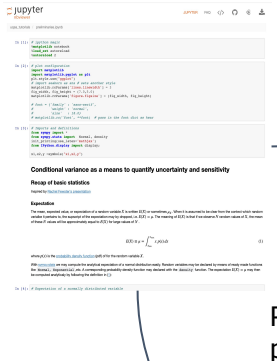
"User-defined" (Advanced Data Products)

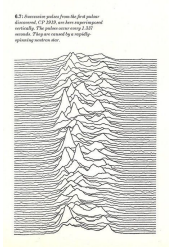
The Role of SRCs: Collaboration platform

SRCs will provide collaborative tools backed up by powerful compute and data management supporting **wide range of use cases**



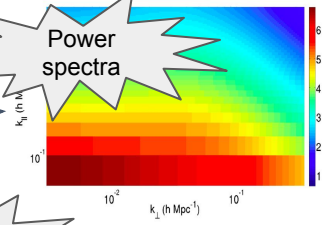
Workflows notebooks



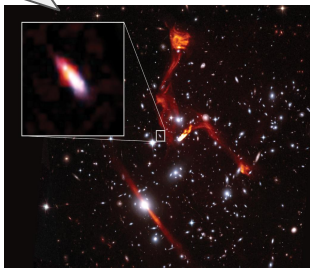


Paul, Sourabh et al. (2016). ApJ 833. 10.3847/1538-4357/833/2/213.

Power spectra



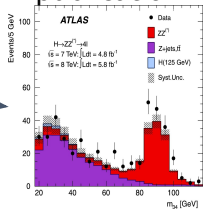
Catalogues / Source List



Credit: Heywood et al.; Sophia Dagnello, NRAO/AUI/NSF; STScI.

Image cut-outs

Plots for publication

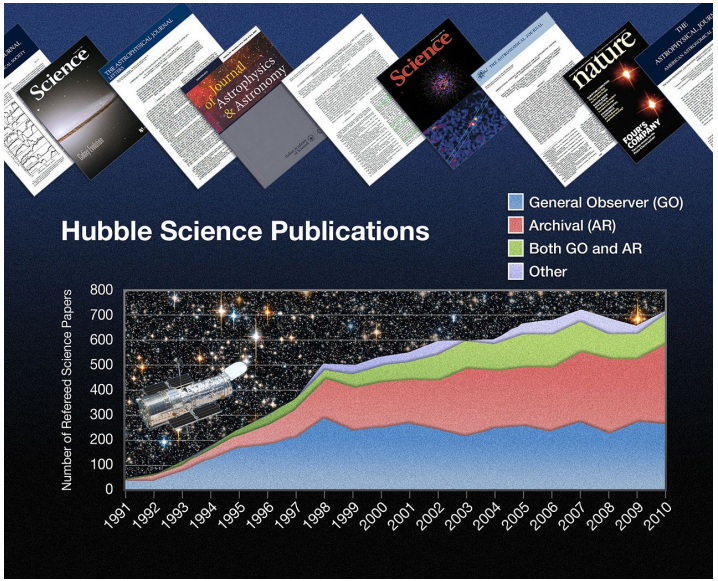


Users will not have access to the SDP or to Raw SKA data!

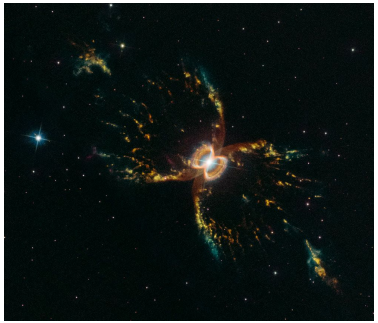
The Role of SRCs: Support data product (re-)use

- All SKA Data Products will (in time) become public - this is likely to be the biggest science generator long term

  - Build SKA science archive around International Virtual Observatory Alliance standards
  - Ensure interoperability with other archives and other experiments
  - Example from Hubble: Science archive users produce more scientific publications annually than users with dedicated observations



FAIR principles
(Find, Access, Interoperate, Re-use)



Southern Crab Nebula imaged by WFC3 - Hubblesite.org

SRC Users

- SKA Programme Users
  - With planned SKA projects
  - User has account in SRCs
  - Estimate SRC workflows and resources in SKA proposal
  - SRC workflows include generating Project-Level Data Products (SKAO responsible for SW for these)
  - Also analysis of OLDPs to produce Advanced Data Products
  - Resources allocated to the collaboration (SKA Project number)
- SRC Users: Public SKA data
  - Requested resources to run big analysis project on SRCNet
- Public users
  - "low" resource requirements, possibly anonymous access to SRC data collections

SRC Capabilities

SKA Regional Centers: Data management

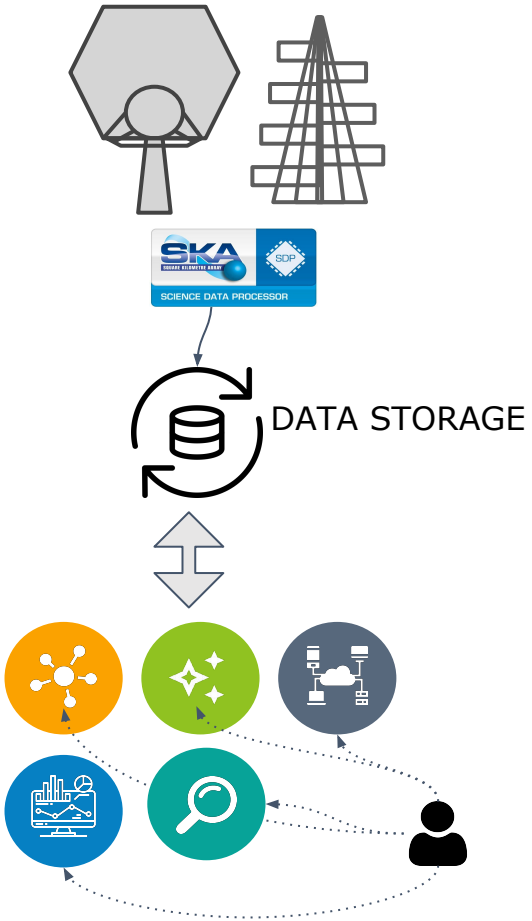
Storing SKAO data growing at up to 700 PBytes each year will be a challenge (plus user-generated data too).

Several million dollars per year in new data, for one copy

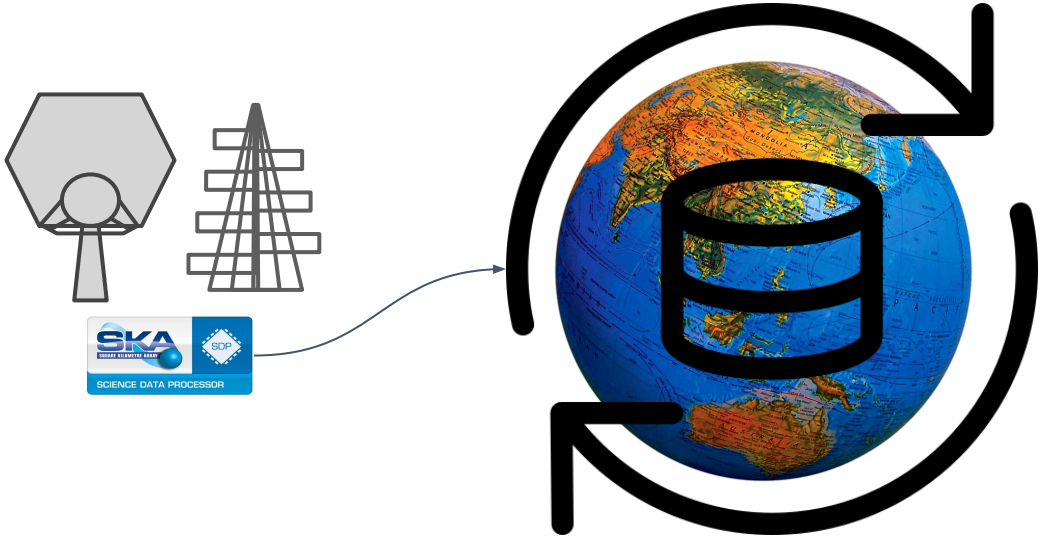
Global data management within SRCNet should enable best possible use to be made of available storage resources

Avoid (reduce) unnecessary duplication

Support mirroring of popular data products to enhance user experience



SKA Regional Centers: Important for Operation of SKA



Pushing data out of SDP staging areas into SRC storage is essential to keep the SKA telescopes running

- If staging area gets full, SKA operations will be impacted

We need to be able to predict link usage

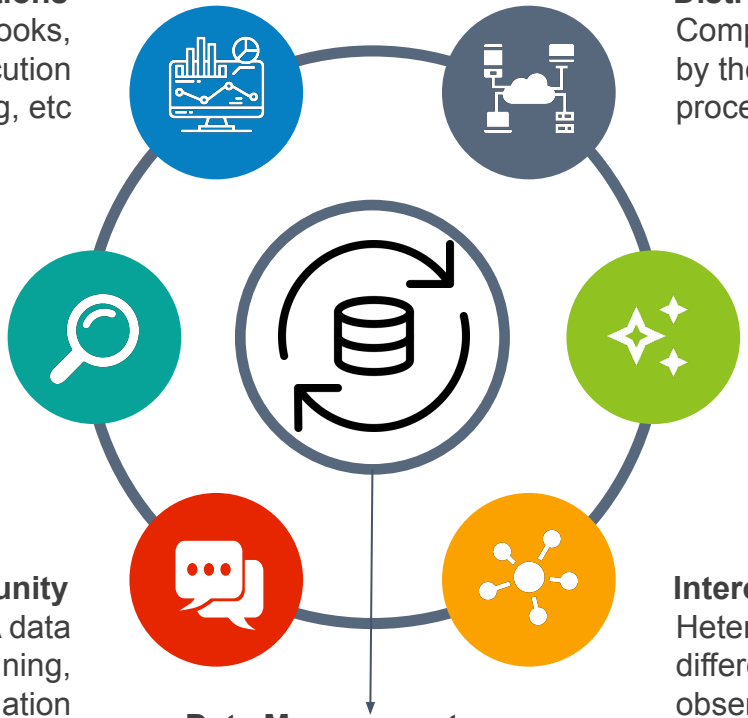
- Which data products are sent where?
- How reliable are links?
- How many copies? From SDP to each SRC or from SDP to one SRC then less time-critical copying

SKA Regional Centre Capabilities

Science Enabling Applications
Analysis Tools, Notebooks, Workflows execution Machine Learning, etc

Distributed Data Processing
Computing capabilities provided by the SRCNet to allow data processing

Data Discovery
Discovery of SKA data from the SRCNet, local or remote, transparently to the user



Visualization
Advanced visualizers for SKA data and data from other observatories

Support to Science Community
Support community on SKA data use, SRC services use, Training, Project Impact Dissemination

Interoperability
Heterogeneous SKA data from different SRCs and other observatories

Data Management
Dissemination of Data to SRCs and Distributed Data Storage

SRC Network global capabilities

Collectively meet the needs of the global community of SKA users

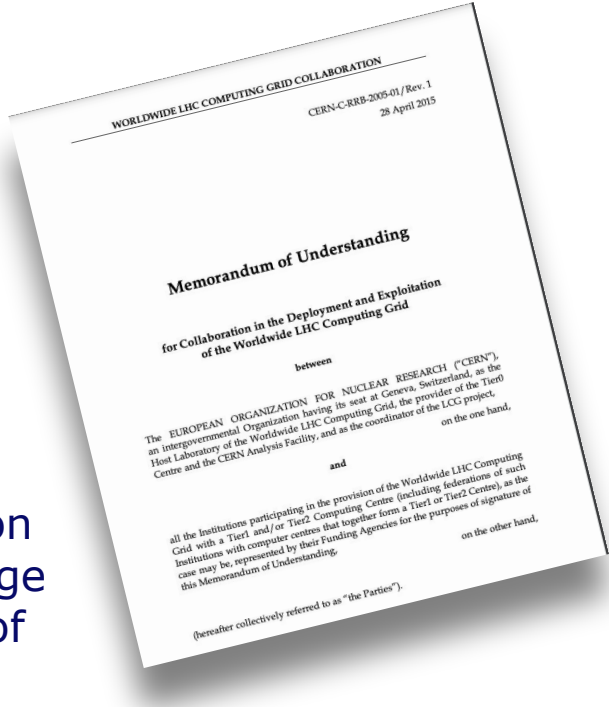
Anticipate heterogeneous SRCs, with different strengths

How might SRC resources be managed?

How might resources be obtained?

We hope that each SRC project will pledge resources into the SRCNet pool

MoU to cover deployment and use of SRCs



WORLDWIDE LHC COMPUTING GRID COLLABORATION

CERN-C-RRB-2005-01/Rev. 1

28 April 2015

Memorandum of Understanding

for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid

between

The EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH ("CERN"), an intergovernmental Organization having its seat at Geneva, Switzerland, as the Host Laboratory of the Worldwide LHC Computing Grid, the provider of the Tier0 Centre and the CERN Analysis Facility, and as the coordinator of the LCG project,

on the one hand,

and

all the Institutions participating in the provision of the Worldwide LHC Computing Grid with a Tier1 and/or Tier2 Computing Centre (including federations of such Institutions with computer centres that together form a Tier1 or Tier2 Centre), as the case may be, represented by their Funding Agencies for the purposes of signature of this Memorandum of Understanding,

on the other hand,

(hereafter collectively referred to as "the Parties").

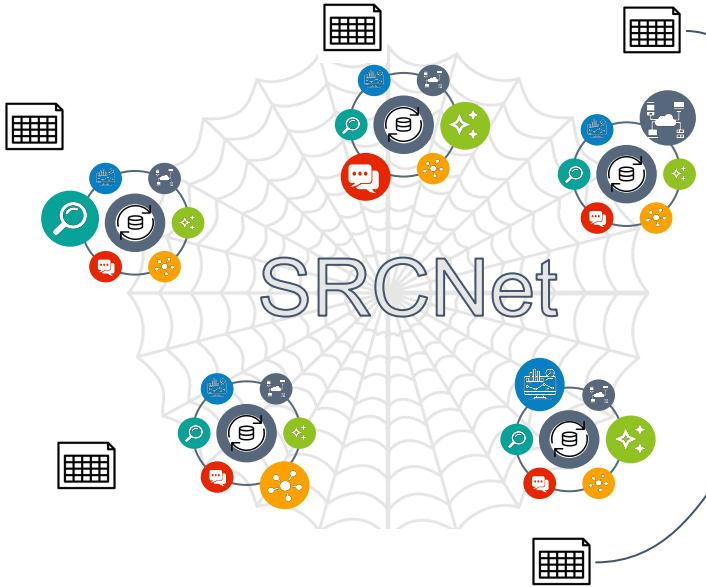
Public pledging system to gather resource offers, rolling (e.g. look ahead 2-5 years, revise each year)

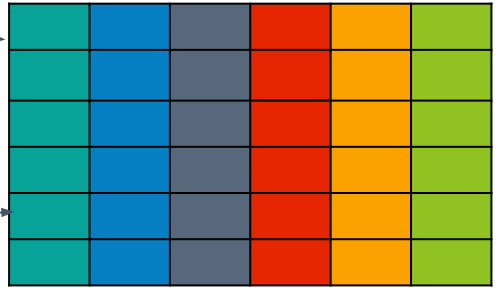
- Programme users should scale with SKA commitment
  - SRC pledges hope to match or exceed this fraction
  - e.g. 5% stakeholder in SKA, we hope would pledge resources at least (or exceeding) estimated 5% of the total required SRCNet resources to support users described previously.

Pledging

Each SRC to pledge resources into global pool to support SRCNet activities

Users can access resources across SRCNet according to their research needs and permissions



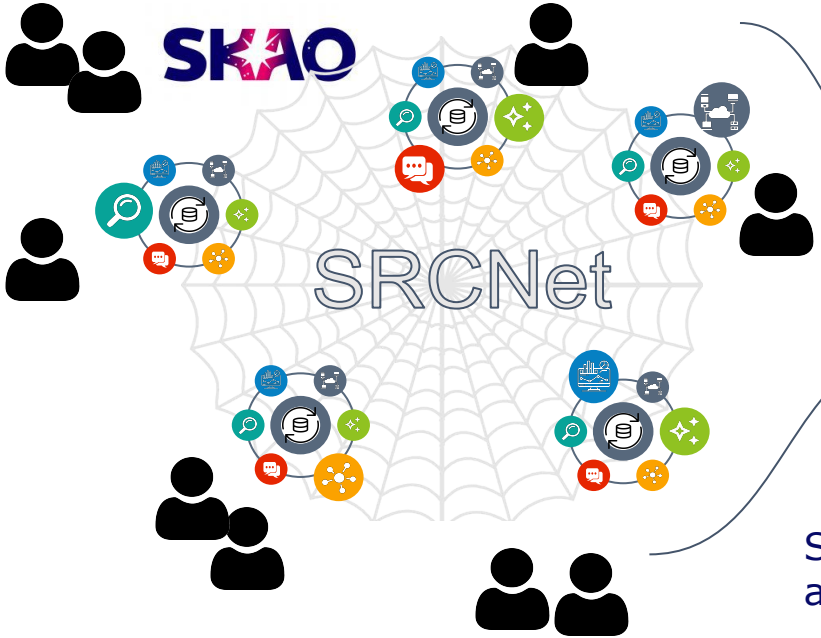


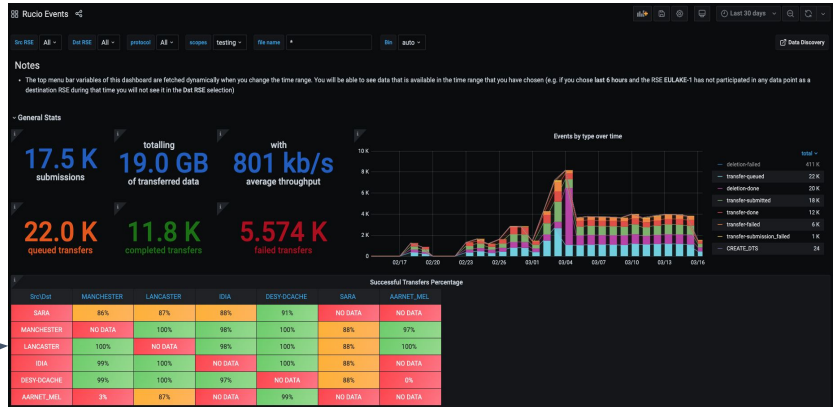
Hope is that each SRC will be able to contribute a total effort that is proportional to their SKA fraction

Additional resources at an SRC could be given to the pool or prioritised to support national interests

Operations

Personnel within each SRC project will be identified to be part of the SRC Operations Group (SOG) - meeting regularly to discuss issues, share tasks, see and test global system health



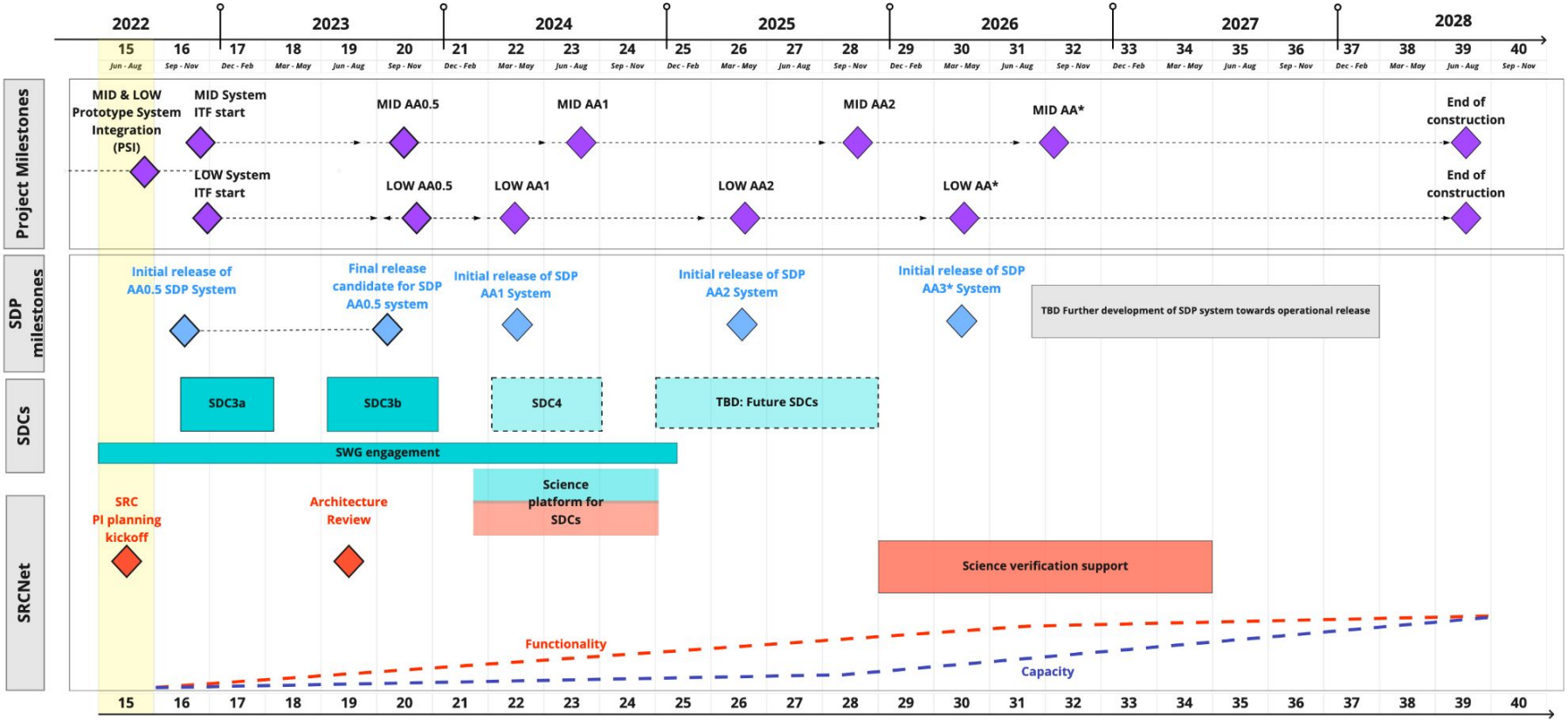


(an example dashboard from our data management prototype, details not important, but nice to see that we are using UK grid storage endpoints in our Rucio prototype which is itself run off **IRIS resources at STFC cloud**)

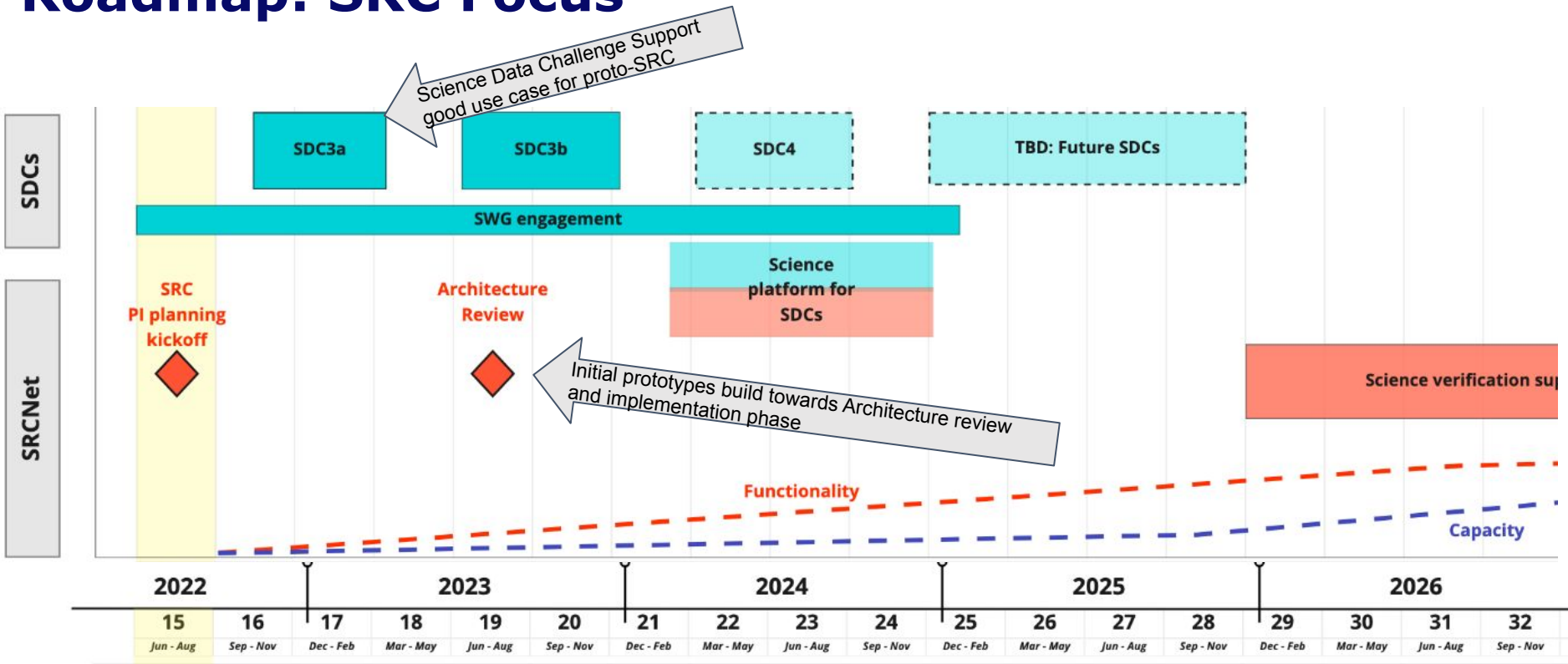
SOG will be led from SKAO Ops, with a team from across each SRC project and SKAO.

Development plans and progress

Roadmap


2022
2023
2024
2025
2026
2027
2028
15 Jun - Aug
16 Sep - Nov
17 Dec - Feb
18 Mar - May
19 Jun - Aug
20 Sep - Nov
21 Dec - Feb
22 Mar - May
23 Jun - Aug
24 Sep - Nov
25 Dec - Feb
26 Mar - May
27 Jun - Aug
28 Sep - Nov
29 Dec - Feb
30 Mar - May
31 Jun - Aug
32 Sep - Nov
33 Dec - Feb
34 Mar - May
35 Jun - Aug
36 Sep - Nov
37 Dec - Feb
38 Mar - May
39 Jun - Aug
40 Sep - Nov
Project Milestones
MID & LOW Prototype System Integration (PSI)
MID System ITF start
MID AA0.5
MID AA1
MID AA2
MID AA*
End of construction
LOW System ITF start
LOW AA0.5
LOW AA1
LOW AA2
LOW AA*
End of construction
SDP milestones
Initial release of AA0.5 SDP System
Final release candidate for SDP AA0.5 system
Initial release of SDP AA1 System
Initial release of SDP AA2 System
Initial release of SDP AA3* System
TBD Further development of SDP system towards operational release
SDCs
SDC3a
SDC3b
SDC4
TBD: Future SDCs
SWG engagement
SRCNet
SRC PI planning kickoff
Architecture Review
Science platform for SDCs
Science verification support
Functionality
Capacity
15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

Roadmap: SRC Focus



First Prototyping Phase: 2022-2023

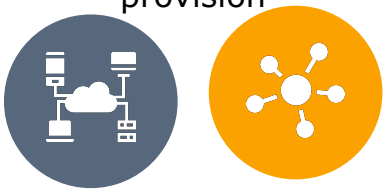
Work *now being undertaken* by development teams to prototype key technologies that will enable selection as SRC functionality and scale grows.

Data Management service: Replication, distribution, synchronisation of data products and location index

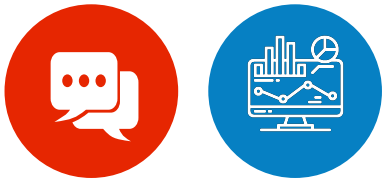
Teams working together with alignment driven by the common roadmap

Federated **Authentication and Authorization**: identity management, compatible with SKAO

Central Services and Software Distribution: SW infrastructure, compute provision



Data Analysis: Science Extraction, Processing in Notebooks



Data Visualisation and discovery - performance at SKA scale

SKA Regional Centers: Data management

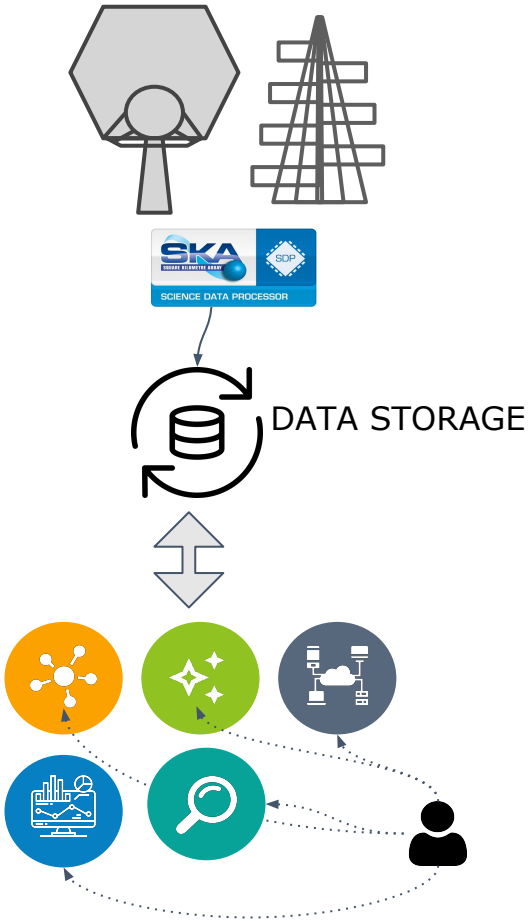
Storing SKAO data growing at up to 700 PBytes each year will be a challenge (plus user-generated data too).

Several million dollars per year in new data, for one copy

Global data management within SRCNet should enable best possible use to be made of available storage resources

Avoid (reduce) unnecessary duplication

Support mirroring of popular data products to enhance user experience



SRC Rucio prototype

Hope for Chinese site by December 2022!

Well suited to centralised Operations model for data management

Performed long-haul transfers, Rucio stress tests, subscriptions (via our automated test framework)

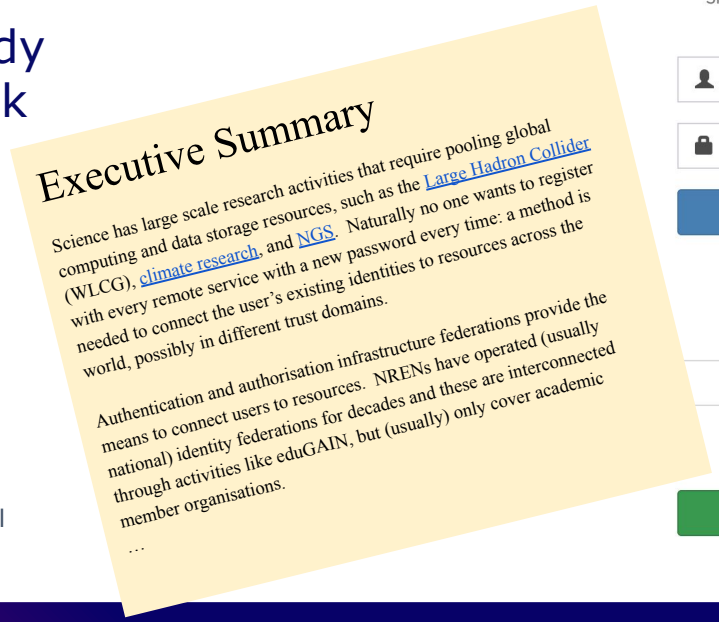
Integrating storage from national SRC efforts to increase understanding and inform assessment

Progress - Identity management (Authentication and Authorisation)

- "IAM"* service deployed in UK (Rutherford Appleton Lab)

https://ska-iam.stfc.ac.uk/login

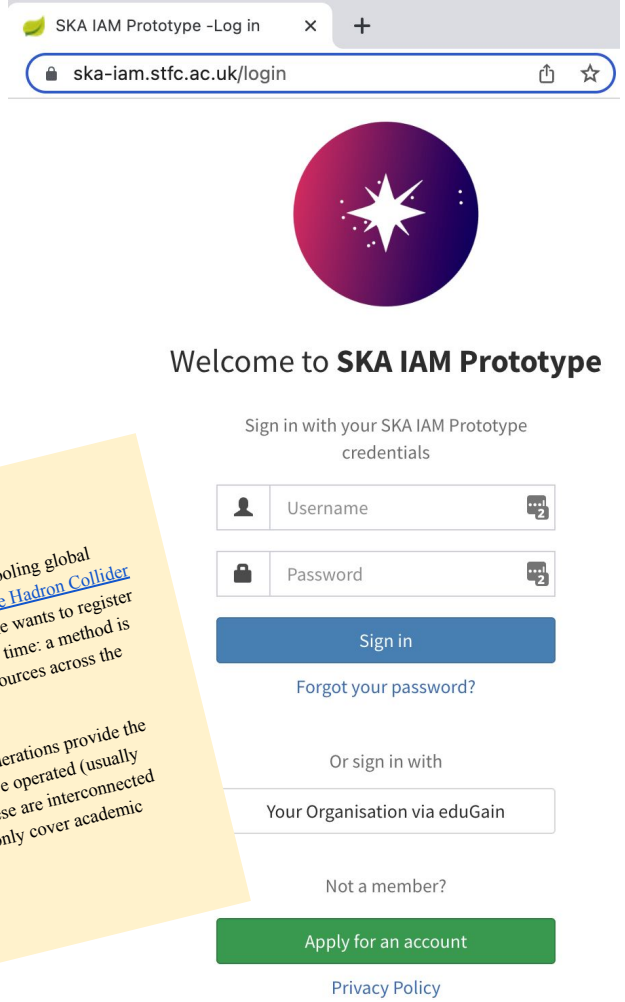
- Landscape report created ready for wider sharing and feedback


Executive Summary

Science has large scale research activities that require pooling global computing and data storage resources, such as the Large Hadron Collider (WLCG), climate research, and NGS. Naturally no one wants to register with every remote service with a new password every time: a method is needed to connect the user's existing identities to resources across the world, possibly in different trust domains.

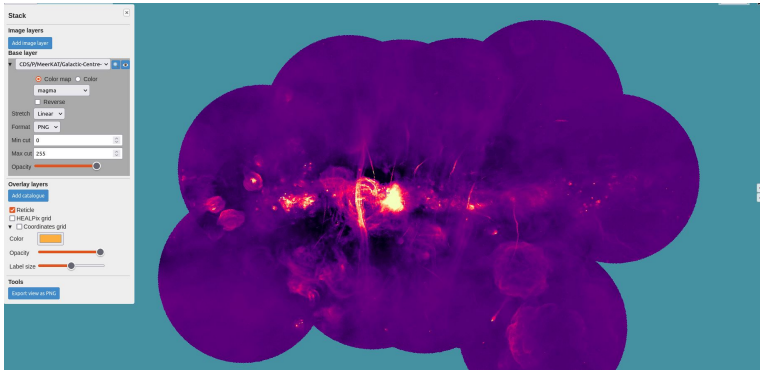
Authentication and authorisation infrastructure federations provide the means to connect users to resources. NRENs have operated (usually national) identity federations for decades and these are interconnected through activities like eduGAIN, but (usually) only cover academic member organisations.

…

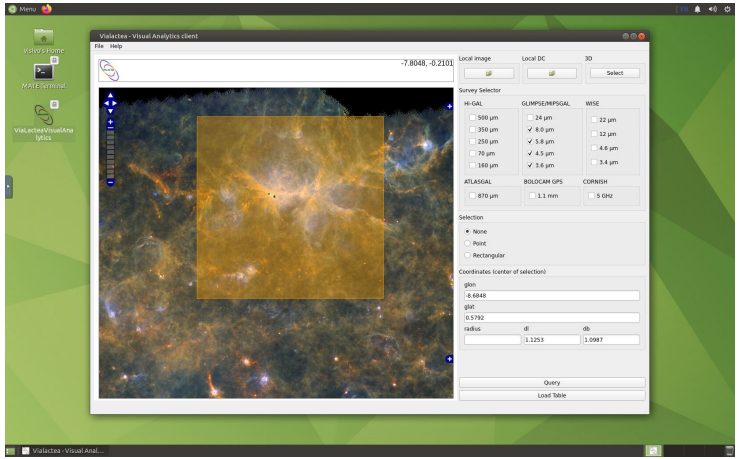


*https://indigo-iam.github.io/docs/v/current/about.html

Progress - Visualisation



- Comparative review of three tools where teams have expertise: CARTA, ALADIN and visIVO
- Collection of data products (with links) for use in demos
- Demo of these tools to the SRCNet team of teams
- Containerised CARTA deployment documented
- All three tools deployed at CHINA SRC, and access given to external users (user pass, IP address whitelist)



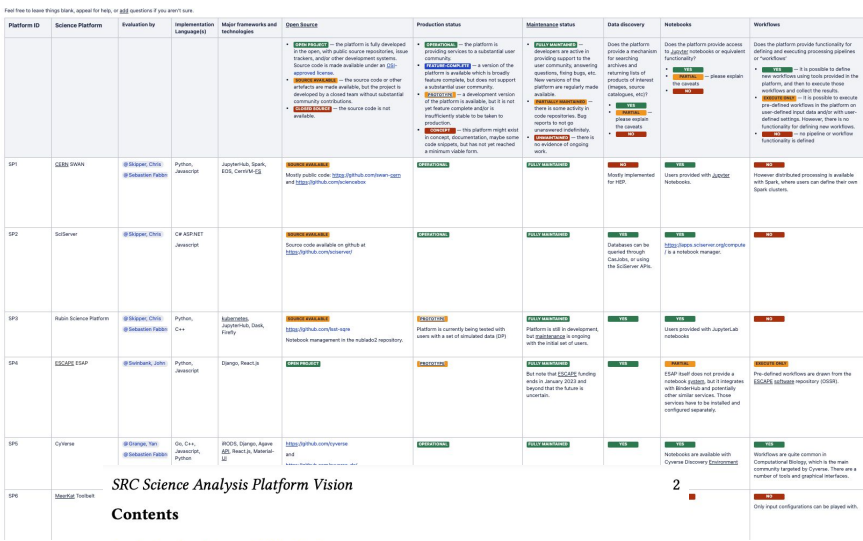
Progress - Science Platform

- Evaluation of (30) existing platforms considering many different criteria

Shortlist of 8 (or so) for further work

- Science Platform vision document ready for review

Next steps are to consider architectural aspects of science platform vision, update vision to incorporate this.



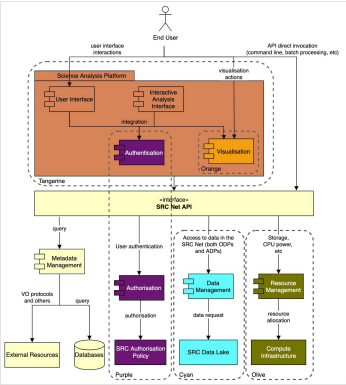
Feel free to leave things blank, appeal for help, or add questions if you aren't sure.

Platform ID	Science Platform	Evaluation by	Implementation Language(s)	Major frameworks and technologies	Open Source	Production status	Maintenance status	Data discovery	Notebooks	Workflows
					• OPEN PROJECT — the platform is fully developed in the open, with public source repositories, issue trackers, and/or other development systems. Source code is made available under an OSI-approved license. • SOURCE AVAILABLE — the source code or other artifacts are made available, but the project is developed by a closed team without substantial community contributions. • SOURCE LOCKED — the source code is not available.	• OPERATIONAL — the platform is providing services to a substantial user community. • BEYOND COMPETE — a version of the platform is available which is broadly feature complete, but does not support a substantial user community. • PROTOTYPE — a development version of the platform is available, but it is not yet feature complete and/or is insufficiently stable to be taken to production. • CONCEPT — this platform might exist in concept, documentation, maybe some code snippets, but has not yet reached a minimum viable form.	• FULLY MAINTAINED — developers are active in providing support to the user community, answering questions, fixing bugs, etc. New versions of the platform are regularly made available. • PARTIALLY MAINTAINED — there is some activity in code repositories. Bug reports to not go unanswered indefinitely. • UNMAINTAINED — there is no evidence of ongoing work.	Does the platform provide a mechanism for searching archives and returning lists of products of interest (images, source catalogues, etc)? • YES • PARTIAL — please explain the caveats • NO	Does the platform provide access to Jupyter notebooks or equivalent functionality? • YES • PARTIAL — please explain the caveats • NO	Does the platform provide functionality for defining and executing processing pipelines or 'workflows' • YES — It is possible to define new workflows using tools provided in the platform, and then to execute those workflows and collect the results. • EXECUTE ONLY — it is possible to execute pre-defined workflows in the platform on user-defined input data and/or with user-defined settings. However, there is no functionality for defining new workflows. • NO — no pipeline or workflow functionality is defined
SP1	CERN SWAN	@Skipper, Chris @Sebastien Fabbro	Python, Javascript	Jupyter/sub, Spark, EOS, CernVM-FS	SOURCE AVAILABLE Mostly public code: https://github.com/swan-cern and https://github.com/sciencebox	OPERATIONAL	FULLY MAINTAINED	NO Mostly implemented for HEP.	YES Users provided with Jupyter Notebooks.	NO However distributed processing is available with Spark, where users can define their own Spark clusters.
SP2	SciServer	@Skipper, Chris	C# ASP.NET Javascript		SOURCE AVAILABLE Source code available on github at https://github.com/sciserver/	OPERATIONAL	FULLY MAINTAINED	YES Databases can be queried through CasJobs, or using the SciServer APIs.	YES https://apps.sciserver.org/compute / is a notebook manager.	NO
SP3	Rubin Science Platform	@Skipper, Chris @Sebastien Fabbro	Python, C++	kubernetes, Jupyter/sub, Dask, Firefly	SOURCE AVAILABLE https://github.com/lsst-sqre Notebook management in the rubladd2 repository.	PROTOTYPE Platform is currently being tested with users with a set of simulated data (DP)	FULLY MAINTAINED Platform is still in development, but maintenance is ongoing with the initial set of users.	YES	YES Users provided with JupyterLab notebooks	NO
SP4	ESCAPE ESAP	@Svenbanck, John	Python, Javascript	Django, React.js	OPEN PROJECT	PROTOTYPE	FULLY MAINTAINED But note that ESCAPE funding ends in January 2023 and beyond that the future is uncertain.	YES	PARTIAL ESAP itself does not provide a notebook system, but it integrates with BinderHub and potentially other similar services. Those services have to be installed and configured separately.	EXECUTE ONLY Pre-defined workflows are drawn from the ESCAPE software repository (OSSR).
SP5	CyVerse	@Orange, Yan @Sebastien Fabbro	Go, C++, Javascript, Python	iRODS, Django, Agave API, React.js, Material-UI	https://github.com/cyverse and	OPERATIONAL	FULLY MAINTAINED	YES	YES Notebooks are available with Cyverse Discovery Environment	YES Workflows are quite common in Computational Biology, which is the main community targeted by Cyverse. There are a number of tools and graphical interfaces.
SP6	MegaKat Toolbelt									NO Only input configurations can be played with.

SRC Science Analysis Platform Vision 2

Contents

Many layers, shared vision



Users

Science Platform

Interface (API layer)

Metadata query - Science Data Discovery

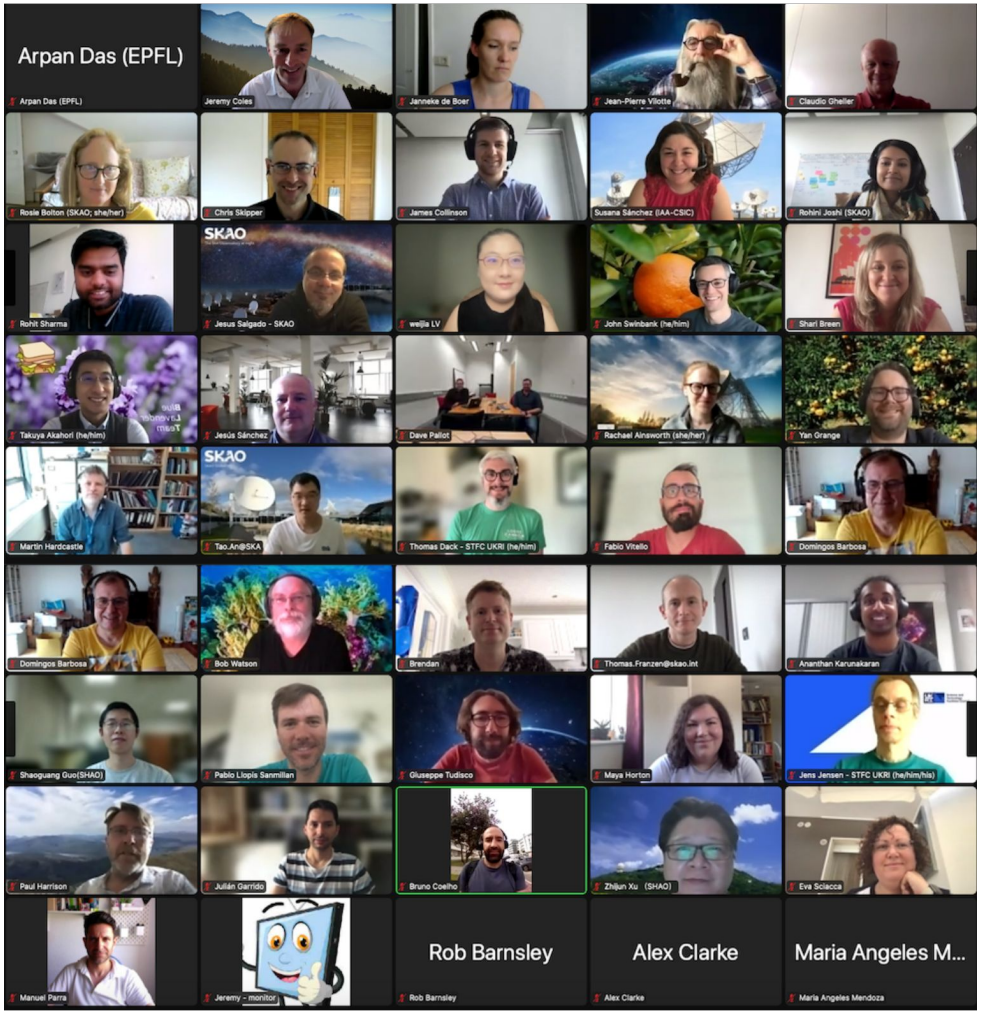
Authentication Who? Permissions?

Data Logistics Globally distributed storage sites

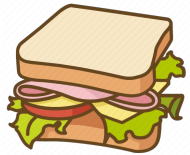
Compute Resource Management Work sharing

SRC Prototyping

- Prototyping started June 2022
- Team members from 12 countries plus SKAO
- about 1000 developer-days per 3-months
- Anticipate growing as national funding to develop SRC nodes is available



Introducing the Blue-Lavender team



Excellent collaboration between **China, Japan, South Korea and Australia** team members

Already achieved:

Deploying Visualisation tools at China SRC

Adding AUS-SRC storage into data management prototype

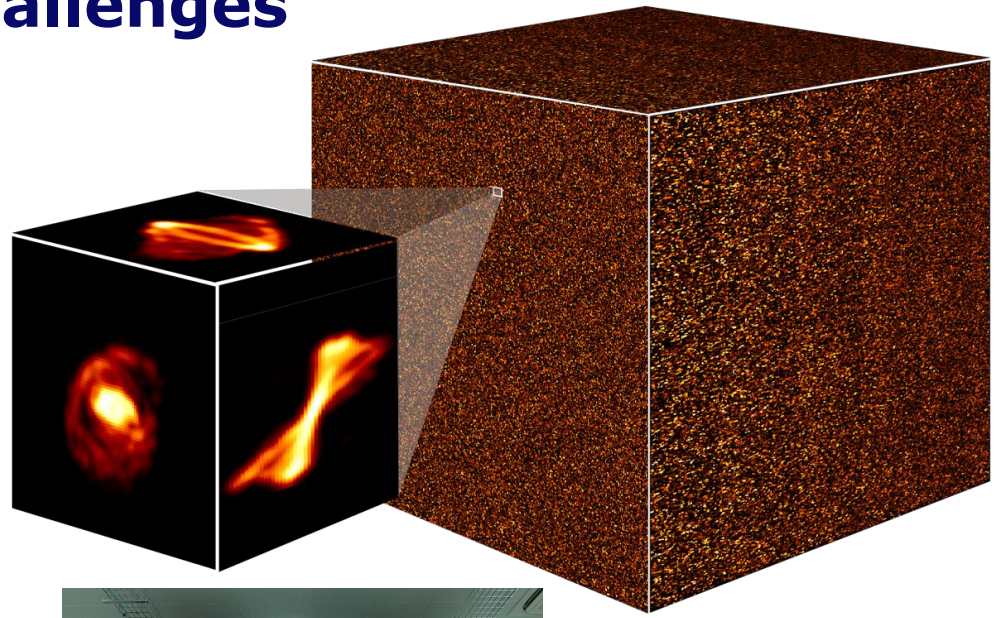
Next: Supporting other teams with visualisation workshop; Adding Chinese and Japanese SRC storage to prototype; defining needs for SDC3 support at China SRC

Selected team member images:
An, Tao
Takuya Akahori
Dave Pallot
Austin Shen
Gordon German
Kang Hyunwoo
Lv, Weijia
Guo, Shaoguang

Link to Science Data Challenges

Science Data Challenges are a great way to involve science community (e.g. research teams - hopefully some of you!) with simulated SKA data products

SRC projects offer a route to supporting challenge participants (including China SRC)



Credit: An, Wu, Hong, Nature Astronomy 3, 1030 (2019)

Summary

- SRCs are essential for connecting users with SKA data
- Collaborative model with Observatory - via MoU
- Pledged resources to provide global access
- Prototyping ideas for components now
- Implementation from end 2023, to keep pace with SKA Telescope development

End