Provenance Requirements for SKA

Jesús Salgado -SKA Regional Centre Architect
Rosie Bolton - Head of Data Operations

SKA1-low – the SKA's low-frequency instrument

The Square Kilometre Array (SKA) is a next-generation radio astronomy facility that will revolutionise our understanding of the Universe. It will have a uniquely distributed character: **one** observatory operating **two** telescopes on **three** continents. Construction of the SKA will be phased and work is currently focused on the first phase named SKA1, corresponding to a fraction of the full SKA. SKA1 will include two instruments – SKA1-mid and SKA1-low – observing the Universe at different frequencies.

SKA SQUARE KILOMETRE ARRAY

Location: Australia

Frequency range: **50 MHz** to **350 MHz**

~131,000 antennas spread between **500 stations**

Total collecting area: **0.4km²**

Maximum distance between stations: **>65km**

SKA1-low

Data transfer rate: **7.2 Terabits** per second

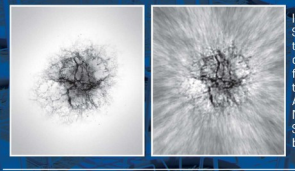


Image quality of SKA1-low (left) versus the best current facility operating in the same frequency range, the LOw Frequency ARray (LOFAR), in the Netherlands (right). SKA-low's resolution will be similar to LOFAR.

Compared to LOFAR Netherlands, the current best similar instrument in the world

25% better resolution

8x more sensitive

135x the survey speed

www.skatelescope.org @SKA_telescope SKAtelescope ska_telescope Square Kilometre Array ska-organisation

SKA1-mid – the SKA's mid-frequency instrument

The Square Kilometre Array (SKA) is a next-generation radio astronomy facility that will revolutionise our understanding of the Universe. It will have a uniquely distributed character: **one** observatory operating **two** telescopes on **three** continents. Construction of the SKA will be phased and work is currently focused on the first phase named SKA1, corresponding to a fraction of the full SKA. SKA1 will include two instruments – SKA1-mid and SKA1-low – observing the Universe at different frequencies.

SKA SQUARE KILOMETRE ARRAY

Location: South Africa

Frequency range: **350 MHz** to **15.3 GHz** with a goal of 24 GHz

197 dishes (including 64 MeerKAT dishes)

Total collecting area: **33,000m²**

or **126 tennis courts**

Maximum distance between dishes: **150km**

SKA1-mid

Data transfer rate: **8.8 Terabits** per second

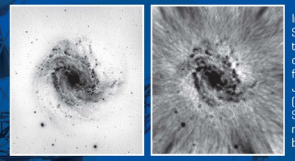


Image quality of SKA1-mid (left) versus the best current facility operating in the same frequency range, the Jansky Very Large Array (JVLA) in the United States (right). SKA-mid's resolution will be 4x better than JVLA.

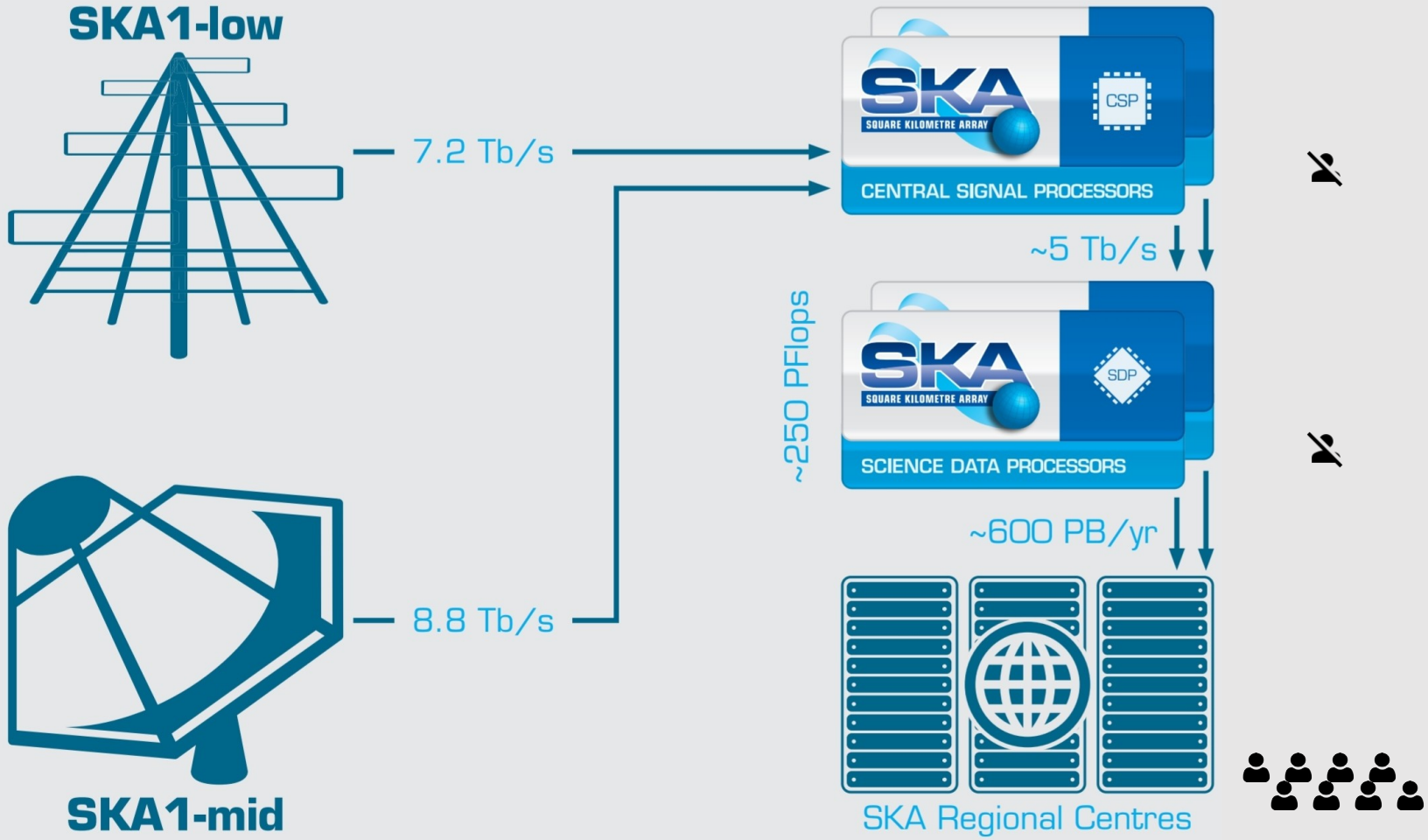
Compared to the JVLA, the current best similar instrument in the world:

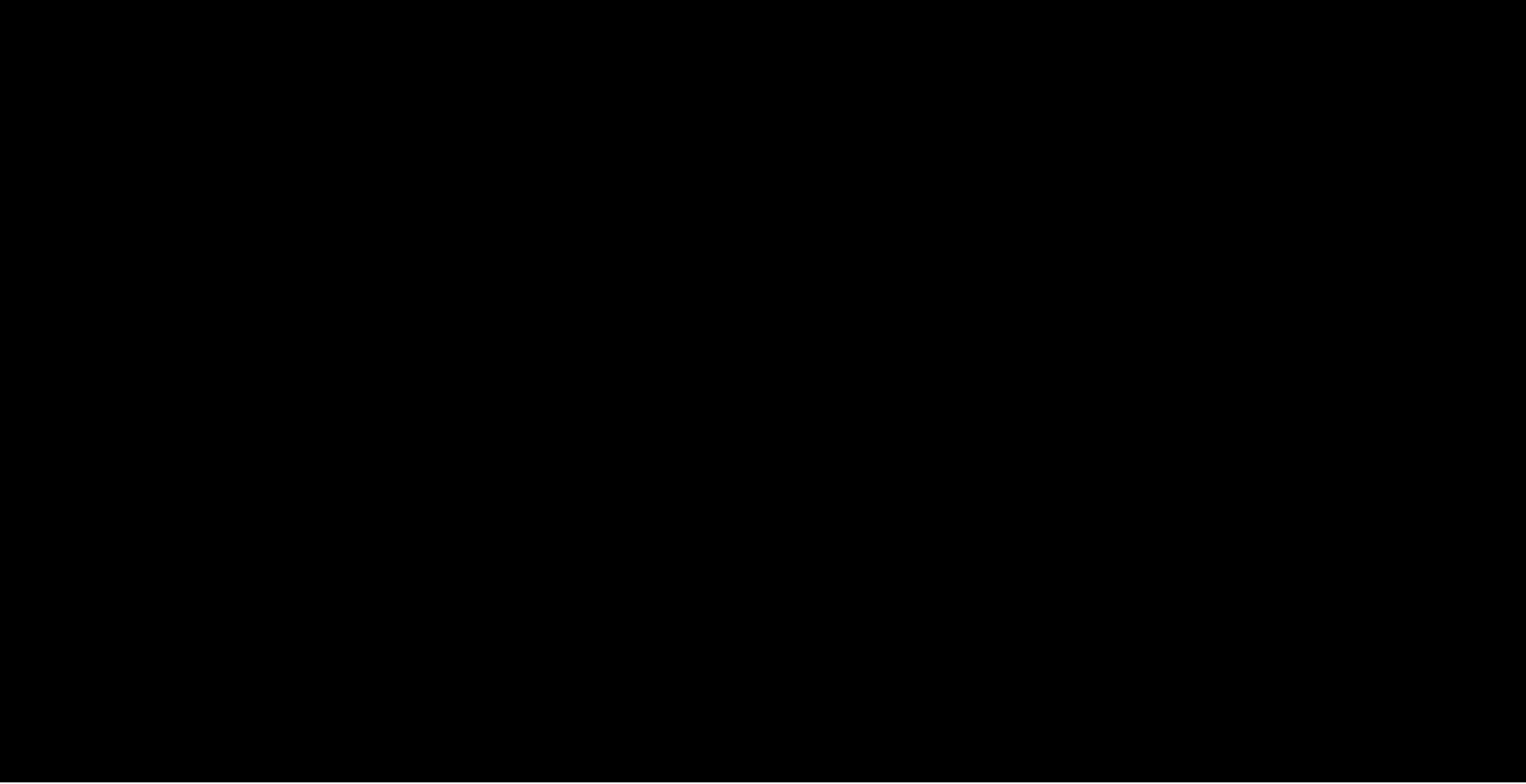
4x the resolution

5x more sensitive

60x the survey speed

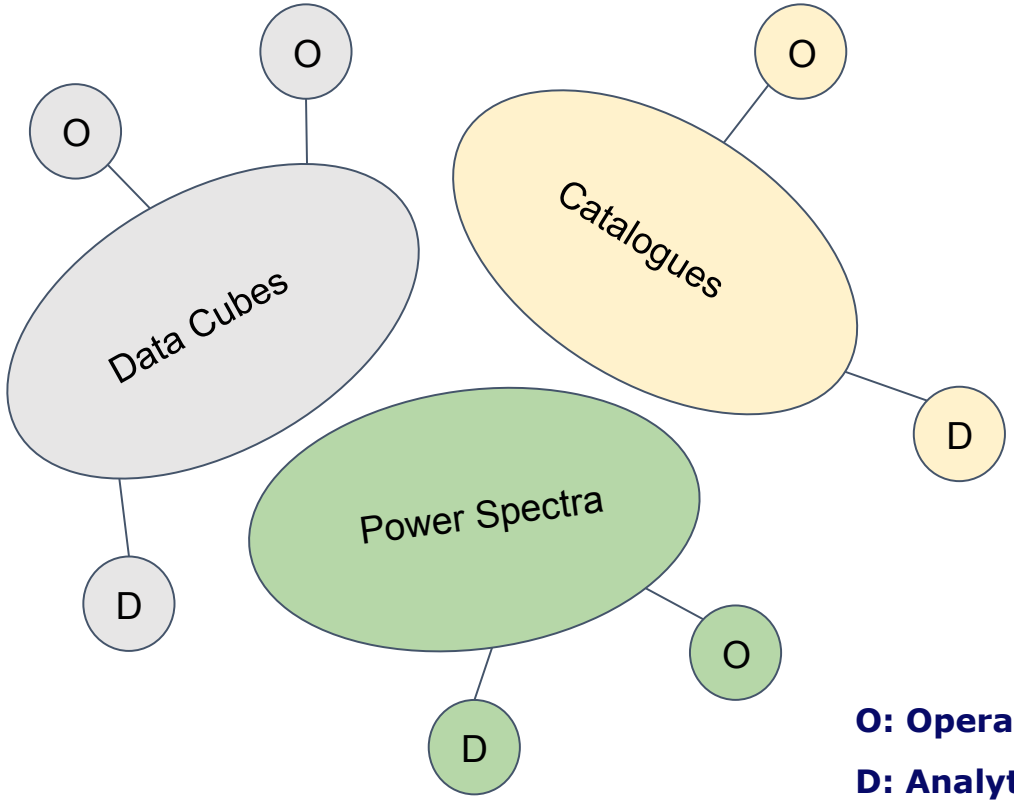
www.skatelescope.org @SKA_telescope SKAtelescope ska_telescope Square Kilometre Array ska-organisation


SKA1-low
7.2 Tb/s
SKA1-mid
8.8 Tb/s
SKA
SQUARE KILOMETRE ARRAY
CSP
CENTRAL SIGNAL PROCESSORS
~5 Tb/s
~250 PFlops
SKA
SQUARE KILOMETRE ARRAY
SDP
SCIENCE DATA PROCESSORS
~600 PB/yr
SKA Regional Centres



(Animation credit: AENEAS)

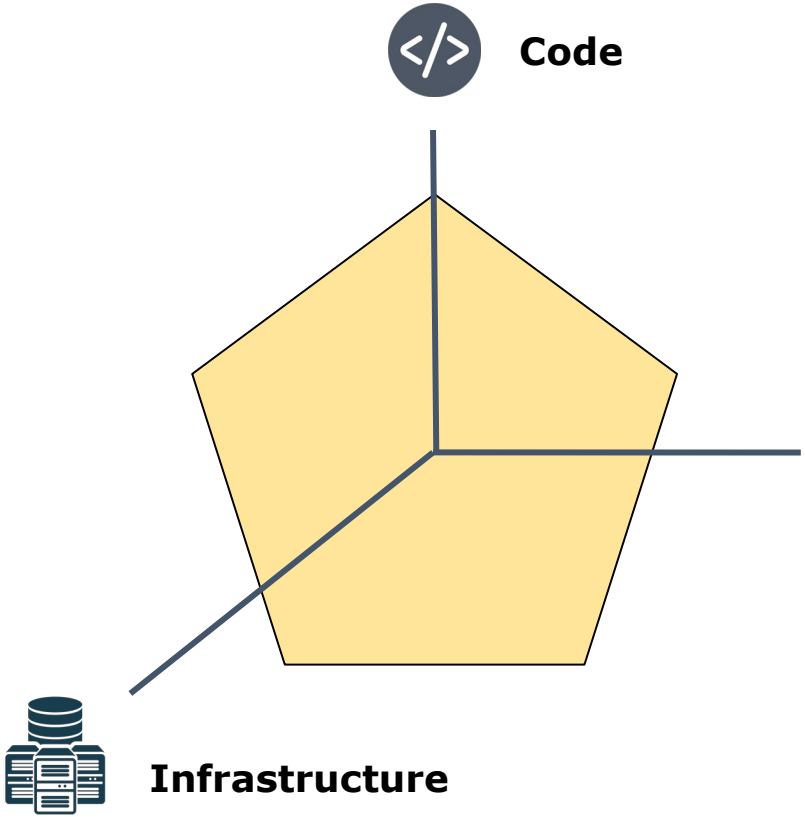
Data Mesh: Domain Oriented



O: Operations

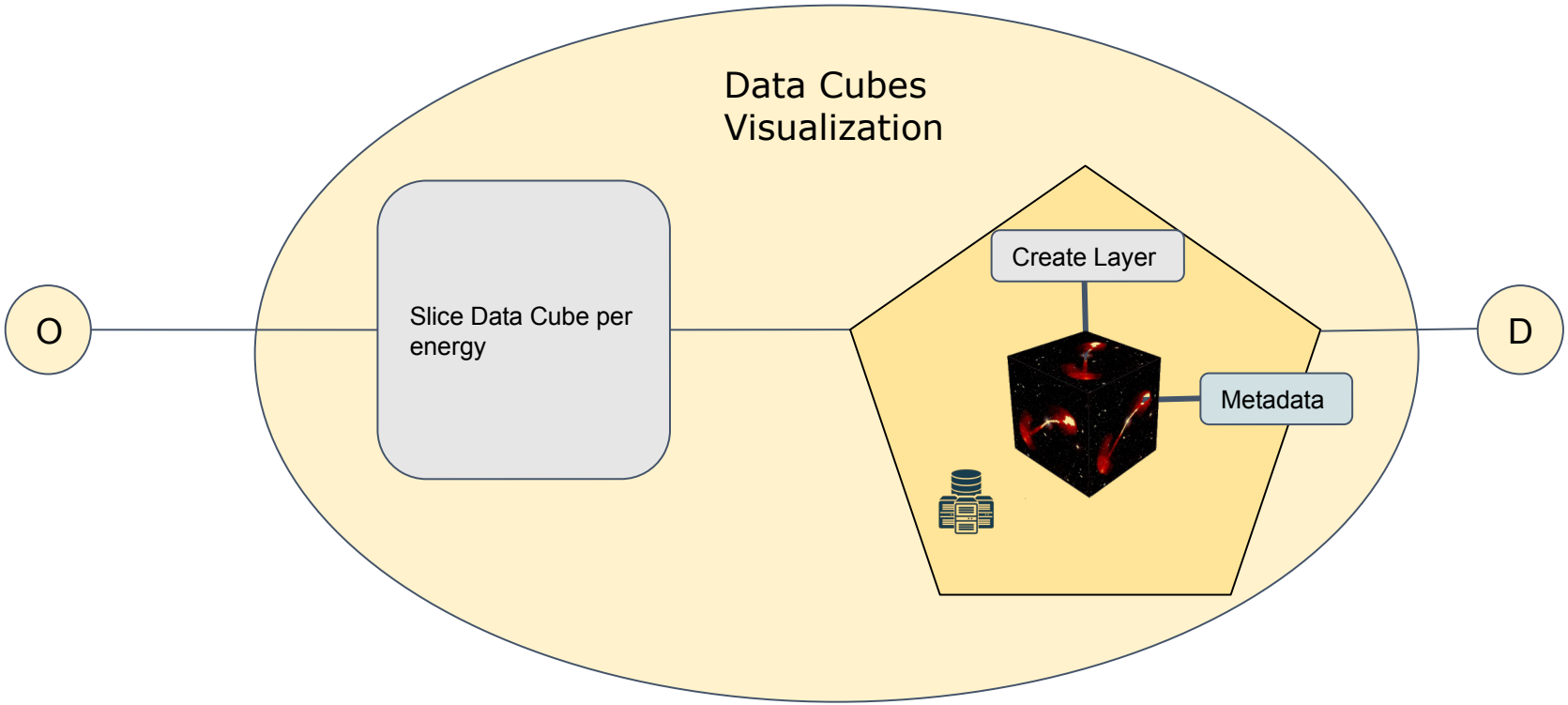
D: Analytics Data

Domain Data as a product



- Discoverable
- Addressable
- Self-describing
- Trustworthy
- Secure
- InterOperable

Data Product in Domain

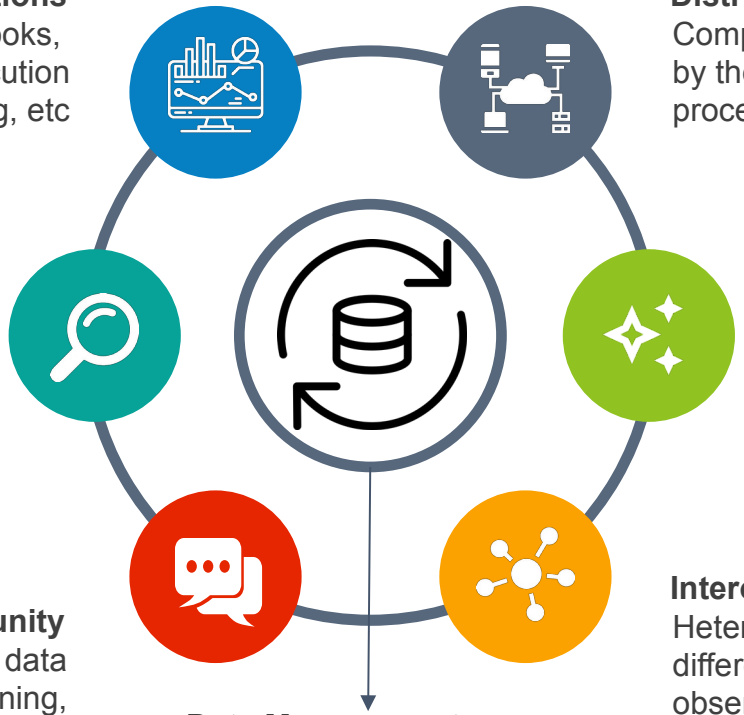



Slice Data Cube per energy

SKA Regional Centre Capabilities Blueprint

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

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



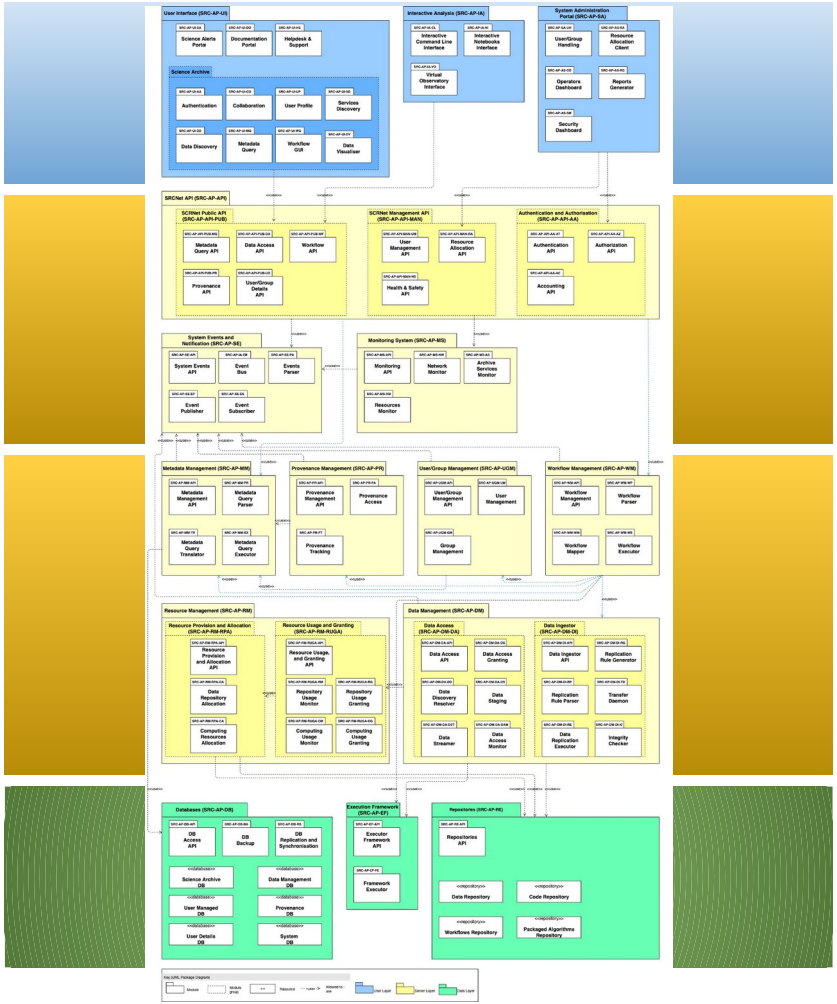
Data Discovery
Discovery of SKA data from the SRCNet, local or remote, transparently to the user

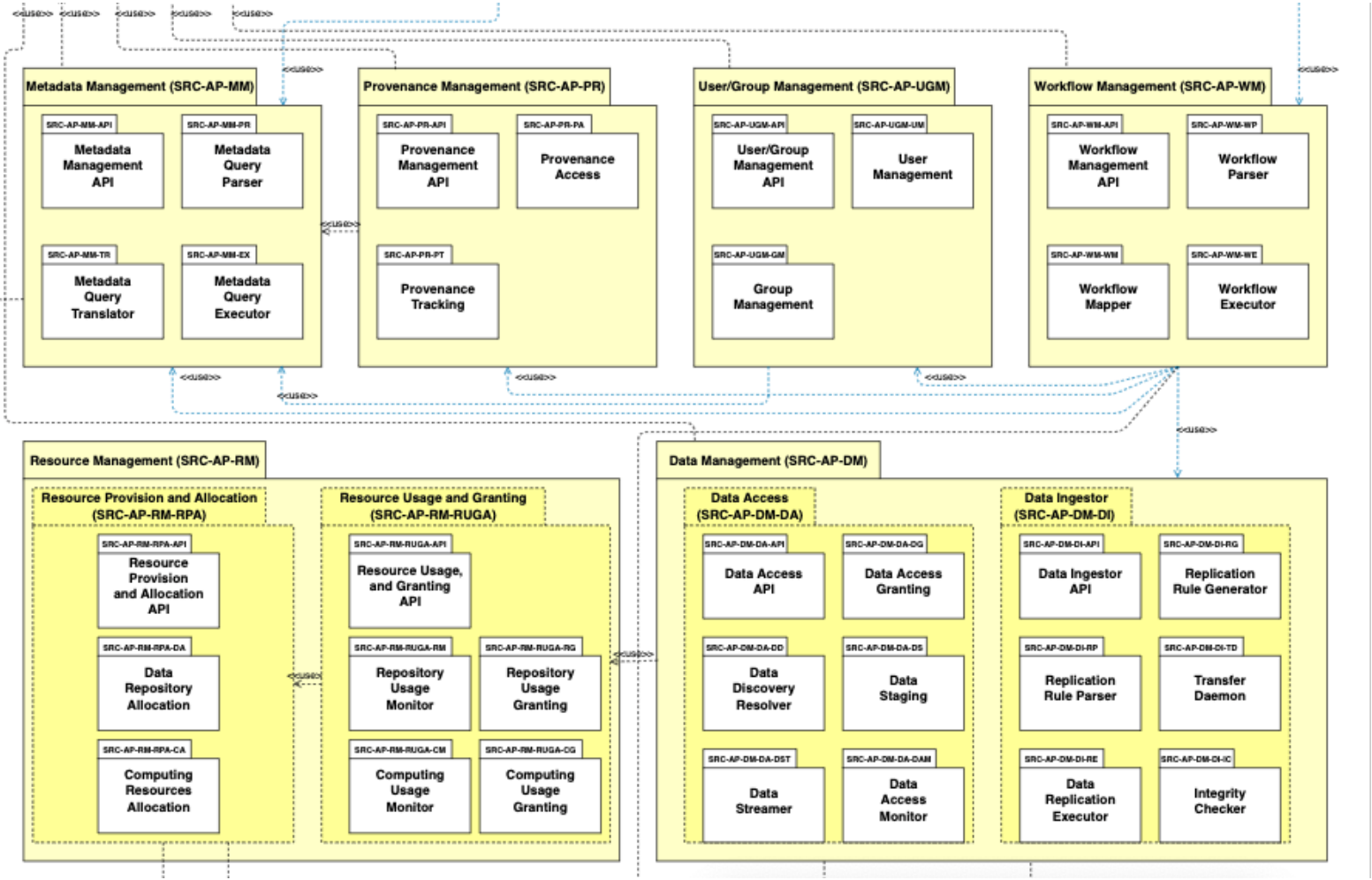
Visualization
Advanced visualizers for SKA data and data from other observatories

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


Client Layer
API + Bus
Management Layer
Resources Layer

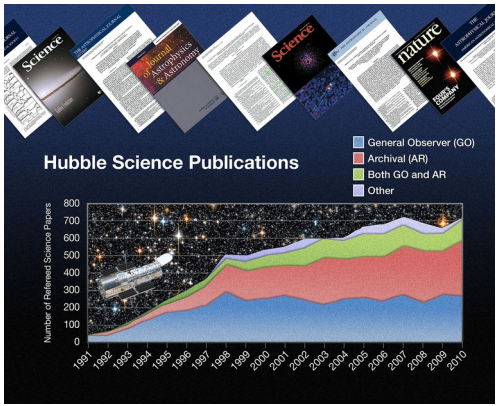

Metadata Management (SRC-AP-MM)
SRC-AP-MM-API
Metadata Management API
SRC-AP-MM-PR
Metadata Query Parser
SRC-AP-MM-TR
Metadata Query Translator
SRC-AP-MM-EX
Metadata Query Executor
Provenance Management (SRC-AP-PR)
SRC-AP-PR-API
Provenance Management API
SRC-AP-PR-PA
Provenance Access
SRC-AP-PR-PT
Provenance Tracking
User/Group Management (SRC-AP-UGM)
SRC-AP-UGM-API
User/Group Management API
SRC-AP-UGM-UM
User Management
SRC-AP-UGM-GM
Group Management
Workflow Management (SRC-AP-WM)
SRC-AP-WM-API
Workflow Management API
SRC-AP-WM-WP
Workflow Parser
SRC-AP-WM-WM
Workflow Mapper
SRC-AP-WM-WE
Workflow Executor
Resource Management (SRC-AP-RM)
Resource Provision and Allocation (SRC-AP-RM-RPA)
SRC-AP-RM-RPA-API
Resource Provision and Allocation API
SRC-AP-RM-RPA-DA
Data Repository Allocation
SRC-AP-RM-RPA-CA
Computing Resources Allocation
Resource Usage and Granting (SRC-AP-RM-RUGA)
SRC-AP-RM-RUGA-API
Resource Usage, and Granting API
SRC-AP-RM-RUGA-RM
Repository Usage Monitor
SRC-AP-RM-RUGA-RG
Repository Usage Granting
SRC-AP-RM-RUGA-CM
Computing Usage Monitor
SRC-AP-RM-RUGA-CG
Computing Usage Granting
Data Management (SRC-AP-DM)
Data Access (SRC-AP-DM-DA)
SRC-AP-DM-DA-API
Data Access API
SRC-AP-DM-DA-DG
Data Access Granting
SRC-AP-DM-DA-DD
Data Discovery Resolver
SRC-AP-DM-DA-DS
Data Staging
SRC-AP-DM-DA-DST
Data Streamer
SRC-AP-DM-DA-DAM
Data Access Monitor
Data Ingestor (SRC-AP-DM-DI)
SRC-AP-DM-DI-API
Data Ingestor API
SRC-AP-DM-DI-RG
Replication Rule Generator
SRC-AP-DM-DI-RP
Replication Rule Parser
SRC-AP-DM-DI-TD
Transfer Daemon
SRC-AP-DM-DI-RE
Data Replication Executor
SRC-AP-DM-DI-IC
Integrity Checker
<<use>>

SRC Network global capabilities

- Every node is an instance of the blueprint
- Interconnections are done using agreed APIs, using FAIR and VO protocols where available
- Collectively meet the needs of the global community of SKA users
- Anticipate heterogeneous SRCs, with different strengths

SRCNet principles: Use of Standards

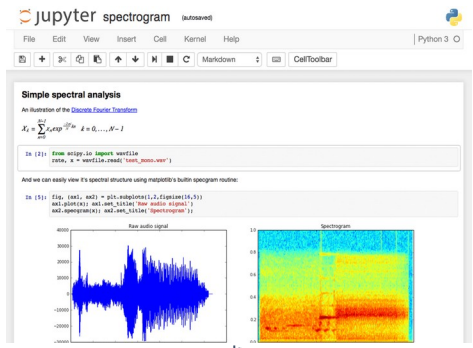
- Build SKA science archive around FAIR and IVOA standards
- Ensure interoperability with other archives and other experiments
- Strong adherence to the FAIR principles
- Give credit appropriately to all contributors to a team

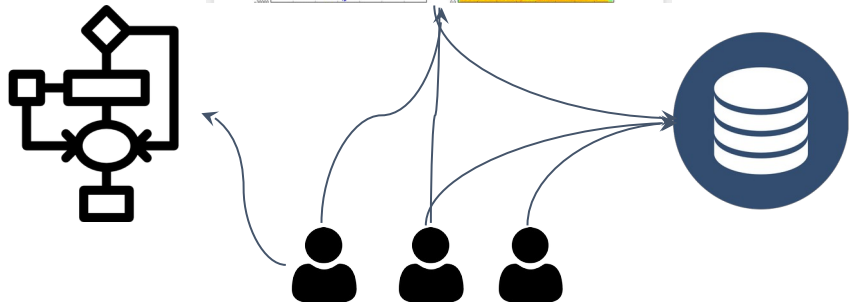


SRCNet principles: Collaboration and Reproducibility

- Most SKA projects will be collaborative
- SRCs will provide collaborative tools
  - Sharing components
  - Single Sign-on

- Support to workflows
- Provenance metadata
- Science Reproducibility at the level of workflows is essential as data should not be downloaded
- 43 requirements talking about provenance for the SRC Net, 26 L1 level





COMMON WORKFLOW LANGUAGE

jupyter

IVOA

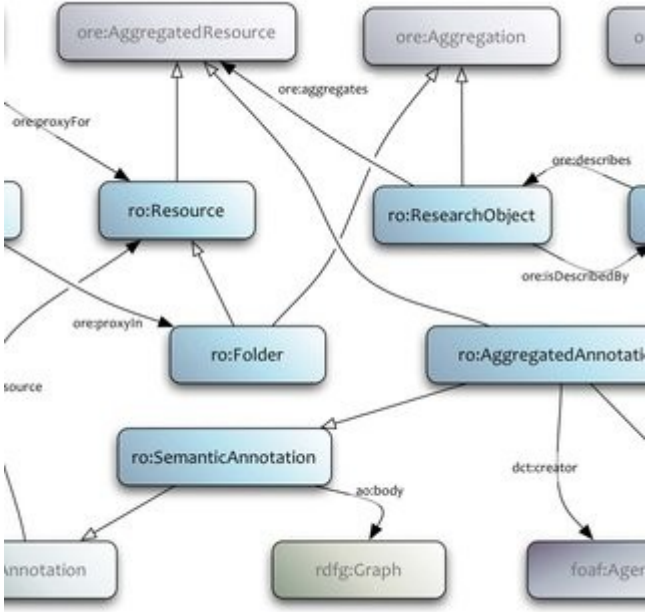
VOSpace

SRC Net Provenance Requirements Examples

- SRC-102 Interactive Analysis: Interactive analysis actions should be permit the serialization in the form of reproducible workflows
- SRC-169 Qualified references to other (meta)data within ADPs: Advanced Data Product would contain provenance metadata to characterize its generation
- SRC-138 and similar requirements. Provenance should be maintained for all products (including movements within the network)
- SRC-118 Uniform vocabulary in the archive: Provenance is part of the vocabulary shared by all the SRC Net node
- SRC-22 Reproducibility of results
- SRC-108 Long term preservation of metadata and provenance metadata for Observatory and Advanced Data Products

SRC Net Provenance non-written yet Requirements

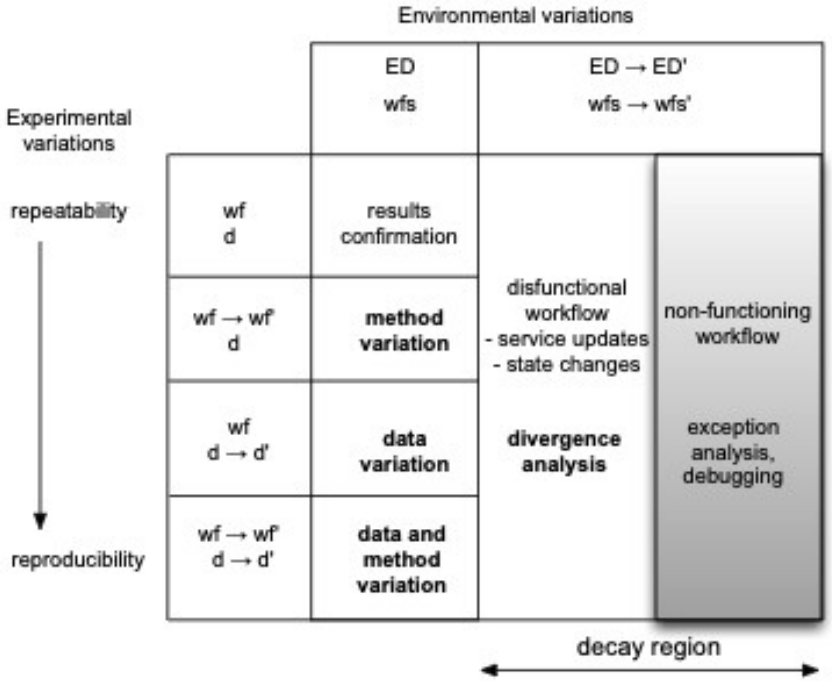
- Provenance mapper
- Workflow editor and visualizer (e.g. Taberna)
- Support to needed execution framework
- Full characterization of software elements
  - Software Versioning
  - Software Access
  - Input/Output parameters/results
  - Error handling
  - Execution
    - Containers?
    - Specific Hardware and execution needs
    - Data Domain



Replicability vs Reproducibility

- **Reproducibility**: Ability to repeat an experiment with minor differences to the original experiment, but still achieving the same qualitative result.
  - E.g. Notebooks or pipelines
- **Replicability**: Ability to exactly reproduce results by running exactly the same experiment.
  - No changes in parameters or environment
  - Frozen data
- First one, quite important to generate pipelines, second to replicate results of a given experiment/paper
- Probably both things are needed for SKA with different level or priority

Mapping the reproducibility space



Provenance and data differencing for workflow reproducibility analysis Missier, P.; Woodman, S.; Hiden, H.; and Watson, P. Concurrency and Computation: Practice and Experience, . 2013

Reproducibility in Cloud

Listing 1.1. HyperFlow workflow description (fragment).

```
processes: [ {
  "name": "alignment_to_reference",
  "function": "k8sCommand",
  "config": {
    "executor": {
      "executable": "bwa-wrapper",
      "args": [ "mem", "-t", "2", "-M",
        "Gmax_275_v2.0.fa",
        "USB-001_1.fastq",
        "USB-001_2.fastq"
      ],
      "cpuRequest": "1",
      "memRequest": "500M",
      "stdout": "20180321-083514-USB-001_aligned_reads.sam"
    }
  },
  "ins": [1,2,3,4,5,6,7,8,9,10,11,12],
  "outs": [13]
}, ... ],
signals: [ {
  "name": "Gmax_275_v2.0.fa",
  "type": "file",
  "size": "990744229",
  "md5sum": "3aa6cf1962f5260cf1405e82efb25c71"
}, ... ]
```

Reproducibility of Computational Experiments on Kubernetes-Managed Container Clouds with HyperFlow, V. Krzhizhanovskaya et al Computational Science – ICCS 2020. 2020 May 26; 12137: 220–233.
Published online 2020 May 26. doi: 10.1007/978-3-030-50371-0_16

Other thoughts

- How to combine data from other missions/data providers and maintain reproducibility?
  - Copy at SRC Net?
- Workflow Serialization Language (CWL?)
- Workflows simplification?
  - Graph analysis and proposal of more optimal execution
  - Analysis of resources
- How to combine provenance from different centres
  - Provenance discovery (Prov-TAP is enough?)
- Specific provenance fields for execution through Provenance extension?

Thanks for your attention