C2CAMP
(A Working Title)

International Coordination for Science Data Infrastructure:
A Symposium

1 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 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

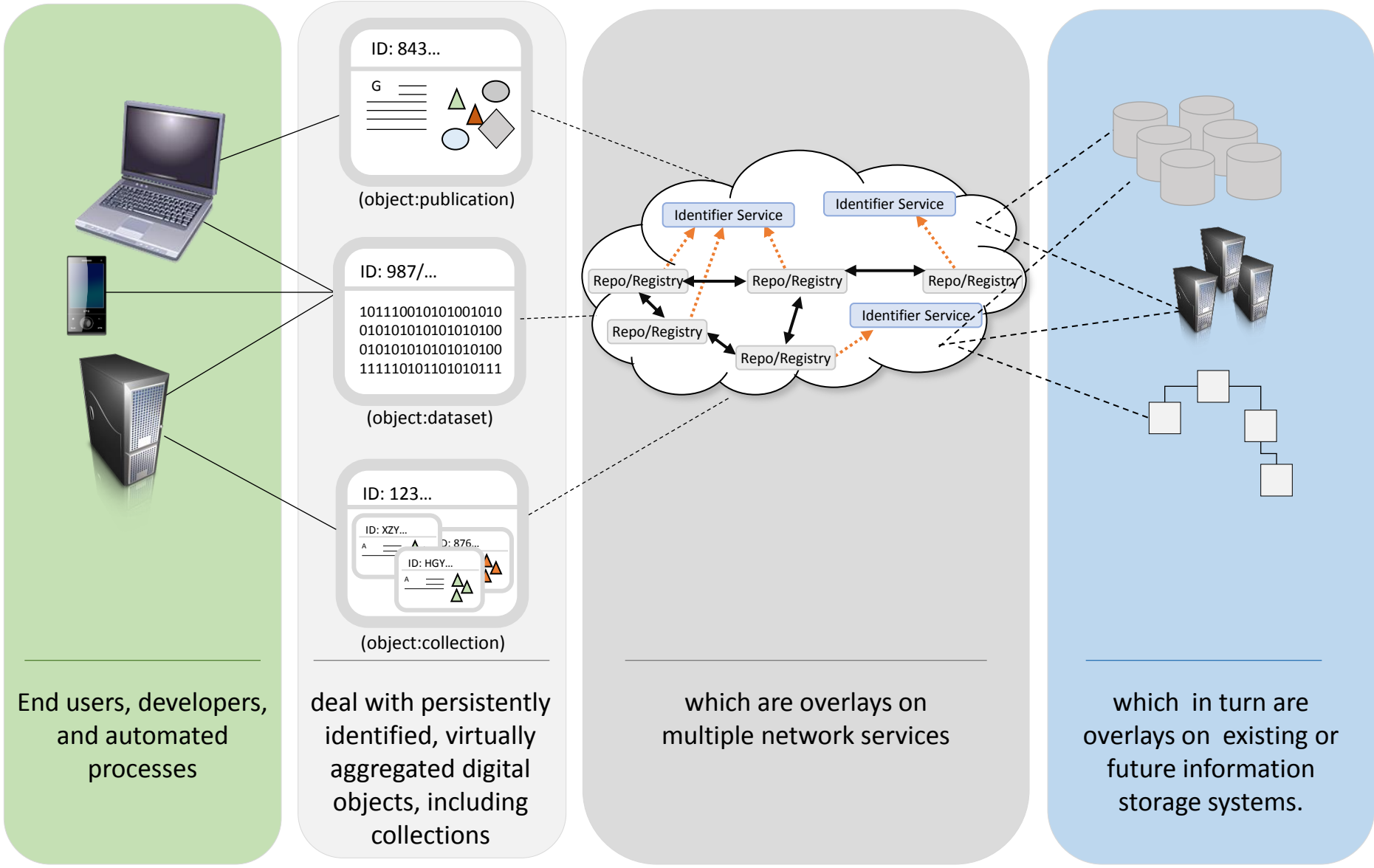
What Problem are We Trying to Solve?

- Wave of Data Coming Quickly in Many Forms (volume/velocity/variety)
  - Scientific output doubling every nine years, as measured by publications
  - And now the data is becoming available
  - Astronomy as example (credit: Richard McMahon, Cambridge Inst. of Astronomy)
    - Petascale data volumes today, exascale in a decade
    - Heterogeneous data; 1000's of different instrumental configurations
    - Poorly documented data models
    - Incorrectly or out of date documented data models
- Availability of data should result in higher levels of re-use, reproducibility, accuracy, but
  - Reproducibility crisis
  - Funding issues
  - More time spent on data than on science
- Need to turn the challenge into an opportunity, change the problem of too much hard to use/find/understand data to the advantage of lots of accessible and understandable data

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 queryable 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)



Why is this a Good Idea?

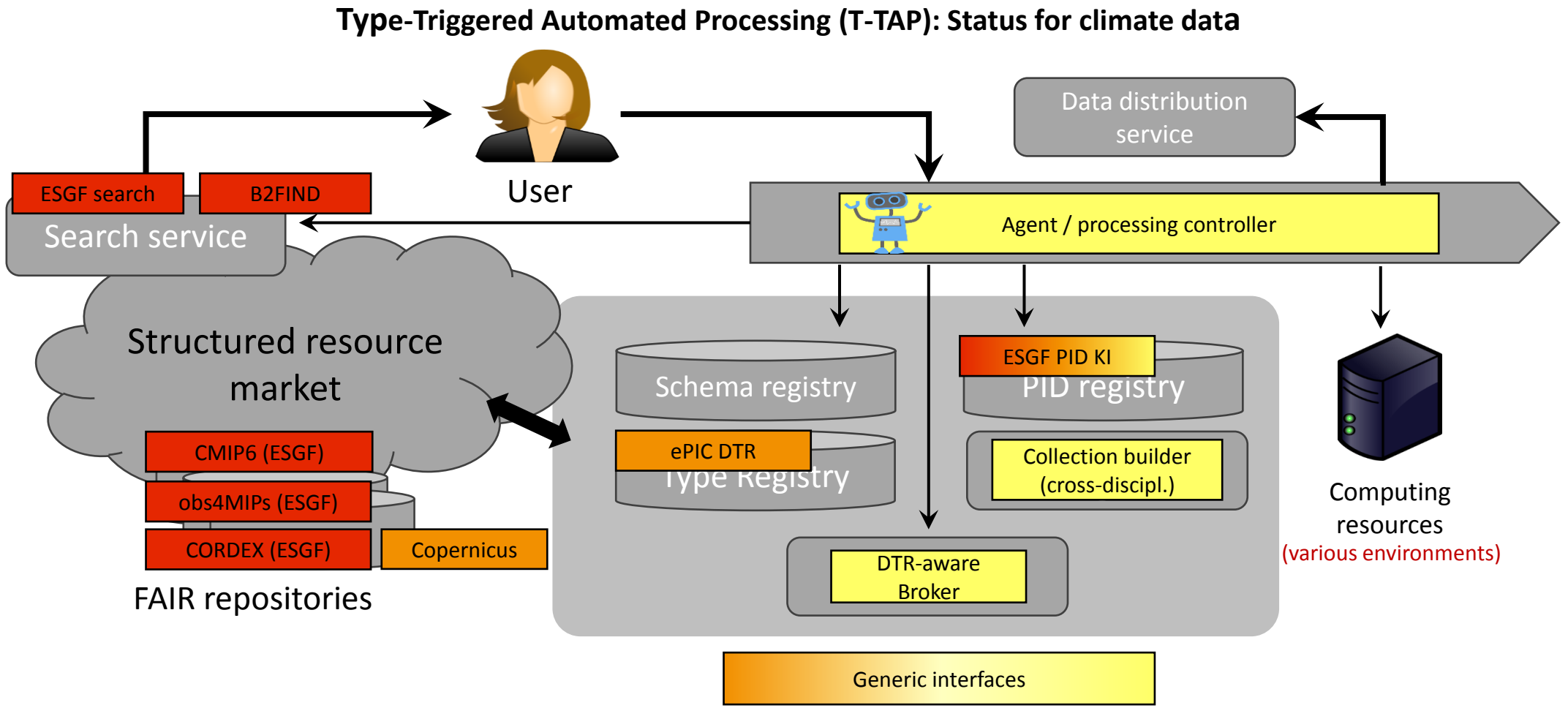
- The Digital Object Model Simplifies the Solution Space
  - Treat every information object the same until you have to differentiate among them to accomplish your purpose
  - Push the current cacophony of information management and storage systems down a level of abstraction
  - Objects are self-describing 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
  - British Museum (biodiversity)
  - Swiss National Computing Center
  - CNRI
  - BRDI
  - Others interested in exploring: DCO, NCAR....

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**