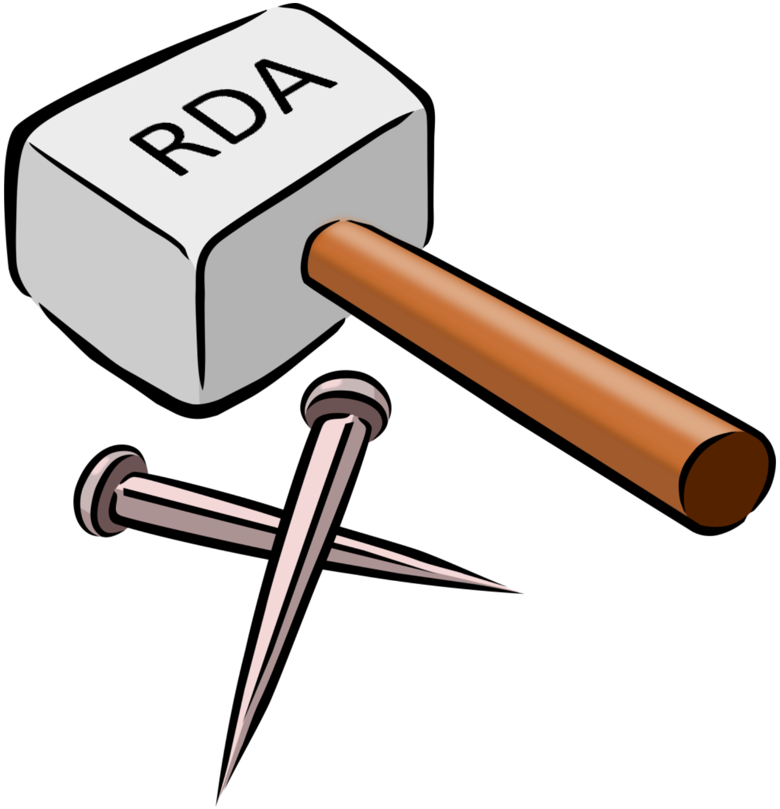
RDA-DE/DINI: Sichtweisen aus der Klimamodellierung

RDA-Deutschland-Treffen

28./29.05.2015, KIT

Tobias Weigel, Stephan Kindermann, Michael Lautenschlager
Deutsches Klimarechenzentrum (DKRZ)

All going well?



We seem to have some tools.
But how are we going to use them now?

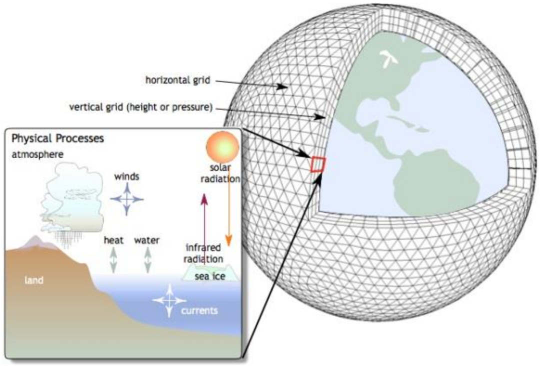
The Earth System Grid Federation and CMIP6

- CMIP6: worldwide coordinated climate simulations (>28 modeling centers, >40 models, CMIP5)

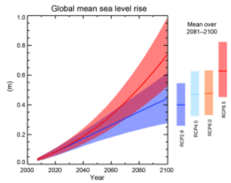
- ESGF data federation: worldwide distributed e-infrastructure for climate data distribution

End users:

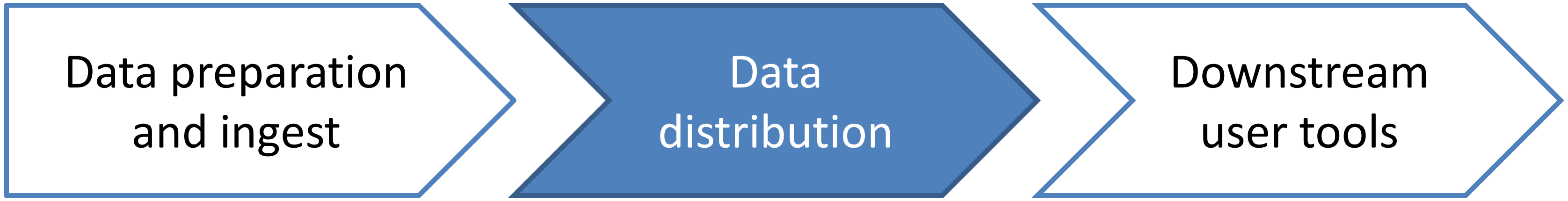
- Climate modeling community
- Climate impact community
- Interdisciplinary



Source:
http://www.cmmap.org/learn/modeling/whatIs2.html

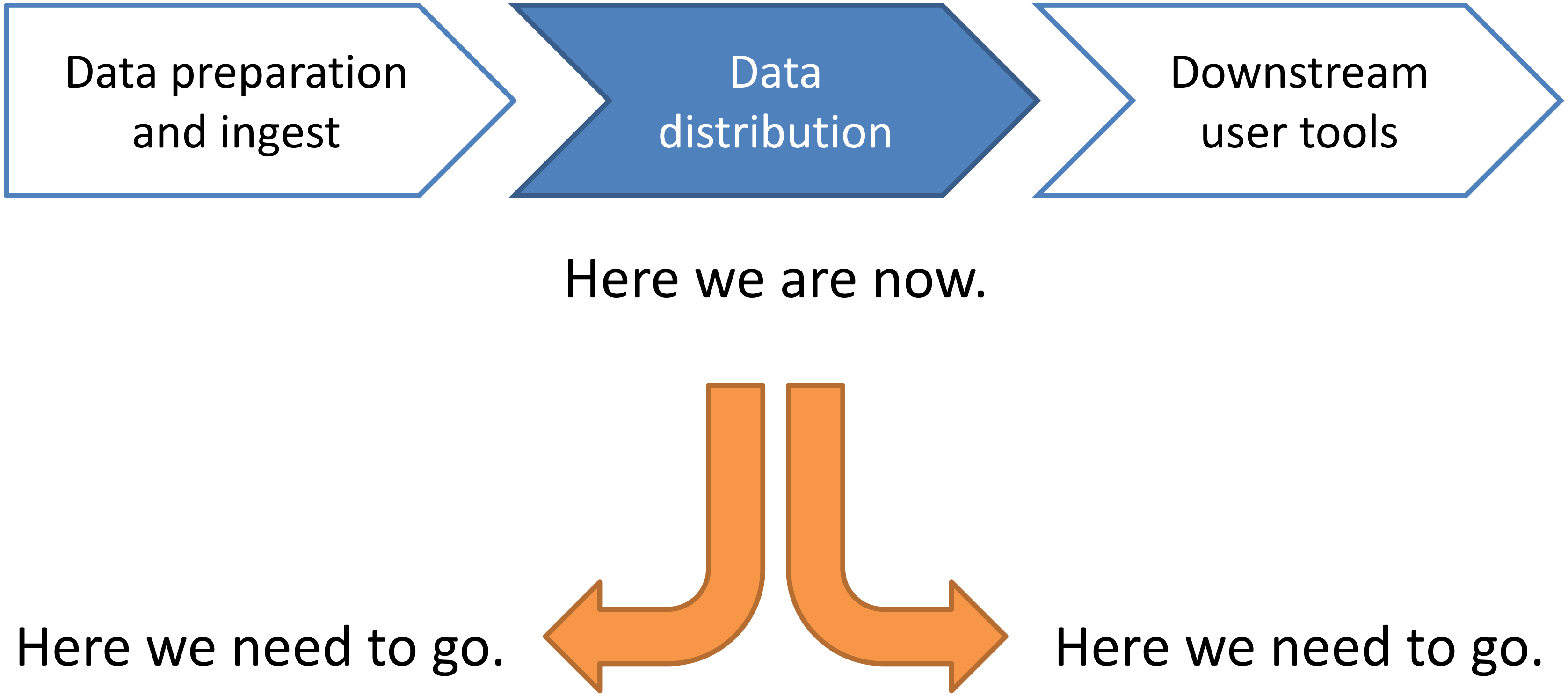


Source: IPCC AR5 Synthesis Report



Data preparation and ingest → Data distribution → Downstream user tools

Future steps: surrounding distribution



Data preparation and ingest

Data distribution

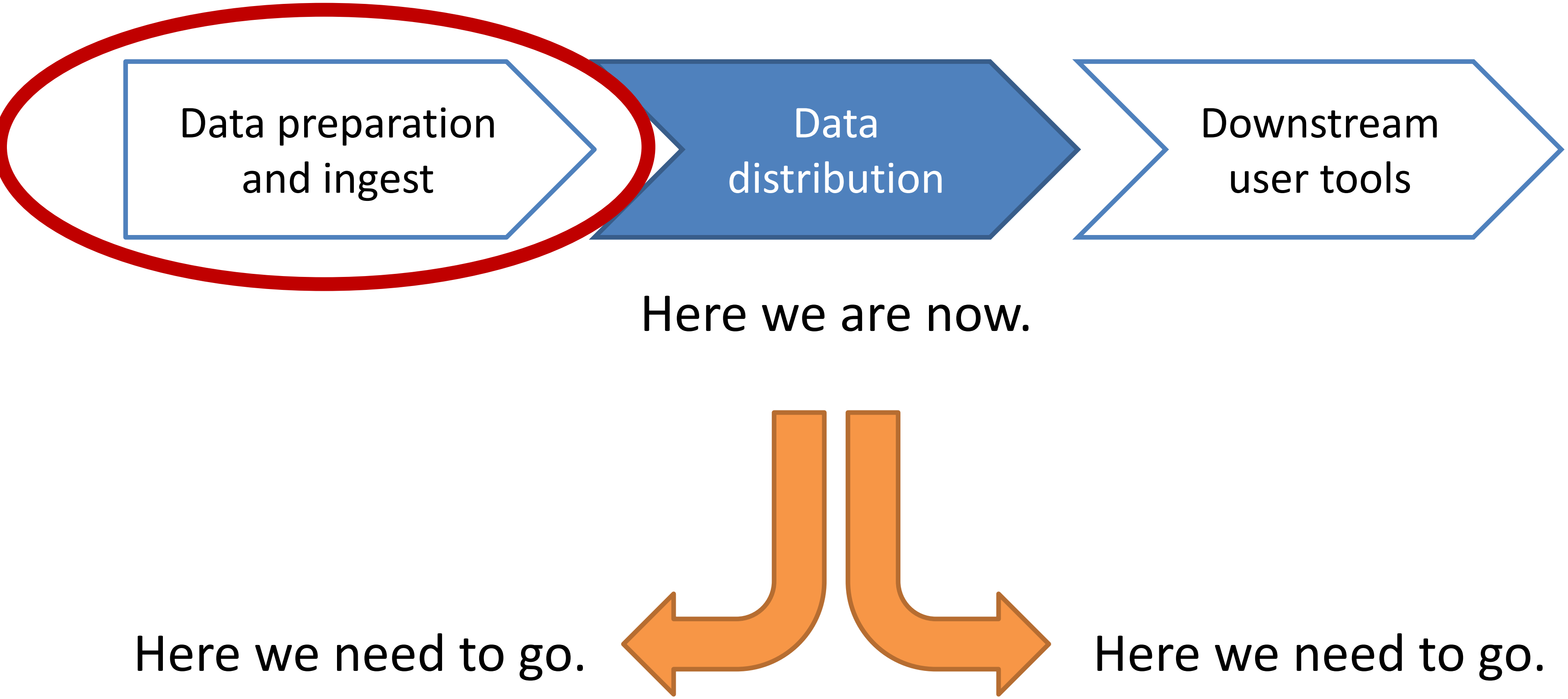
Downstream user tools

Here we are now.

Here we need to go.

Here we need to go.

Future steps: surrounding distribution



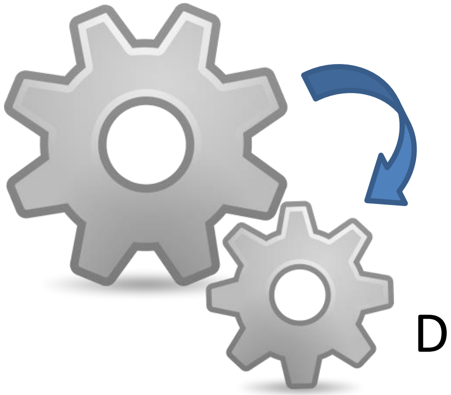
Here we are now.

Here we need to go.

Here we need to go.

Data preparation: Unified workflows

Modelling groups workflow



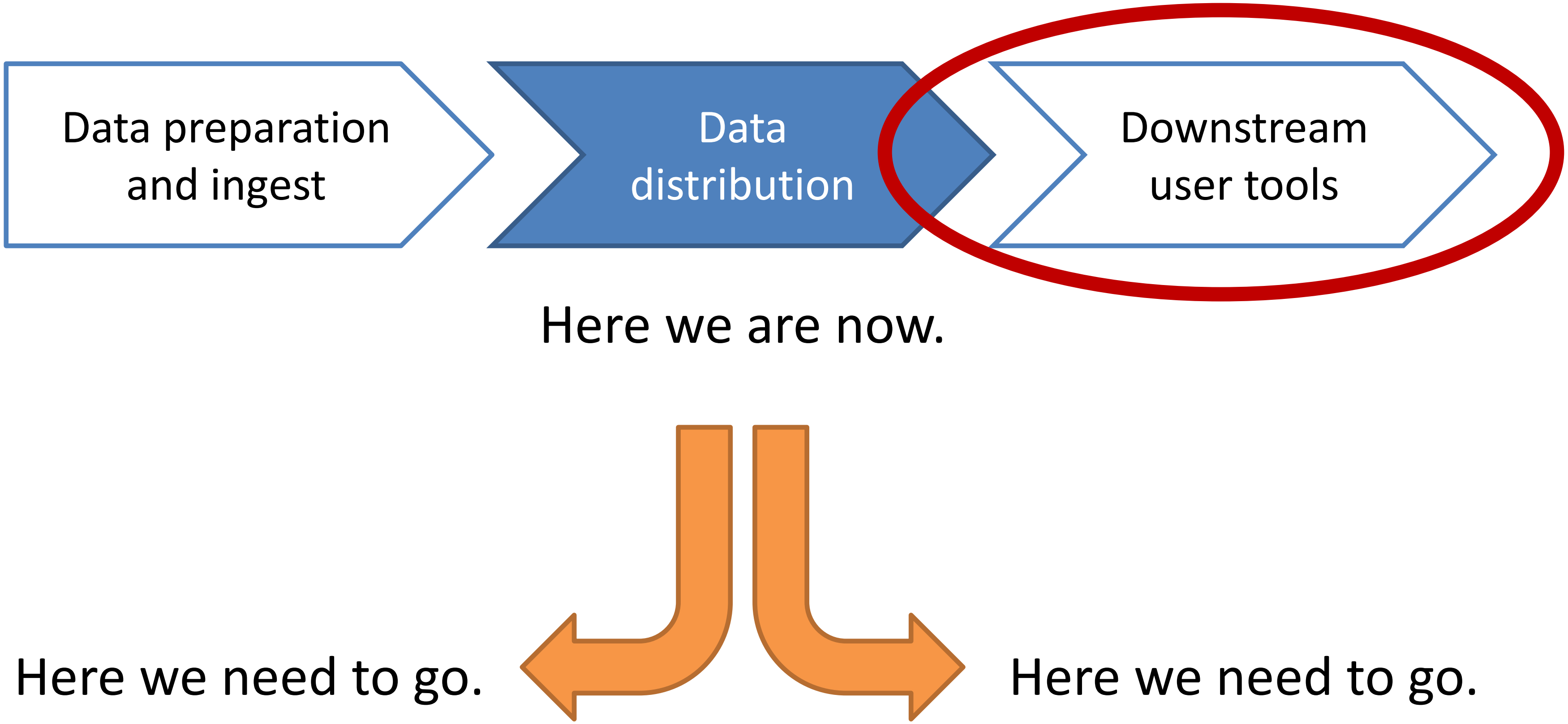
Data center workflow

- We observed friction and workflow disruptions
  - Modellers complain about too high effort for most fundamental data management tasks
  - Data centers suffer from lack of standardization and traceability of data
- Establish “RDA compliant” workflows
  - The Data Fabric metaphor is essential! (Datengewebe?)

Data preparation: Prefabricated policy modules

- Offered by data center, to be used by data producers
- Example: Final data preparation / standardization, including PID assignment and checksum verification
- What is needed: black-boxing, documentation adequate for scientists, high quality – ease of use!
  - Modules must be of operational quality and thoroughly tested, otherwise the users will reject them
  - Scientists will not develop or test modules; we have to do it for them
  - Modules most likely not used directly, but wrapped in existing or customized modules

Future steps: surrounding distribution



Data preparation and ingest

Data distribution

Downstream user tools

Here we are now.

Here we need to go.

Here we need to go.

Downstream user tools

- ESGF provides data and metadata distribution
  - Metadata includes e.g. scientific model descriptions, quality, attribution/citation
- This is a necessary task, but users outside the core community need to have additional benefits directly at their fingertips
  - PIDs, PITs, registries etc. provide great potential
- Also, unification must go beyond large projects like CMIP6 and cover the long tail

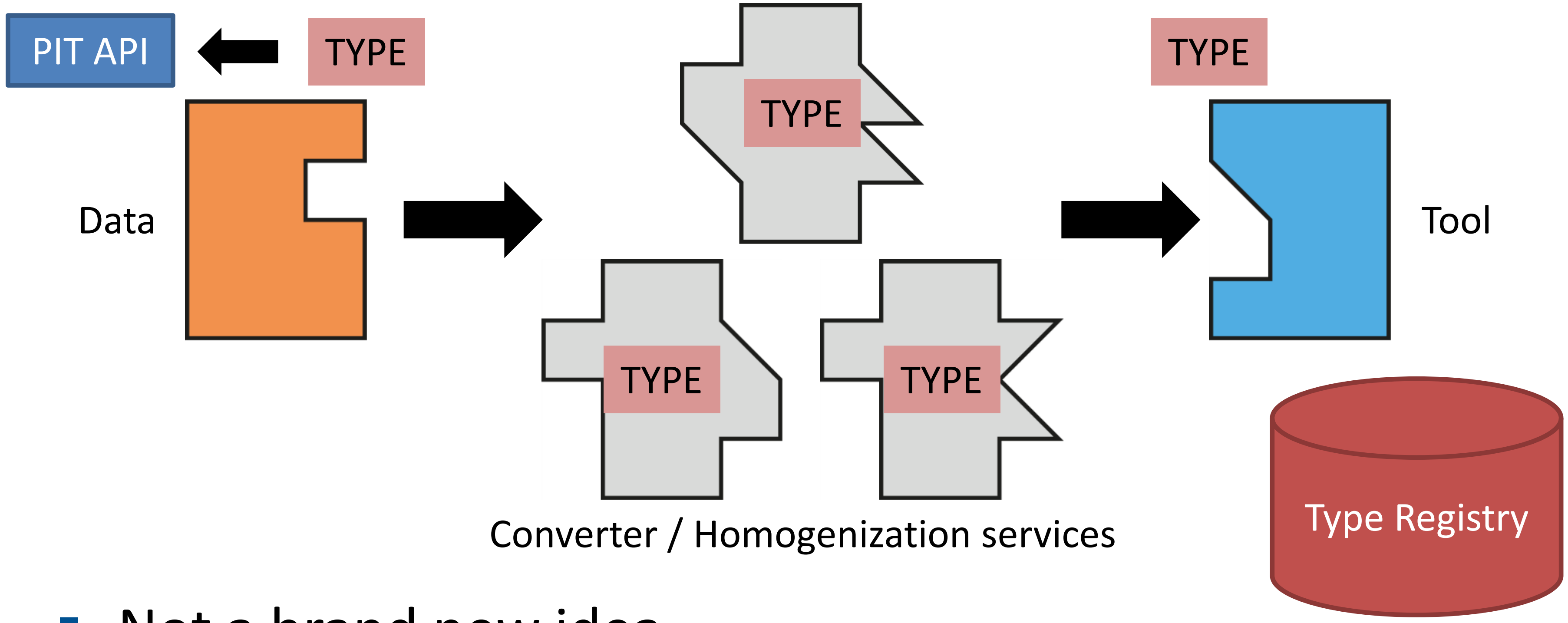
Downstream user tools: Collection building

- Standard way of arranging data for dissemination does not always match the scientist‘s perspectives
  - Hierarchies built according to different criteria
  - There is more than one way to arrange data
- Physical arrangement of files is not a good option – virtual collections are required
- Virtual collections must become primary objects
  - The physical location of a collection member should not matter – members can come from different sources and cross institutional boundaries; use PIDs!
  - Collections can grow over time, be versioned, annotated – they have their own life cycle independent from their members

Downstream user tools: service interoperability

- Incoming data should bear identifiers and data types (via the type registry)
  - Main distinctions could be between model output and sensor data, most common data formats, some data details such as grids
- The downstream user communities can be highly interdisciplinary!
- Analysis tools require data to be in specific format, grid, ...
  - Example tools currently used in the community: ESMValTool, MIKLIP tool; GIS-based tools
- Small converter services required as intermediaries
- First evolution: Documenting, cataloging, manual discovery
- Second evolution: Automated orchestration

Service discovery via the Type Registry



- Not a brand new idea...
- But: limited description complexity, possibility to make progress across disciplines via their respective e-infrastructures
- Contribute to a larger conversion tool registry

What is needed?

- Collaboration between data centers, modelling groups and downstream users
  - There is a wide range of downstream users from the climate impact research community, but also others
  - Provenance tracing from modellers through dissemination to analysis as a long-term goal (10+ years)
- RDA groups help to communicate and explore ideas.
- Developing a larger number of prototypes with more users however requires project work.
- Feed experience back into RDA processes.

Make efficient use of funding

- Basic PID service infrastructure at EU/global level
  - Problem is too large to solve locally
  - Open challenge: High scalability, elastic federation
- Innovation for individual users and local communities at national level
  - Stay closer to the source – our users and the long tail

Thank you for your attention.

cliparts from opencliparts.org