Data Discovery - Introduction

Why (benefits of reusing data)
How EUDAT's services help with this (in general)

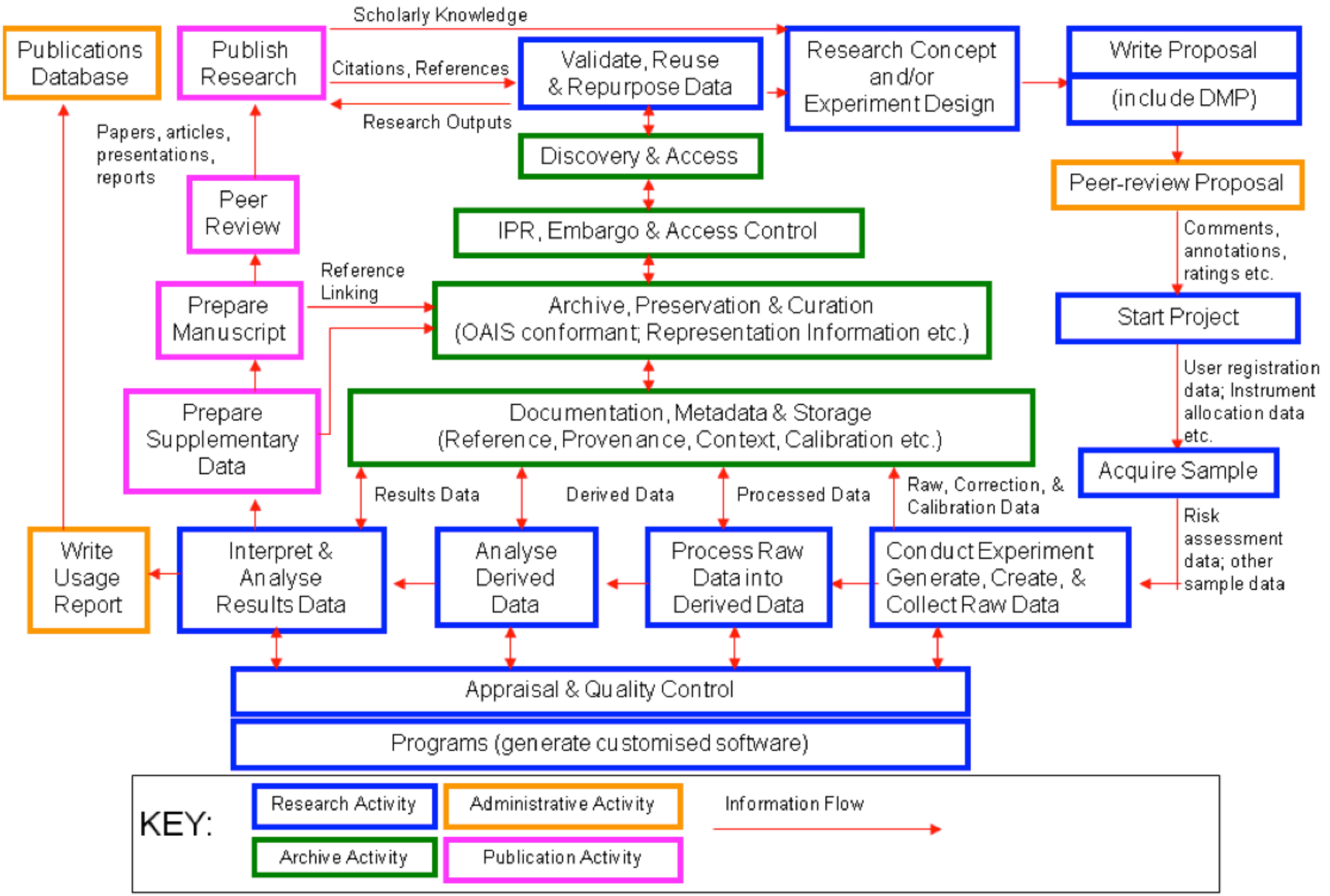
Adam Carter

EUDAT

Getting Your Data

- In days gone by:
  - Design an experiment
  - Conduct the experiment
    - In the lab
    - Real-world observation
    - In silica (Computational Science)
  - Obtain (your own) data
- Now, more and more, there is a large amount of data that has already been collected, and stored

An Idealised Scientific Research Activity Lifecycle Model



Publications Database

Publish Research

Scholarly Knowledge

Citations, References

Research Outputs

Validate, Reuse & Repurpose Data

Research Concept and/or Experiment Design

Write Proposal (include DMP)

Papers, articles, presentations, reports

Discovery & Access

Peer-review Proposal

Peer Review

IPR, Embargo & Access Control

Comments, annotations, ratings etc.

Reference Linking

Prepare Manuscript

Archive, Preservation & Curation (OAIS conformant; Representation Information etc.)

Start Project

User registration data; Instrument allocation data etc.

Prepare Supplementary Data

Documentation, Metadata & Storage (Reference, Provenance, Context, Calibration etc.)

Acquire Sample

Results Data

Derived Data

Processed Data

Raw, Correction, & Calibration Data

Risk assessment data; other sample data

Write Usage Report

Interpret & Analyse Results Data

Analyse Derived Data

Process Raw Data into Derived Data

Conduct Experiment Generate, Create, & Collect Raw Data

Appraisal & Quality Control

Programs (generate customised software)

KEY:

Research Activity

Administrative Activity

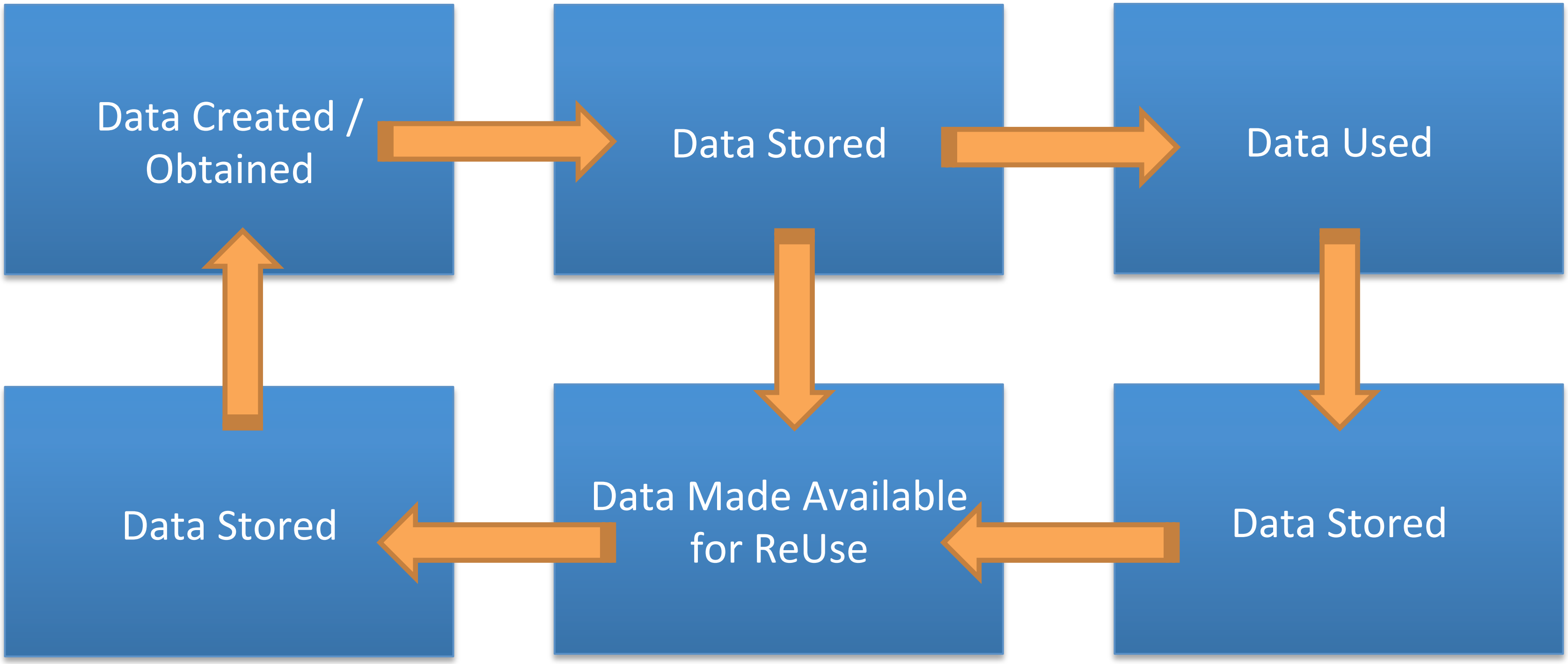
Information Flow

Archive Activity

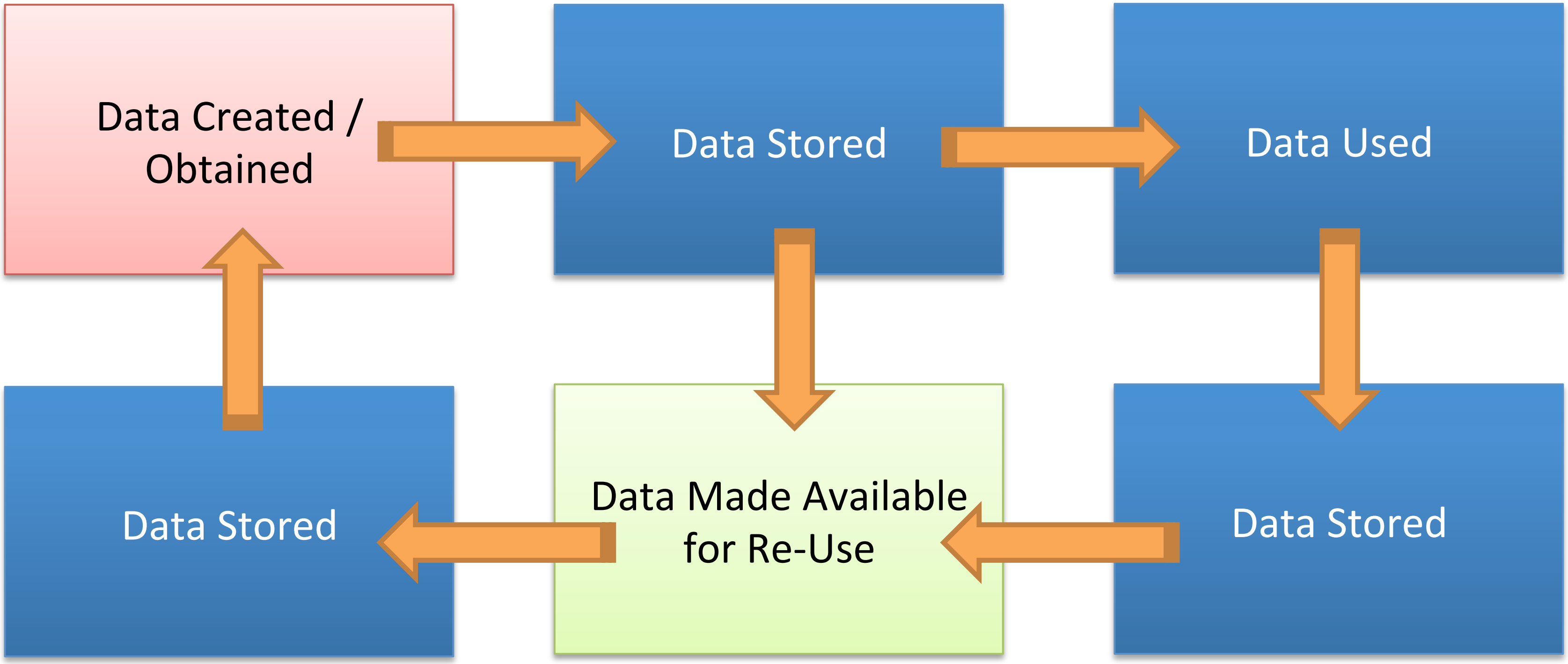
Publication Activity

Source: **I2S2 Idealised Scientific Research Activity Lifecycle Model**
http://www.ukoln.ac.uk/projects/I2S2/documents/I2S2-ResearchActivityLifecycleModel-110407.pdf

A (simplified) Data Lifecycle



A (simplified) Data Lifecycle



Data Discovery & Sharing: Two Sides of the Same Coin

- Why re-use data?
  - Avoids duplication of effort
  - Easier/cheaper than collecting your own
  - It may not be possible to re-measure (e.g. climate data)
  - Validate/test previous results

- Why make your data re-usable?
  - Allow others to build on your efforts and use your data in new ways
  - Allow others to validate/test your results
  - Credit? Reputation?
  - Obligation to funders?

EUDAT

Data Discovery & Sharing: Two Sides of the Same Coin

- How to discover data
  - Web Search
  - Metadata Search
  - Follow links from other data and publications
  - Search popular repositories
  - Ask your twitter followers

- How to make your data discoverable
  - Give it a Persistent Identifier
  - Link it to other data, and cite it
  - Associate it with Metadata
  - Put it somewhere where people can get it easily (e.g. online)
  - Put it in a trusted repository which will look after it beyond when you'd look after it

EUDAT

Allow me to introduce… EUDAT

- A partnership of leading European Data Centres and Research Communities working towards a **Collaborative Data Infrastructure**

AUSTRIA

umweltbundesamt

CZECH REPUBLIC

FRANCE

CINES

CERFACS

maatG

ITALY

CINECA

INGV

SPAIN

Barcelona Supercomputing Center

RedIRIS

NETHERLANDS

SURF SARA

POLAND

PSNC

SWEDEN

SNIC

NORWAY

UNINETT Sigma

FINLAND

CSC

UNITED KINGDOM

Science & Technology Facilities Council

epcc

UCL

Trust-IT Communicating ICT to Markets www.trust-itservices.com

SWITZERLAND

CERN

GERMANY

DKRZ Deutsches Klimarechenzentrum

JÜLICH Forschungszentrum

KIT Karlsruhe Institute of Technology

Eberhard Karls Universität Tübingen

Max-Planck-Institut für Meteorologie

rzg Rechenzentrum Garching

EUDAT

EUDAT: Vision & Architecture

- EUDAT began with the concept of the Collaborative Data Infrastructure
  - See “Riding the Wave” (High Level Expert Group on Scientific Data, Final Report, 2010)
- This identified a handful of core Service Cases
- And the implementation of the Service Cases led to our current distributed Architecture
  - See later

What is the EUDAT CDI?

- The EUDAT Collaborative Data Infrastructure is
  - a ***pan-European***, ***cross-disciplinary*** domain of research data for both ***big community*** researchers and ***"long tail"*** scientists
  - where data are ***registered, preserved***, ***accessible*** and made ***re-usable***

What does this mean?

- ***Pan-European***
  - Fundamentally, a wide-area distributed architecture

- ***Cross-disciplinary***
  - Five core stakeholder communities, many other interested; many sources of conflicting requirements!
  - Including simplified services to encourage the “long tail” to participate
  - All implies a significant systems integration challenge!

What does this mean? (2)

- ***Registered*** means EUDAT data are
  - Globally identified and discoverable (the PID Service)

- ***Preserved*** means EUDAT data are
  - Stored at big European HPC and data centres
  - Replicated for safety (**B2SAFE**: the Safe Replication Service)
  - Governed by policy rules (the Policy Management Service)

What does this mean? (3)

- ***Accessible*** means EUDAT data are
  - Identifiable and findable (the PID Service)
  - Retrievable efficiently (**B2STAGE**: the Data Staging Service)
  - Governed by suitable access control (the AAI Service)

- ***Re-usable*** means EUDAT data are
  - Findable (the PID Service)
  - Comprehensible (**B2FIND**: the Joint Metadata Service)
  - Composable and combinable (future workflow and computational services)

What does this mean? (4)

- For both ***big communities*** and ***"long tail"*** means
  - Stable, web-service APIs for existing tool-stacks to use (the Common Service Layer Interface)
  - Low barriers to use (the Simple Store Service)

- Hence the core EUDAT service cases

- Identifying solutions for these cases *that work with our stakeholder communities' existing solutions* led us to the current CDI architecture

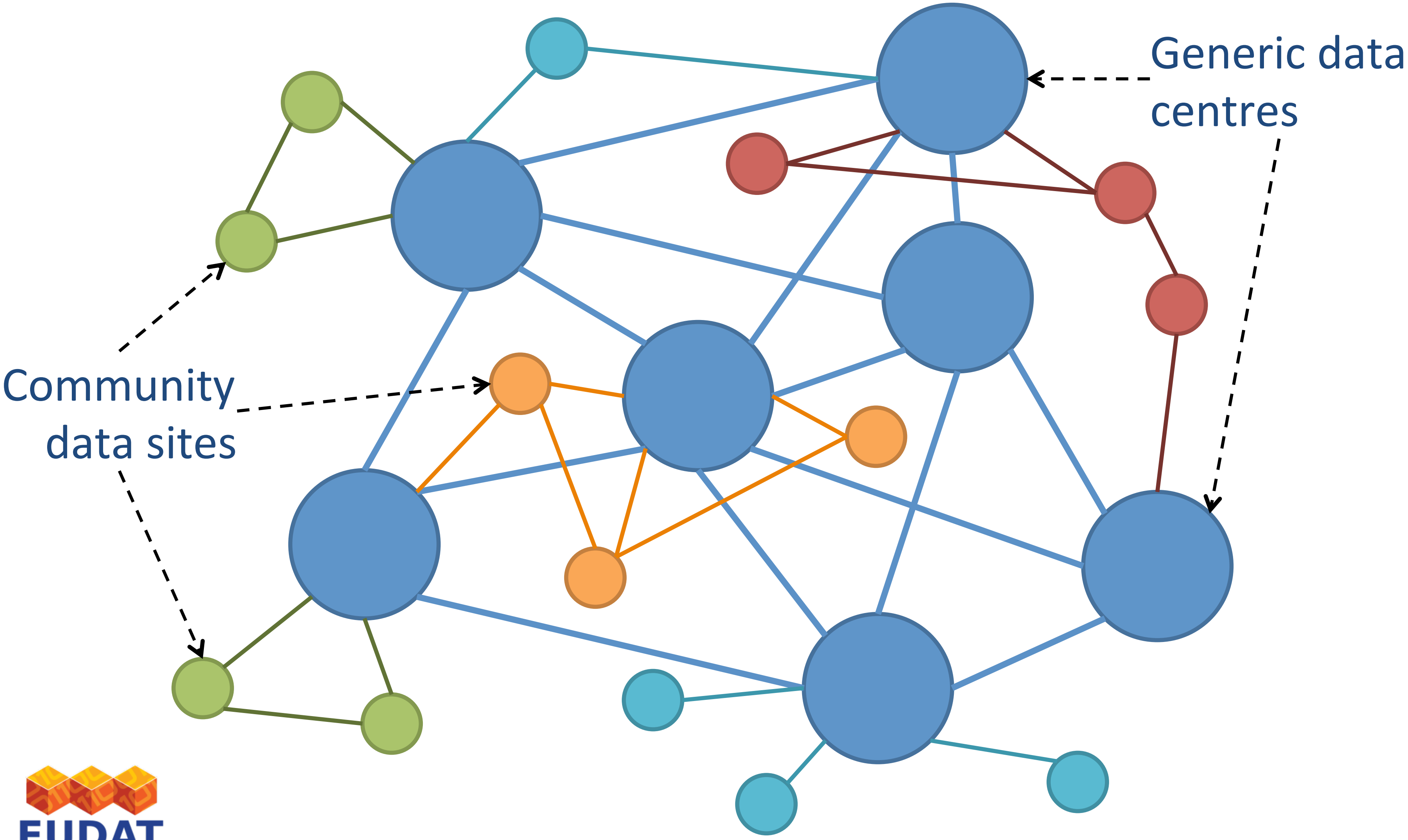
EUDAT

The CDI network architecture

- The CDI is a connected network of European research institutions and data centres (collectively *Nodes*) each offering one or more common EUDAT data services to both participating research communities and independent researchers

- Data centre Nodes have lots of connections
- Research community Nodes need only one
- Connections have both technical & policy agreement aspects

The CDI network architecture



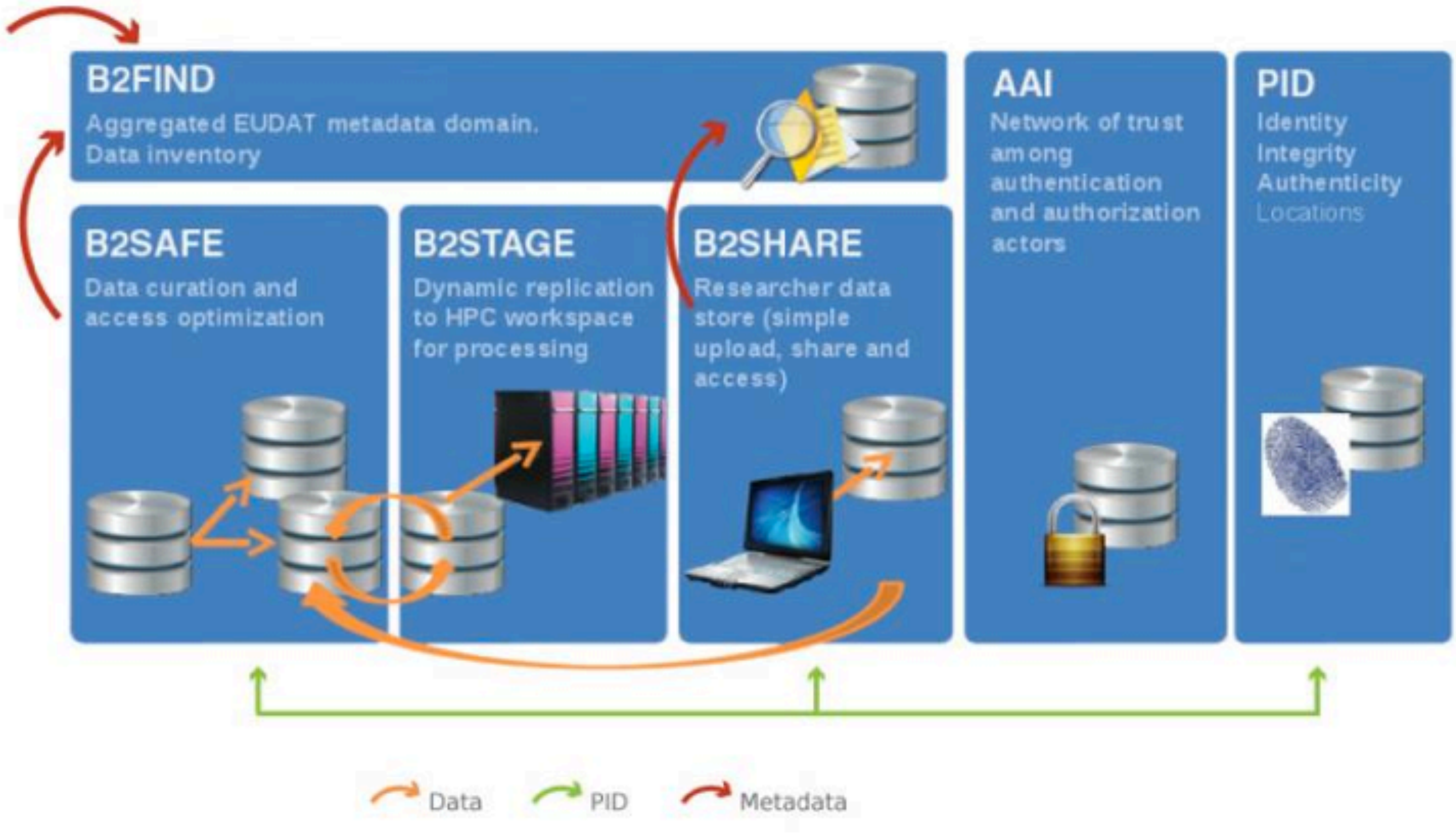
Generic data centres

Community data sites

EUDAT

CDI Node architecture

- Nodes run parts of the CDI Node software suite, depending on which services they want to offer
- All Nodes should offer Safe Replication and PID
  - This is really what being in the CDI is all about
- Others are optional
  - Depends on what a Node’s expected user base requires
- (Some data centre Nodes also need to run the Operational Services suite)



B2FIND
Aggregated EUDAT metadata domain.
Data inventory

B2SAFE
Data curation and access optimization

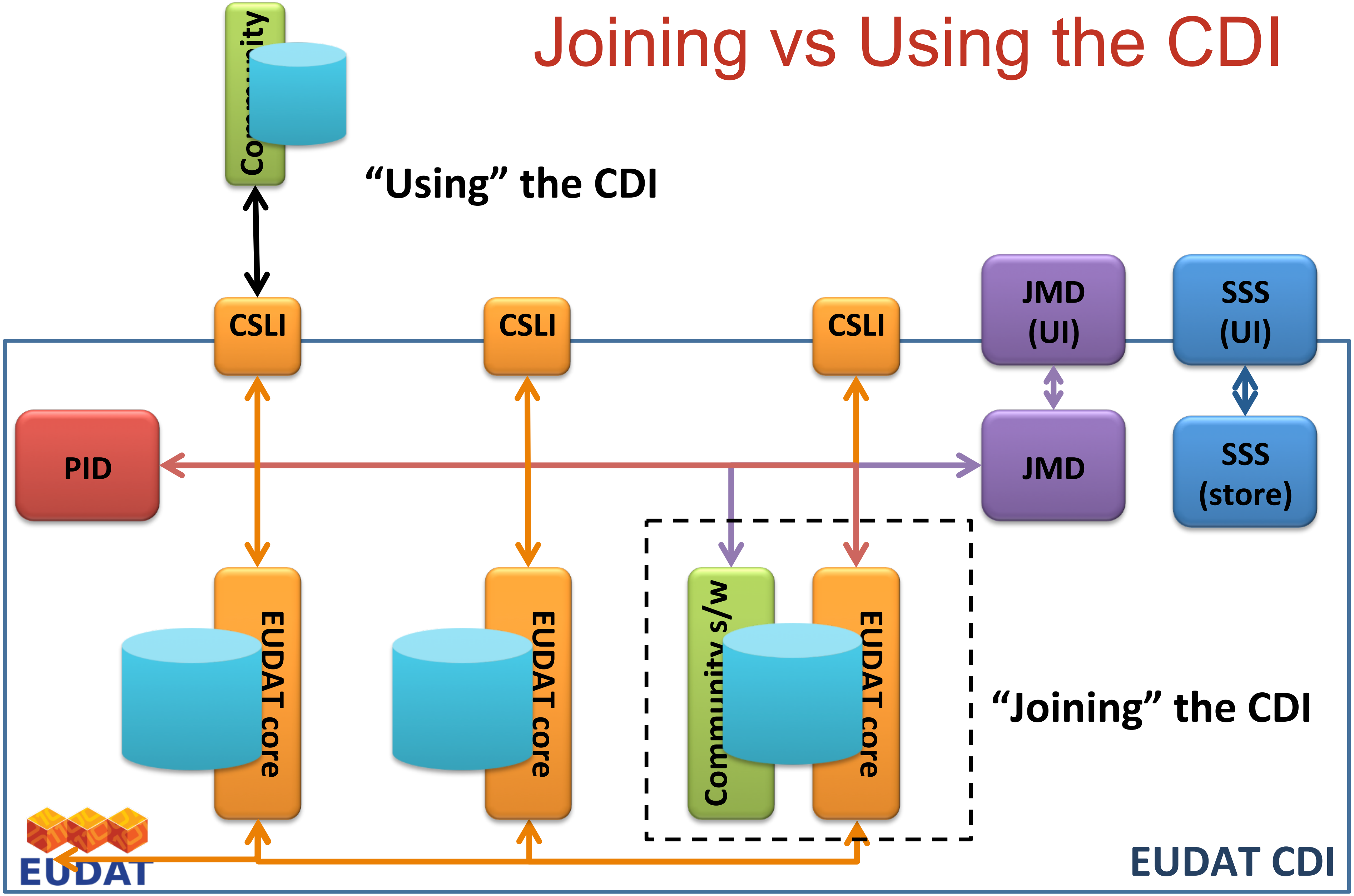
B2STAGE
Dynamic replication to HPC workspace for processing

B2SHARE
Researcher data store (simple upload, share and access)

AAI
Network of trust among authentication and authorization actors

PID
Identity
Integrity
Authenticity
Locations

Joining vs Using the CDI



What’s Next for EUDAT?

Working Groups

- Data Access & Re-Use Policies
- Dynamic Data
- Semantics
- Workflows

New Services? *Under Consideration*

- B2NOTE? – Annotation
- B2DROP? – File Workspace & Synchronisation
- B2HOST? – Hosting of (Application Specific) Data Services