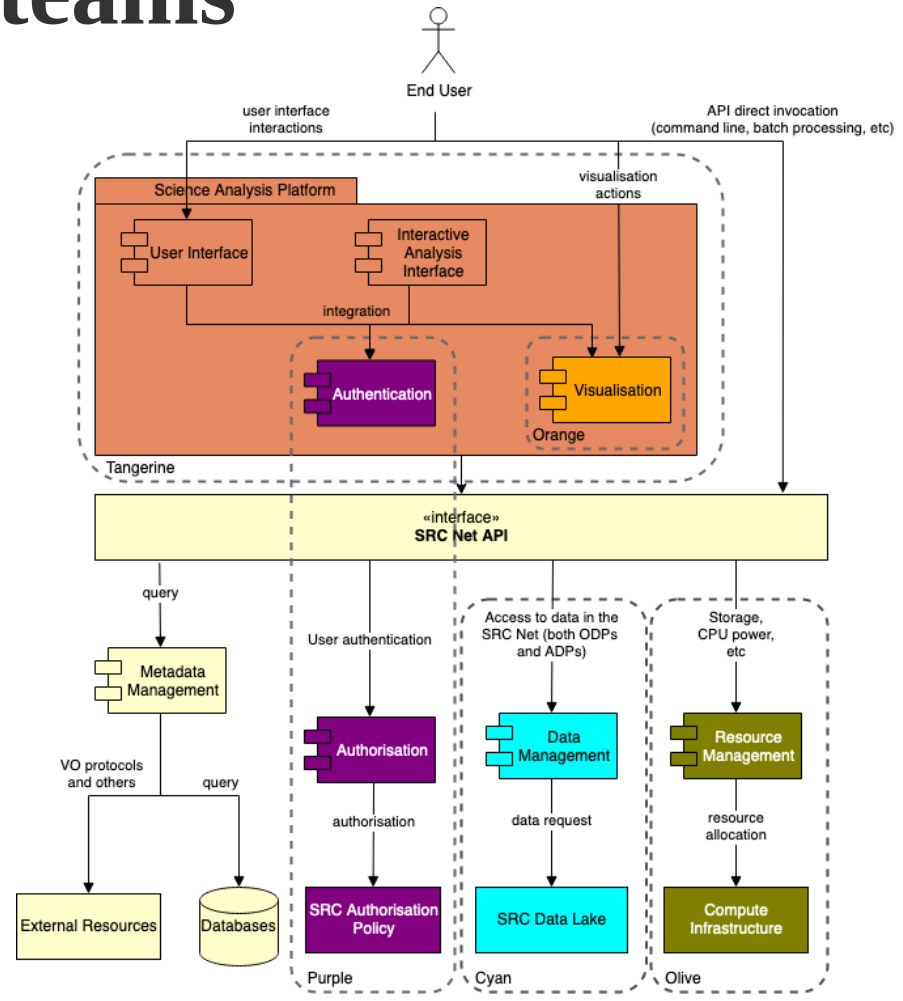
Evaluating existing science platforms for SKA development

Team Tangerine, SKAO Prototyping Team
(Total members ~ 40)

- Presented by: Rohit Sharma, FHNW, Windisch

12th Jan 2023
Winter SKACH Meeting, University of Basel, Switzerland

Prototyping teams



Contents of the talk

- What is a Science platform?
- Requirements of SRCNet
- Challenges
- List of existing science platforms
- Down listing the science platforms
- Evaluating Criteria for system architecture
- Evaluating selected platforms

What is a Science platform?

- Analysis platform permitting scientists a collaborative handling of large and diverse datasets with access to large-scale computing resources.
- Linked to a specific science projects
- Must provide features that are relevant to the user base and avoid any pitfalls

General Criteria for a Science Platform

- Consistency
- Scalable
- Reproducibility
- Usabilty
- Reliability
- Accessibility

Some Requirements at high-level SRC system requirement

- Provide accessible and straightforward interfaces / must be available to widest possible user base
- Provide expert-compute users with maximal flexibility
- Provide appropriate interfaces to the lower-level services and tools deployed across the SRCNet
- Platform will only provide users with access to services and/or data products to which those users are entitled and will include limited, public, anonymous user access.

Challenges Involved

- Data Volume: 700 PB/year
- Data Distribution of ODPs – image cubes, uv-grids, calibrated vis, pulsar timing solutions, calibration data and metadata
- Typical design and working of the platform is driven by a scientific projects, but SKA has many science projects

Existing Science Platforms

List of various suggested science platforms

	ID	Science Platform	Acronym (if any)	Links, references, etc	Notes
1	SP1	Service for Web-based ANalysis	SWAN	• ScienceDirect Article: SWAN • https://www.sciencedirect.com/science/article/abs/pii/S0167739X16307105?via%3Dihub • https://swan.web.cern.ch/swan/ • https://swan.docs.cern.ch/intro/what_is/	
2	SP2	SciServer		• ScienceDirect Article: SciServer - paywall-free access via arXiv at SciServer (Taghizadeh-Popp et al, 2020) • https://www.sciserver.org/	
3	SP3	Rubin Science Platform		• LSST Science Platform Vision Document • LSST UK Plans (LIneA Workshop): https://youtu.be/nwHd4RWCgYU?t=2302 • Rubin Science Platform: https://data.lsst.cloud/ • Rubin (LIneA Workshop 2021 link): https://youtu.be/nwHd4RWCgYU?t=1050	
4	SP4	European Strategy Forum on Research Infrastructure - Science Analysis Platform	ESCAPE ESAP	• https://projectescape.eu/services/esfri-science-analysis-platform • Rucio + JupyterHub + OpenData: https://indico.in2p3.fr/event/23333/sessions/14433/attachments/62070/84839/lapp-escape-wp2-arturos-220121.pdf • ADASS proceedings – need to track down a link. • LIneA Workshop 2021 link: https://youtu.be/YHUn93BwZoE?t=2935	
5	SP5	CyVerse		• Original paper: PLOS • https://cyverse.org/ • Science APIs: https://cyverse.org/Science-APIs • Data Commons Serivice: https://datacommons.cyverse.org/ • LIneA Workshop 2021 link: https://youtu.be/YHUn93BwZoE?t=4944	
6	SP6	MeerKat Toolbelt		• Introduction to MeerKat Toolbelt • LIneA Workshop 2021 link: https://youtu.be/jTDj-YSUU-k?t=123	
7	SP7	China Virtual Observatory	China VO	• ScienceDirect Article: China-VO (arXiv) • http://www.china-vo.org/doc.html	
8	SP8	Canadian Advanced Network for Astronomical Research	CANFAR		
9	SP9	French Space Agency - CNES Platforms			
10	SP10	CDS Astronomy Data Centre	CDS		
11	SP11	NASA HEASARC			
12	SP12	NOIRLab Astro Data Lab			
13	SP13	European Space Astronomy Centre - Science Data Centre	ESAC - SDC		
14	SP14	Spanish SRC	SPSRC		
15	SP15	European Space Agency	ESA		
16	SP16	WholeTale			
17	SP17	ESO Science Archive	ESO		
18	SP18	Pangeo			

16	SP16	WholeTale		• https://wholetale.org	
17	SP17	ESO Science Archive	ESO	• http://archive.eso.org/cms.html	• The ESO Archive doesn't (as far as I know...) have notebooks or workflows, but it does provide data discovery, exploration, access, etc.
18	SP18	Pangeo		• https://pangeo.io	• Geophysics, not astronomy, but many of the technologies and considerations are likely similar.
19	SP19	NEWT	NERSC	https://ieeexplore.ieee.org/document/5676125 https://github.com/NERSC/newt-2.0/	• Project looks quite abandoned (last commit is from 2019 across all forks) so may be interesting architecture-wise.
20	SP20	Astronomy Commons Platform	UW DIRAC	https://arxiv.org/abs/2206.14392	Small scale but automated deployment on AWS
21	SP21	REANA		• https://www.reanahub.io	• Another CERN effort. Maybe complementary to SWAN – this provides workflows, while SWAN provides notebooks (I think...).
22	SP22	Galaxy		https://galaxyproject.org/ https://sciencegateways.org/index.php?option=com_content&view=article&id=405:galaxy-project-enabling-an-active-global-research-community&catid=32:w	Galaxy is a science platform aimed primarily at the life sciences.
23	SP23	WLCG: a distributed infrastructure for large-scale scientific computing	WLCG	• https://wlcg.web.cern.ch/ • Technical Documents: https://wlcg-docs.web.cern.ch?dir=technical_documents • LIneA Workshop 2021 link: https://youtu.be/YHUn93BwZoE?t=1015 • https://twiki.cern.ch/HEPTape • https://wlcg.web.cern.ch/activities/working-groups	• Deployed and used in CERN
24	SP24	nanoHUB		https://nanohub.org/simulate/	Modelling and simulation gateway for nanotechnology scientists and engineers. Jupyter notebook-type interface running transparently in a scientific computing cloud.
25	SP25	The Virtual Imaging Platform	VIP	Paper at: https://docs.google.com/a/nd.edu/viewer?a=v&pid=sites&srcid=bmQuZWR1fGI3c2cyMDE4fGd4OjNmOGRkMDIzOTEyNTRmZjI Platform at: https://vip.creatis.insa-lyon.fr/documentation/	A medical imaging platform allowing new applications to be imported through Docker.
26	SP26	InterMine Science Gateway		http://intermine.org/documentation/	A science platform for biology data, providing web-enabled access to large databases via SQL, with tools for building queries. No notebook element, but provides extensive API access from Perl, Python, Ruby, and Java.
27	SP27	HubZero		https://hubzero.org https://ieeexplore.ieee.org/document/5432299	The generalised version of the nanohub platform. It is somewhat unclear to me what HubZero exactly is but I think it is a framework that can be used to build a Science Analysys Platform/Gateeway.

Down Selecting the List

- Galaxy:
- Cyverse:
- ESA Datalabs
- WholeTale
- CERN REANA
- CERN SWAN
- CANFAR
- ESCAPE ESAP
- SciServer

- Open Source
- Maintenance statu
- Data discovery
- Notebooks
- Workflows
- Software Distribution
- API access
- Deployment Platform
- Usability and Accessibility
- Production status

Scoring	Platform ID	Science Platform	Implementation Language(s)	Major frameworks and technology	Open Source
18,00	22	Galaxy	Python, javascript, others	RStudio, Jupyter	2
16,33	14	Spanish SRC		Openstack, On-Demand Elastic Clusters, On-Dema	1
16,00	5	CyVerse	Go, C++, Javascript, Python	iRODS, Django, Agave API, React.js, Material-UI	0
14,67	13	ESA Datalabs	most is in Python in the JWST section		2
14,00	16	WholeTale		The Whole Tale platform leverages and extends a v	2
14,00	21	CERN REANA	Python	Kubernetes, HTCondor, Slurm, Docker, Singularity	2
13,67	4	ESCAPE ESAP	Python, Javascript	Django, React.js	2
13,00	24	nanoHUB		Linux, Apache web server, a MySQL database, PH	1
12,00	2	SciServer	C# ASP.NET Javascript		1
11,67	8	CANFAR	Java	kubernetes on OpenStack, Harbor, 0	2
11,67	17	ESO Science Archive	JSky - Java Dataset - web based access		1
11,00	1	CERN SWAN	Python, Javascript	JupyterHub, Spark, EOS, CernVM-FS	1
10,33	12	NOIRLab Astro Data Lab	Python	JupyterHub, Postgres+Q3C	1
10,00	25	The Virtual Imaging Platform	Java, Javascript, Python, HTML		2
10,00	28	AMDA http://amda.irap.omp.eu	PHP, Javascript, HTML	C	1
9,67	3	Rubin Science Platform	Python, C++	kubernetes, JupyterHub, Dask, Firefly	1
9,67	9	French Space Agency -CNES Platforms There are a bu	Java, R	Vue.js, Leaflet.js	1
9,00	29	VirES	Python, javascript HTML	WPS	1
8,00	26	InterMine Science Gateway	Java, SQL	PostgreSQL	2
8,00	27	HubZero	PHP, Javascript, HTML	OpenVZ (www.openvz.org) Apache Httpd Web Se	2
7,00	18	Pangeo Cloud	Python	xarray, iris, Dask, Jupyter	1
5,33	6	MeerKat Toolbelt	Â Python 3.8+ (for CASA 6.5+)	CASA Singularity container for parallel processing	1

Want/must have:

- Fully maintained & well documented (easy to find install, details etc)
- Source code available / open-source
- Rely on the summary of vision doc for criteria taken from vision doc
- System to be able to interact with remote content

Desired platforms that have:

- Modularity workflows and notebooks

Cross off list the platforms that have:

- Features/modules that can't be developed together OR not useful for us(eg. the platform is intended to be running on laptop)

Matrix Evaluation

Reference: https://confluence.skatelescope.org/display/SRCSC/The+Selection+of+the+Science-Platform+Shortlist

Assessing Architecture of SP

- **General**

- E.g. Does the platform use virtual machines? Does the platform use containerisation to provide functionality to the user, if so what tool is used (Docker/Singularity)?

- **Federation**

  - Is the platform distributed between several sites? Are users registered at individual sites or are they registered to use resources at all sites?

- **Computational Layer**

  - How does the computational layer connect to the server layer & presentation layer? What is the underlying infrastructure (HPC, public cloud, private cloud, hybrid)?

- **Presentation Layer**

  - Does the presentation layer have direct access to the data layer? Does the presentation layer make use of some kind of pattern (e.g. Gateway pattern) to allow extensibility?

- **User Customization**

  - What support is provided for user-supplied software and notebooks? If notebooks are used, can the users add their own dependencies?

- **Server Layer**

  - How can the platform be scaled to allow for an increase in the number of users and/or data? Does the platform use orchestration? If so, what tool does it use (e.g. Kubernetes)?

Evaluation Criteria (under Progress)

- Interfacing with data archive
- Data Quality Control
- Management of User Accounts
- Management of Compute Resources
- Visualization
- Tabular data
- Software access and data processing tools
- User support

Q. What do we want to get out of the assessment?

- To find the perfect platform

- To find components to build upon/from

- To gather ideas

... other

Q. What headers for the matrix (columns)

- High level summary of tasks (bullet points = +1)

- Maybe extra points for fancy stuff

Q. Do we need to score all the use cases for the existing platforms, for example the use case 9a: Creating a user account. Is this useful for our assessment?

Marks to be given out of 10, for example:

0 = No, not possible

5 = Yes, exactly what was wanted but with moderate difficulty / Can do something similar with ease

10 = Yes, exactly what was wanted and can be done easily

Summary

- Defined high level essential criteria of SKA science platform
- Downlisted initial science platforms list
- Accessed some science desirable science platforms
- Architecture assessment under progress
- Future: Deployment and testing