The Why and What of Federated Infrastructure

Federated Infrastructures

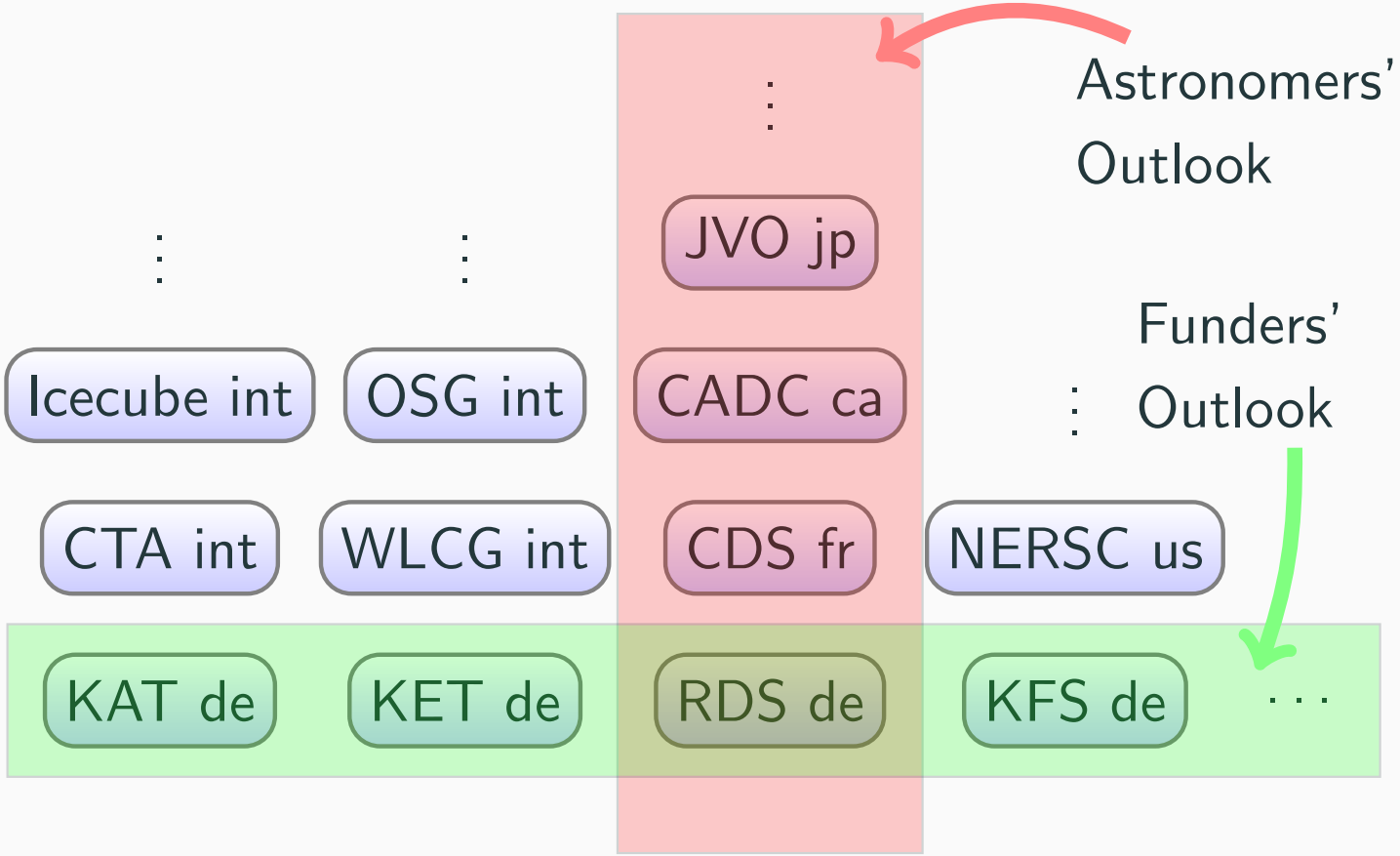
What? Networked IT systems following common standards with shared facilities for discovery, authentication, etc.

Why? Uniform interfaces let users freely switch between data sets, operators and lets software work consistently across them. This fosters resource utilisation and re-use and avoids lock-in.

But Working out the standards and convincing individual operators to open up their systems means work for the implementors.

Hence there is a long-term need for dedicated funding and incentives.

The International Perspective



Infrastructures relevant to researchers must federate with their international peers. Hence, federation efforts necessarily reach in two dimensions.

Infrastructures to Federate

Authentication and Authorisation

What? Recognising actors and their affiliations as a basis of almost everything else.

Why? Limiting access, preserving environments ("uploaded table is still there"), enforcing quota etc.

How? X.509, Oauth, etc. [Helmholtz-AAI]

Challenges Interoperable auth, in particular outside of web browsers, stable, cross-discipline, internationally federated infrastructure.

Storage

What?	Distributed storage of data sets that usually are too large for a single site; provision of local caching for working sets.
Why?	Facilitate data sharing between researchers and institutions, enable processing of huge data sets.
How?	dCache, XRootD, VOSpace; rucio for management.
Challenges	Federation, unified namespace, automatic data placement and replication.

Compute

What? Remote access to large-scale shared computing resources with transparent access to federated storage resources.

Why? Enable processing of data sets too large to move and/or process locally. Improved and sustainable utilisation of opportunistic or previously isolated resources.

How? WLCG middleware; COBaID/TARDIS for job orchestration; AUDITOR for monitoring and accounting.

Challenges Roll-out, community adaptation, continued operation and support, scaling methodologies for green operations.

Analysis Platforms

What?	Interactive analysis of large-scale data on federated resources.
Why?	User workstations are too small or have too little network bandwidth to process huge data volumes with sufficiently short response time.
How?	Jupyter Hub, Apache Spark, etc.
Challenges	Avoid lock-in to individual solutions; scalability, interconnectivity to other federated resources.

Archiving

What? Making previous results and observations Findable, Accessible, and Interoperable...

Why? ...for Re-usability, i.e., to enable new science with existing data (re-use for sustainability)

How? There's the VO with a decade-old large federation; Zenodo; cross-disciplinary discovery portals like b2find, openAIRE.

Challenges Disciplinary vs. cross-disciplinary engines (and metadata schemes $\rightarrow$ RDM), interoperable access to data, validation, monitoring. And of course: Clients!

Vision North Sea

The more shared, interoperable infrastructures become standard: What about concentrating much of the hardware where there is lots of green power?

Something like this will need to happen (see Sustainability Paper). But what we build in Federated Infrastructures arguably is a necessary condition to make it happen when we do.