C2CAMP
(A Working Title)

Managing Digital Objects in an Expanding Science Ecosystem

15 Nov 2017

Larry Lannom
Corporation for National Research Initiatives

C2CAMP
(Cross-Continental Collection & Management Pilot)

- Proposed multi-party distributed test bed based on open specifications across a minimal set of (mostly) existing components and interfaces allowing users to deal with Digital Objects efficiently
- Data producers and managers invited to prototype their work flows and other processes in the distributed test bed
- Solicit the creation of additional components and interfaces as needed to meet the requirements evolving from prototypic use of the test bed
- Demonstrate complex scientific workflows for data processing harmonized and automated across communities by using interchangeable infrastructure components and a structured resource market approach

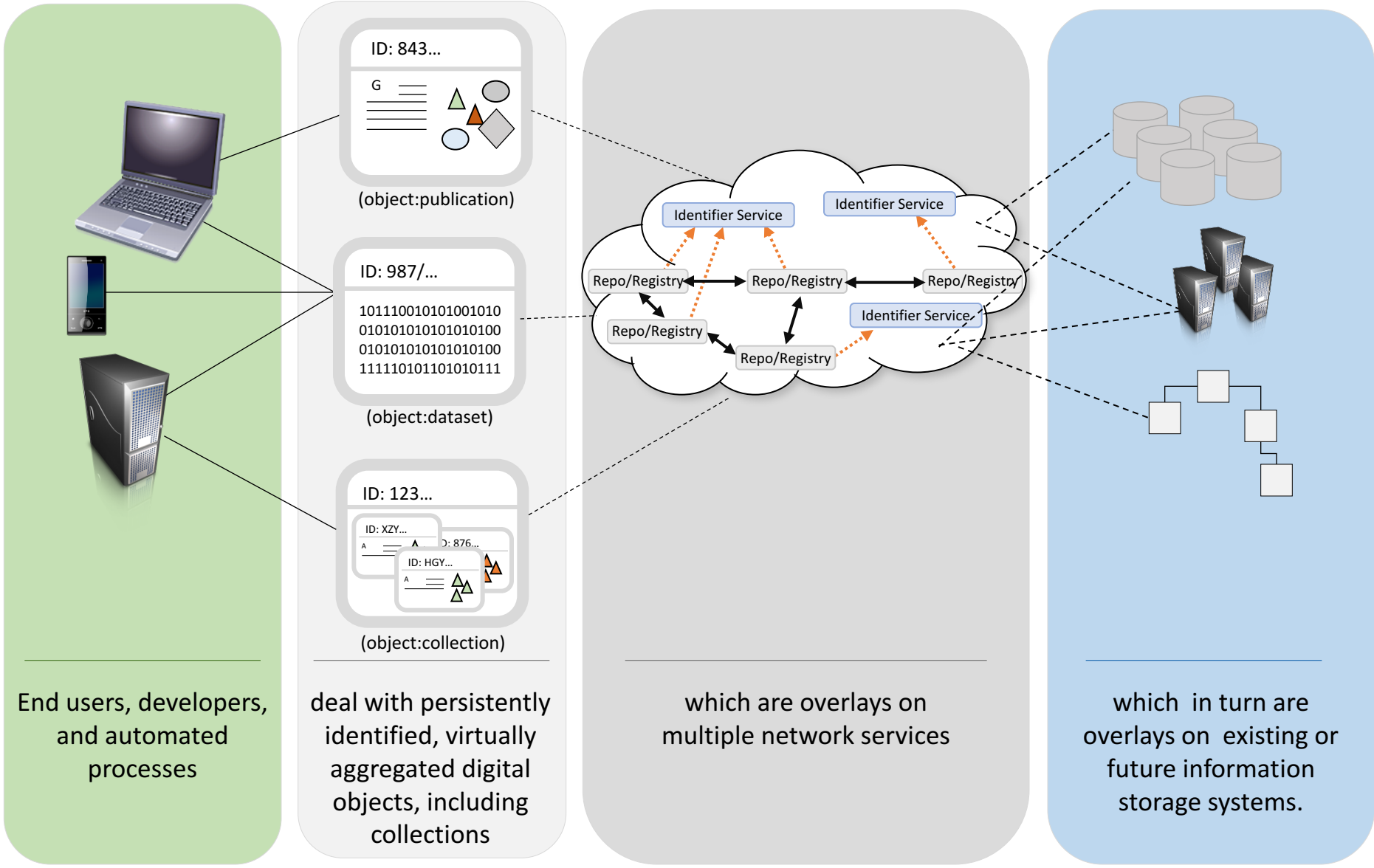
What Problem are We Trying to Solve?

- On the processing side – the vast amounts of data now being collected and generated require new levels of automation.
  - Point solutions are terribly inefficient
  - Core data processing is the same in physics as it is in medicine
  - We need common tools and processes for the levels at which all data is the same
- On the access side – simply making data available (DOI to repository), challenging as that may sometimes be, is not sufficient for widespread sharing and re-use of scientific data.
  - How can users interpret the data? What do they know other than format? Provenance? Documentation?
  - Can they combine it with other data sets? What tools can they use?
  - Where is the detailed metadata?
- We see the need for a common middle layer between clients of all kinds and storage of all kinds, allowing focus on the difficult semantics/knowledge challenges.

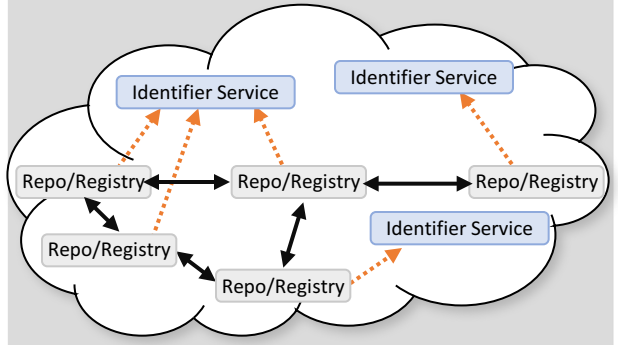
What Exactly are we Proposing to Do?

- Implement a prototype distributed environment based on the digital object model
  - Everything in the environment is a digital object
  - For basic information management tasks every object can be treated the same, regardless of information content
    - Every object has a globally unique and actionable identifier
    - Every object is typed
    - Every object has tightly associated metadata
    - Every object has a query-able set of operations that can be performed on it
- Start with the minimal set of components and services that enable the DO model
  - Identifiers + Resolution System
  - Types + Type Registries
  - DO Repositories, including repositories of metadata, aka, registries
  - Mapping/brokering software & services to map existing data storage and management systems to DOs
  - Digital Object Interface Protocol, implemented by DO Repositories
- Open the environment to as many use cases as possible to hone the core infrastructural pieces

Global Digital Object Cloud (GDOC)

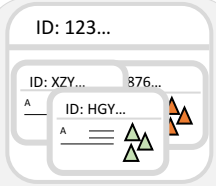


Global Digital Object Cloud (GDOC)

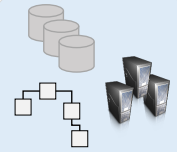


which are overlays on multiple network services

End users, developers, and automated processes



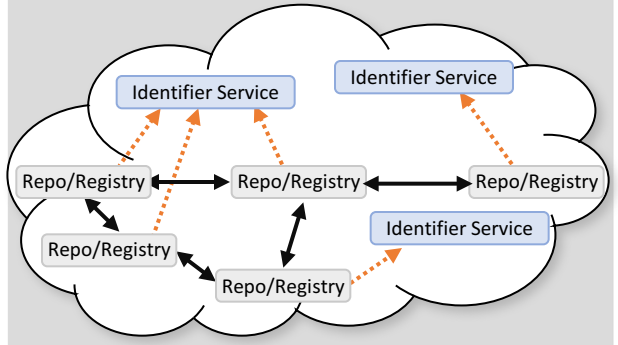
deal with persistently identified, virtually aggregated digital objects, including collections



which in turn are overlays on existing or future information storage systems.

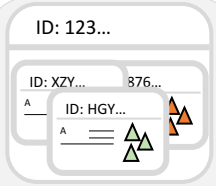
Global Digital Object Cloud (GDOC)

These services can be orchestrated to provide an object view of underlying storage, e.g., file systems, or basic data management systems, e.g., databases.

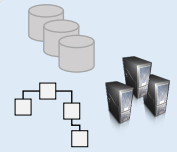


which are overlays on multiple network services

End users, developers, and automated processes

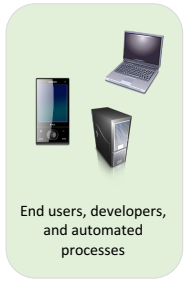


deal with persistently identified, virtually aggregated digital objects, including collections

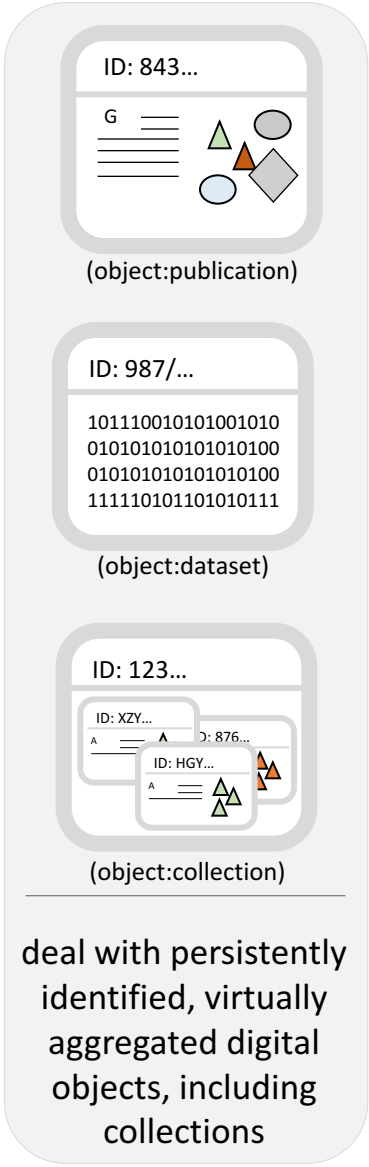


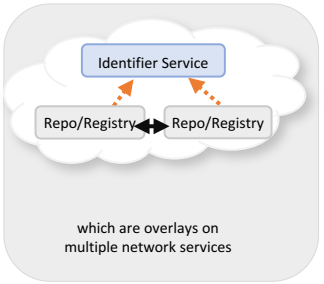
which in turn are overlays on existing or future information storage systems.

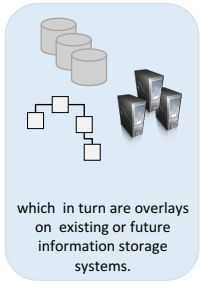
Global Digital Object Cloud (GDOC)



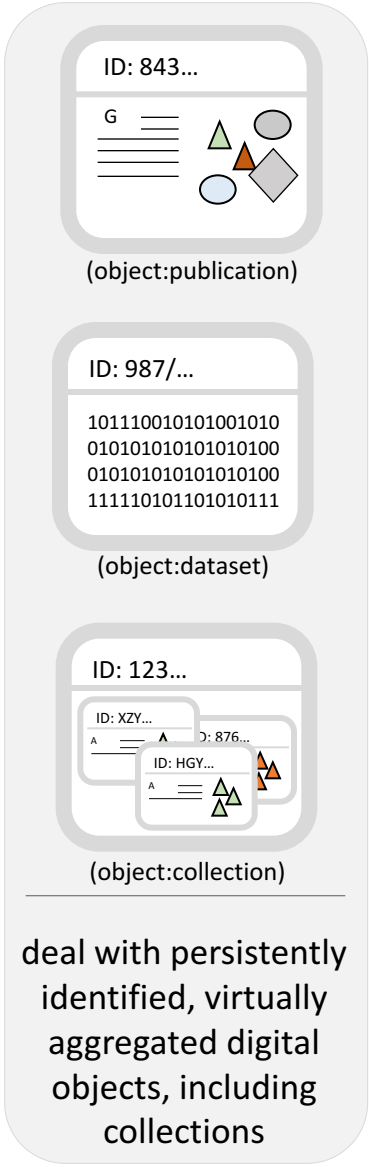
End users, developers, and automated processes



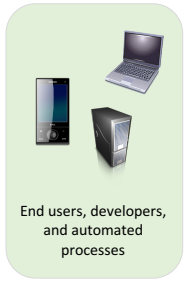


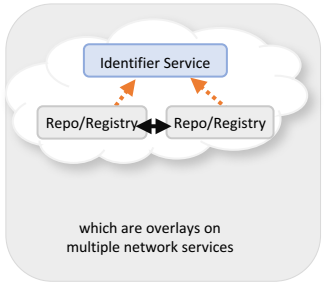


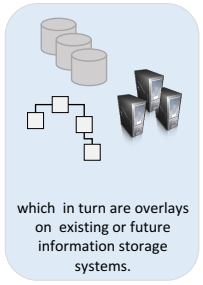
Global Digital Object Cloud (GDOC)



The resulting set of identified and well-structured objects provide a common, and constant, view and ‘remote control’ management of data distributed in various locations and systems, which can change without changing the virtualized object.

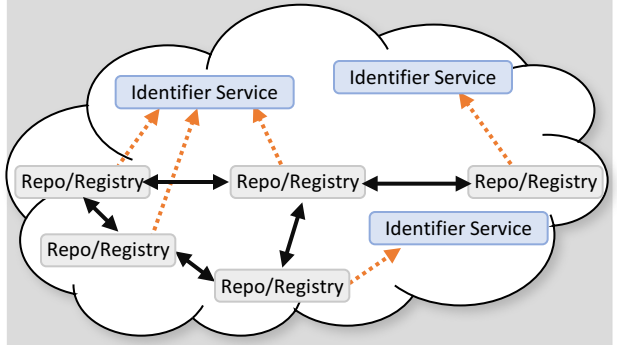






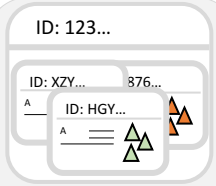
Global Digital Object Cloud (GDOC)

All of these services exist today in one form or another, but some are not yet widely used and few are tightly coordinated and orchestrated in the way that is needed.

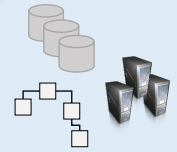


which are overlays on multiple network services

End users, developers, and automated processes



deal with persistently identified, virtually aggregated digital objects, including collections



which in turn are overlays on existing or future information storage systems.

Why is this a Good Idea?

- The Digital Object Model Simplifies the Solution Space
  - Treat every information object the same until you have to differentiate among them to accomplish your purpose
  - Push the current cacophony of information management and storage systems down a level of abstraction
  - Objects are self-describing in that they carry their type information independent of their current system location
- The prototype will let us test the above assertions
- The prototype will be based on open standards and proven technology
- The proposed project has already gathered significant support and is coming out of the Research Data Alliance, broadly representing the international research data community

Who Is We?

- Digital Object Model based on CNRI's Digital Object Architecture
- RDA Data Fabric Interest Group
  - Reusable components, Automated Work Flows, Type-based operations
  - Supporting Output "Recommendations for Implementing a Virtual Layer for Management of the Complete Life Cycle of Scientific Data"
- Brainstorming meeting Nov 16
  - German Climate Center
  - UK Natural History Museum (biodiversity)
  - Swiss National Computing Center
  - CNRI
  - BRDI
  - Others interested in exploring: DCO, NCAR....

PIDs for Data – Why Bother?

- Managing increasing amounts of primary and secondary data on the Net over long periods of time
- Managing increasingly complex data relationships on the Net over long periods of time
- When the attributes of that data such as location(s), responsible parties, and the underlying systems may change dramatically over time
- Science builds on past work and increasingly relies on collaboration within virtual distributed communities
- All of this absolutely requires reliable, long-term persistent references to bind together the distributed data, processes, and parties involved – referential integrity

PID Considerations – Big Picture

- No lack of unique identifiers in the world – that part is easy
  - Unique identification is NOT a technical challenge (U.S. SS# - 1935)
- Strength in numbers – at this point you would need a very good reason to start yet another PID scheme
  - Smaller independent schemes will be more fragile and vulnerable to a small group moving on in any fashion, i.e., less persistent
  - Reliable well-run systems will tend to grow (nobody gets fired for assigning DOIs?)
  - If there is some aspect of a current widely used scheme that doesn't work for your case, talk to that community
- What problem are you trying to solve?
  - Don't start with deciding on a scheme, start with defining the requirements
- Resolution Systems – basic decision point
  - Single authoritative resolution system (≠ single point of failure): DNS, Handle
  - No single authoritative system (but controlled minting): ISBN, SS#

Why Do We Care About Data Types?

- Problem: Implicit Assumptions in Data
- Data sharing requires that data can be parsed, understood, and reused by people and applications other than those that created the data
- How do we do this now?
  - For documents – formats are enough, e.g., PDF, and then the document explains itself to humans
  - This doesn't work well with data – numbers are not self-explanatory
    - What does the number 7 mean in cell B27?
- Data producers may not have explicitly specified certain details in the data: measurement units, coordinate systems, variable names, etc.
- Need a way to precisely characterize those assumptions such that they can be identified by humans and machines that were not closely involved in its creation

What is a Data Type?

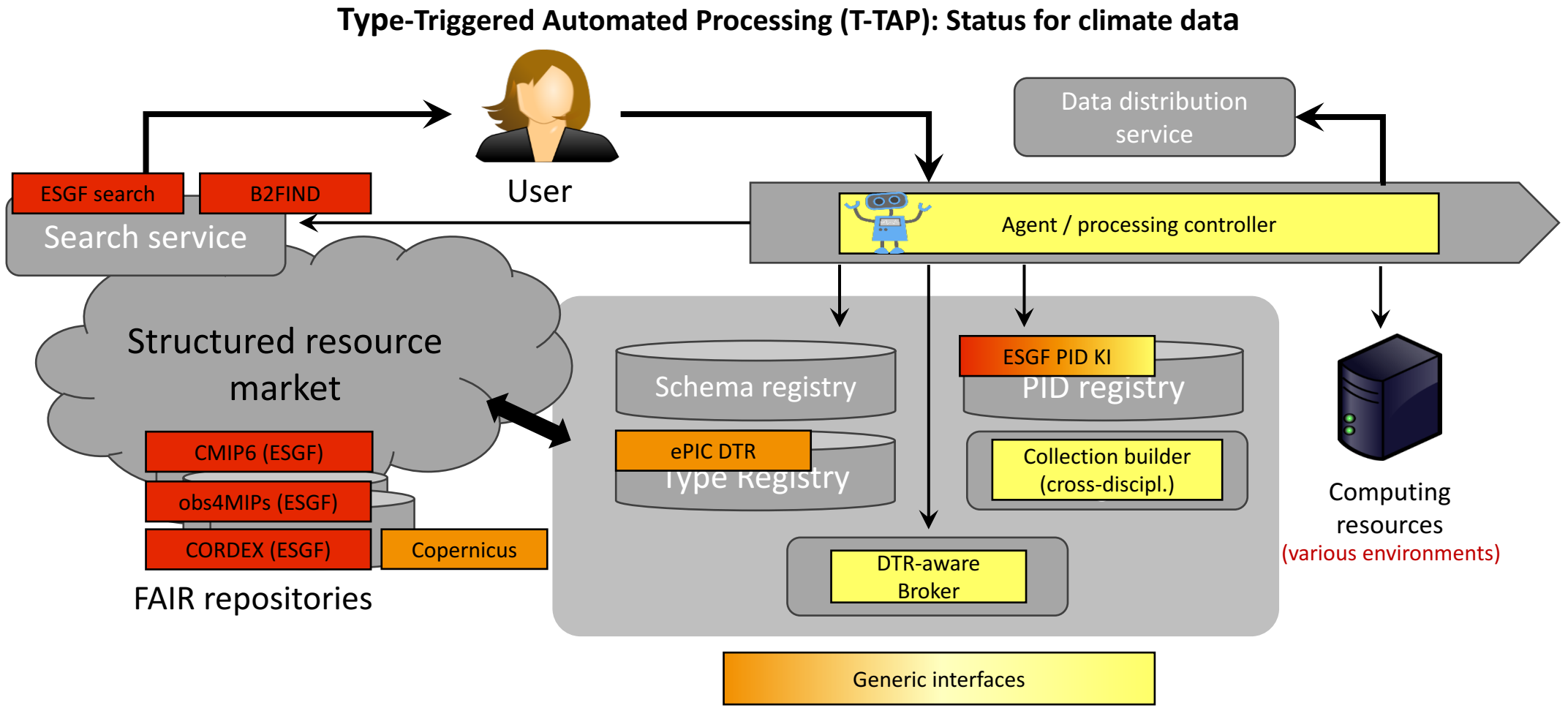
- A unique and resolvable identifier
  - Which resolves to characterization of structures, conventions, semantics, and representations of data
  - Serves as a shortcut for humans and machines to understand and process data
- File formats and mime types have solved the ‘representation’ problem at a ‘unit’ level
- Examples of problems we aim to solve with data types:
  - It is a number in cell A3, but is it temperature? If so, in Celsius?
  - It is a dataset consisting of location, temperature, and time, but what variable names should I look for?
  - Is it all packaged as CSV or NetCDF? And as a single unit or a collection of units?
- Such types are in use and a standard record structure is evolving (ISO study group) – not finished, but functioning

Uses of Data Types

- Discovery
  - Find data of a specific type, e.g., for ’mash-ups’
- Interpretation
  - Parse, understand, and analyze / re-use data created by others
- Processing
  - Where does this data go next in my automated work-flow?
  - Type-triggered Automated Processing (T-TAP)

One Usage of Types

Type Triggered Automated Data Processing (T-TAP): Object type triggers processing, e.g., by an agent in workflow control



red: operational / ready
orange: under construction (e.g. via confirmed projects), but likely to become operational
yellow: more work to be done

Proposed Evaluation of T-TAP for Climate Data Processing (DKRZ)

- Come to a full understanding of the **type-driven workflow**, with necessary component interfaces specified, scope and limitations of types understood and orchestration of type-referenced services demonstrated
- Define more precisely the capabilities of the **agent** and evaluate iterative implementations based on typical user tasks with the available data sources and also along concrete user feedback
- Define the **hand-over steps** among agents and the processing controller (are they different?), based on multiple data center environments
- Explore the concept of the **structured resource market**