Opening up Climate Research:
a linked data approach to publishing data provenance

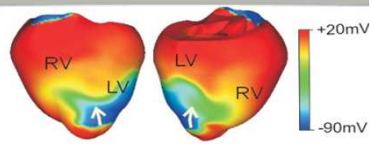
Brian Matthews,
STFC e-Science

Arif Shaon (STFC ESC), Sarah Callaghan (STFC - CEDA), Bryan Lawrence (STFC - CEDA), Andrew Woolf (B. of Met, Aus),Tim Osborn (UEA - CRU), and Colin Harpham (UEA - CRU)

Science & Technology
Facilities Council

Science & Technology Facilities Council

Science and Technology Facilities Council



- Provides large-scale scientific facilities for UK Science

- RAL is the home of two key UK environmental data centres,
  - BADC and NEODC
- E-Science Centre
  - Involved in Digital Curation activities
  - Data management
  - Active contributor in the international arena of Environmental Informatics, e.g. OGC(OWS 6), INSPIRE and ESA

Science & Technology Facilities Council

ACRID

- **Advanced Climate Research Infrastructure for Data**
  - Developed a linked-data approach to publishing complex climate research datasets
  - Collaboration between:
    - Climatic Research Unit, University of East Anglia
    - STFC e-Science Centre, Rutherford Appleton Laboratory
    - Met Office (unfunded partner)
  - Funded by the JISC Managing Research Data (MRD) Programme
  - Completed on 1 August 2011

Data Publication

- Publishing scientific datasets as scientific resources
  - validation and reproducibility
  - credit

- Data by itself is not sufficient
  - verification of data provenance
  - detailed workflow information
  - Including analysis software

Science & Technology Facilities Council

House of Commons Report

"(data sharing) actions were in line with common practice in the climate science community"

"it is not standard practice in climate science to publish the raw data and the computer code in academic papers".

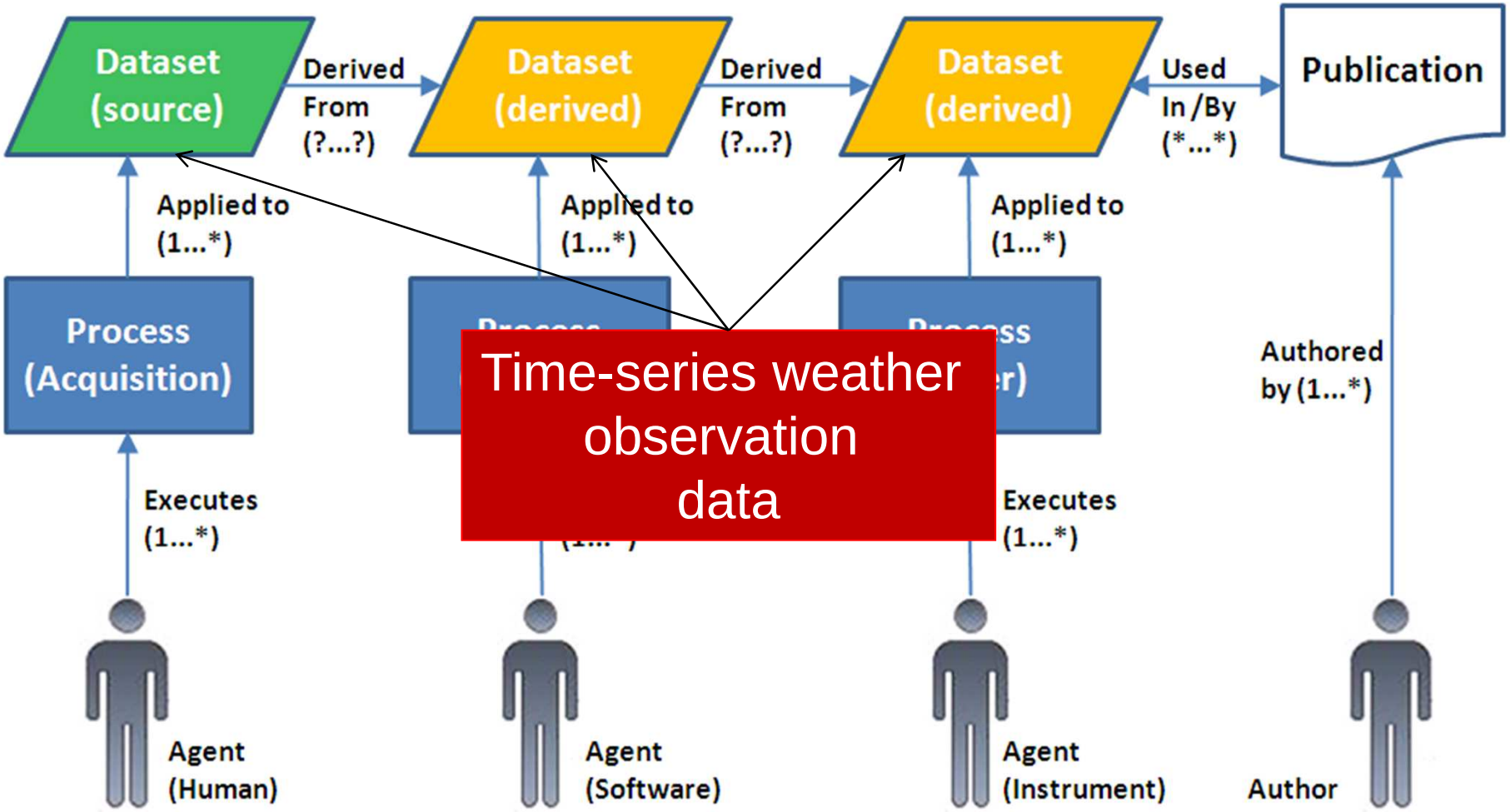
"...that climate scientists should take steps to make available all the data that support their work (including raw data) and full methodological workings (including the computer codes)".

Main Challenges for ACRID

- Repeatability and Reusability
  - Facilitating traceability of the provenance of published data
  - Enabling re-enactment and reusability of the workflows
- Common information model
  - Interoperability with existing systems and tools
  - Address dynamic/evolving datasets, e.g. versioning
- Citable in scholarly communication
- Efficient metadata curation strategy
  - Accurate collection of metadata
  - Efficient management and storage
  - Querying and Searching
  - Keep metadata in sync with data

Science & Technology Facilities Council

ACRID Information Model (Workflow Analysis)



ACRID Information Model (Main Concepts)

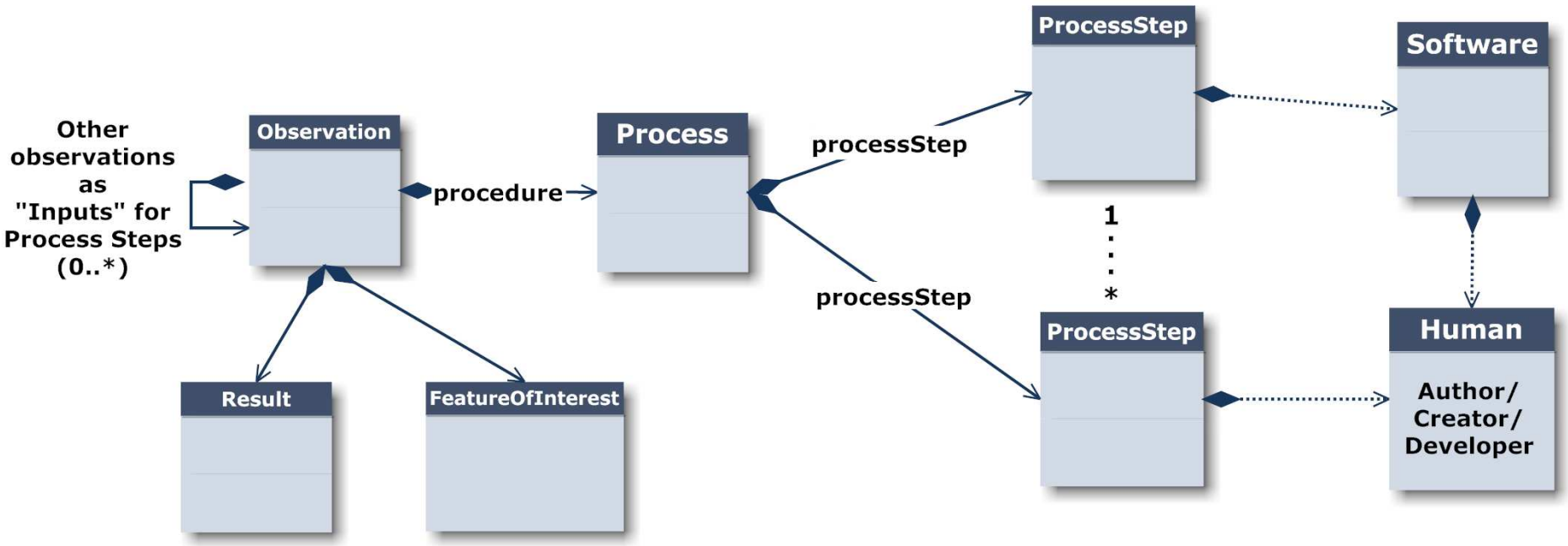
- Observation
  - *The act of measuring or calculating a particular property (e.g. temperature) associated with a certain feature of interest (e.g. air) over a discrete period of time*
- Process
  - *an action or a set of actions performed to produce the result (i.e. dataset) of an observation*
  - *e.g. an algorithm, a computation, a manual procedure*
- Processor
  - *an entity or a set of entities that performs and/or controls a process in order to produce the result of an observation*
  - *e.g. a human, computer software, an instrument etc.*

ACRID Information Model

- Need to be compatible with existing models in use in Environmental science
- Developed as:
  - an application schema (profile) of the ISO 19156 Observations & Measurements (O&M) Model
  - with the observation related concepts derived from the Climate Science Modelling Language (CSML) *TimeSeriesObservation* classes
- Available in UML, GML schema and RDF Ontology

Science & Technology Facilities Council

ACRID Information Model – Overview



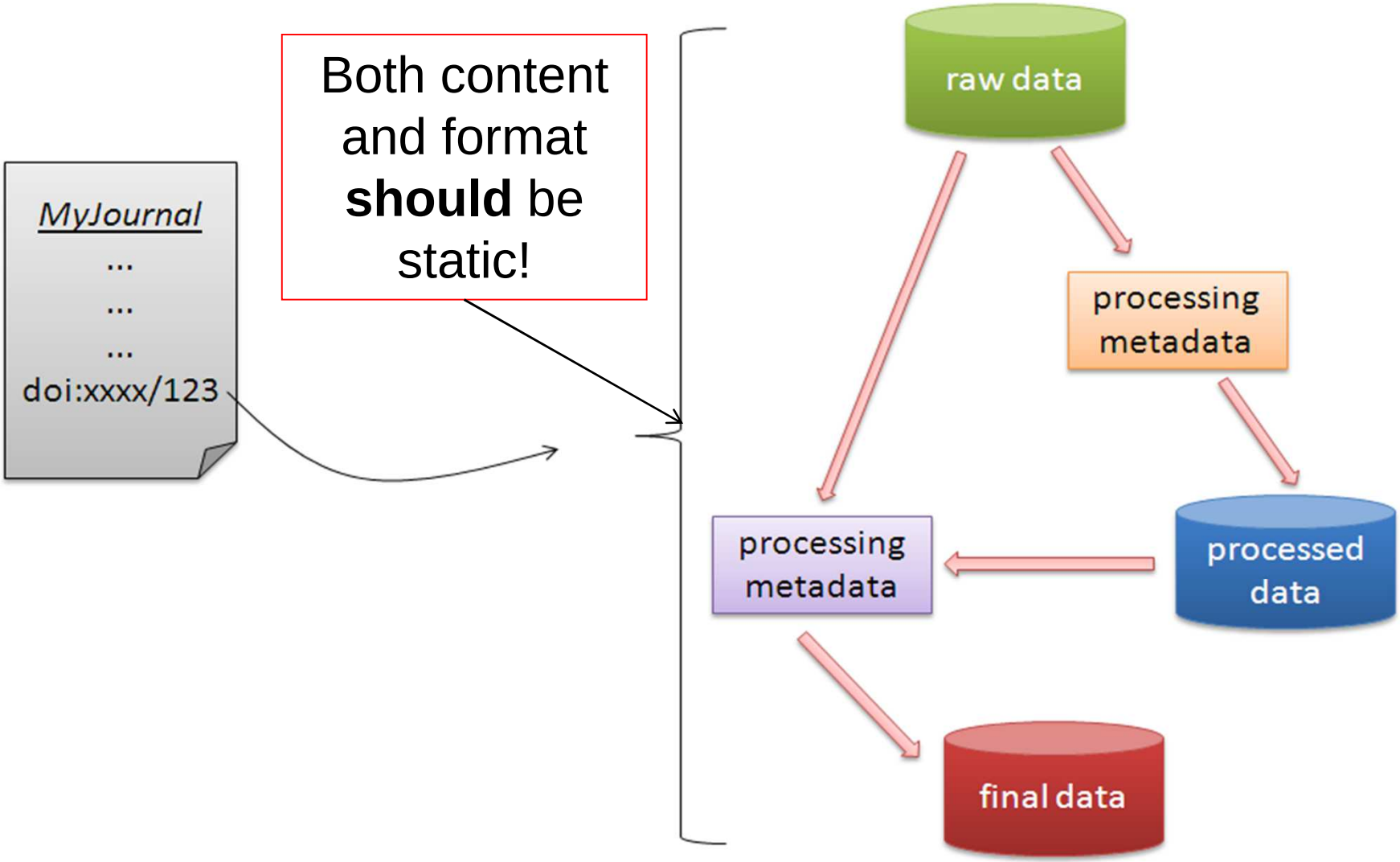
Science & Technology Facilities Council

Data Citation (Identification)

- Digital Object Identifier (DOI) for dataset identification
- *DataCite* initiative
  - International consortium, incl. British Library, assigning DOIs to datasets
  - Earth System Science Data journal issues DOIs to data publications
- The main DOI "question" – *What does a DOI point to?*
  - We propose "linked-data".

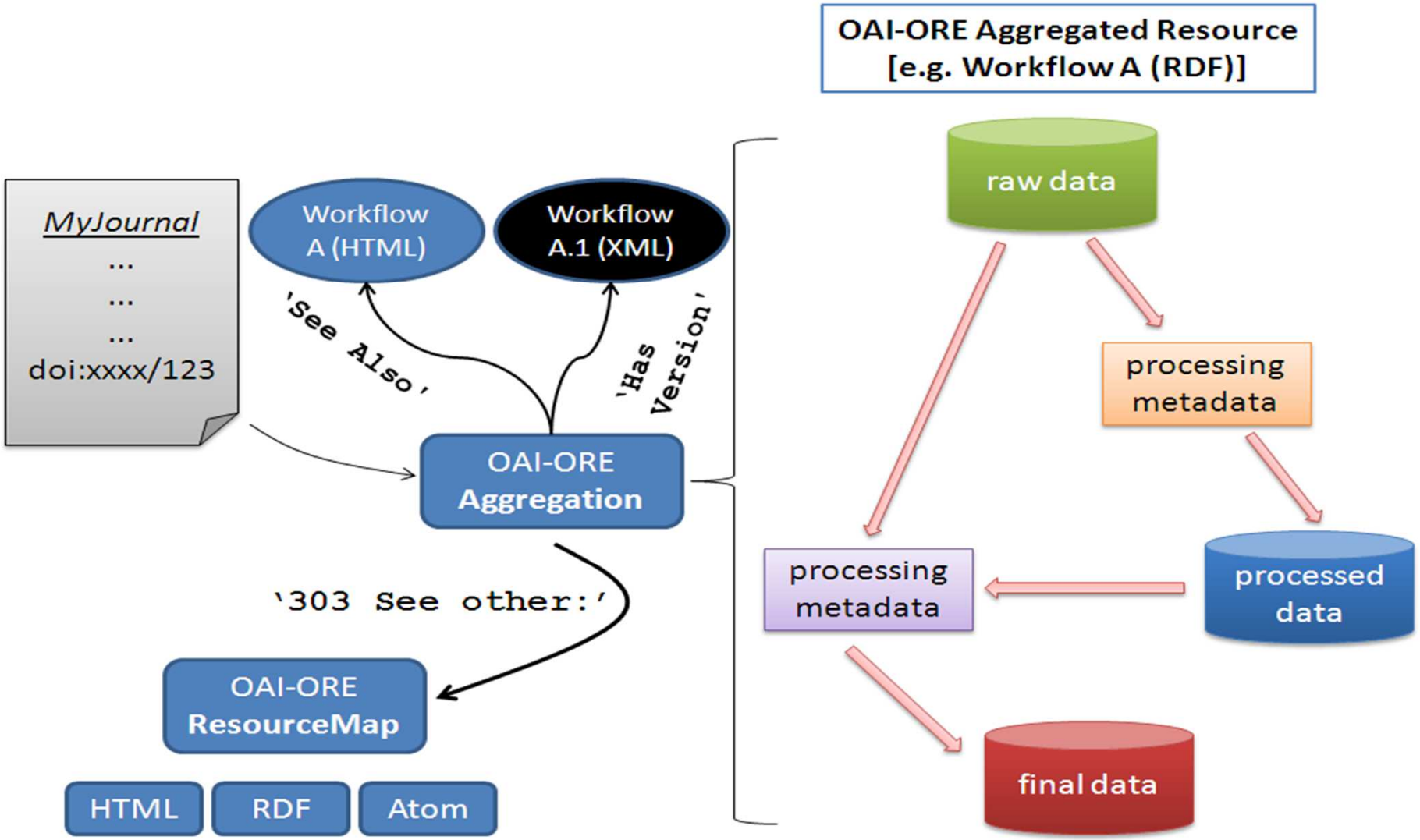
Science & Technology Facilities Council

A Published Linked Workflow



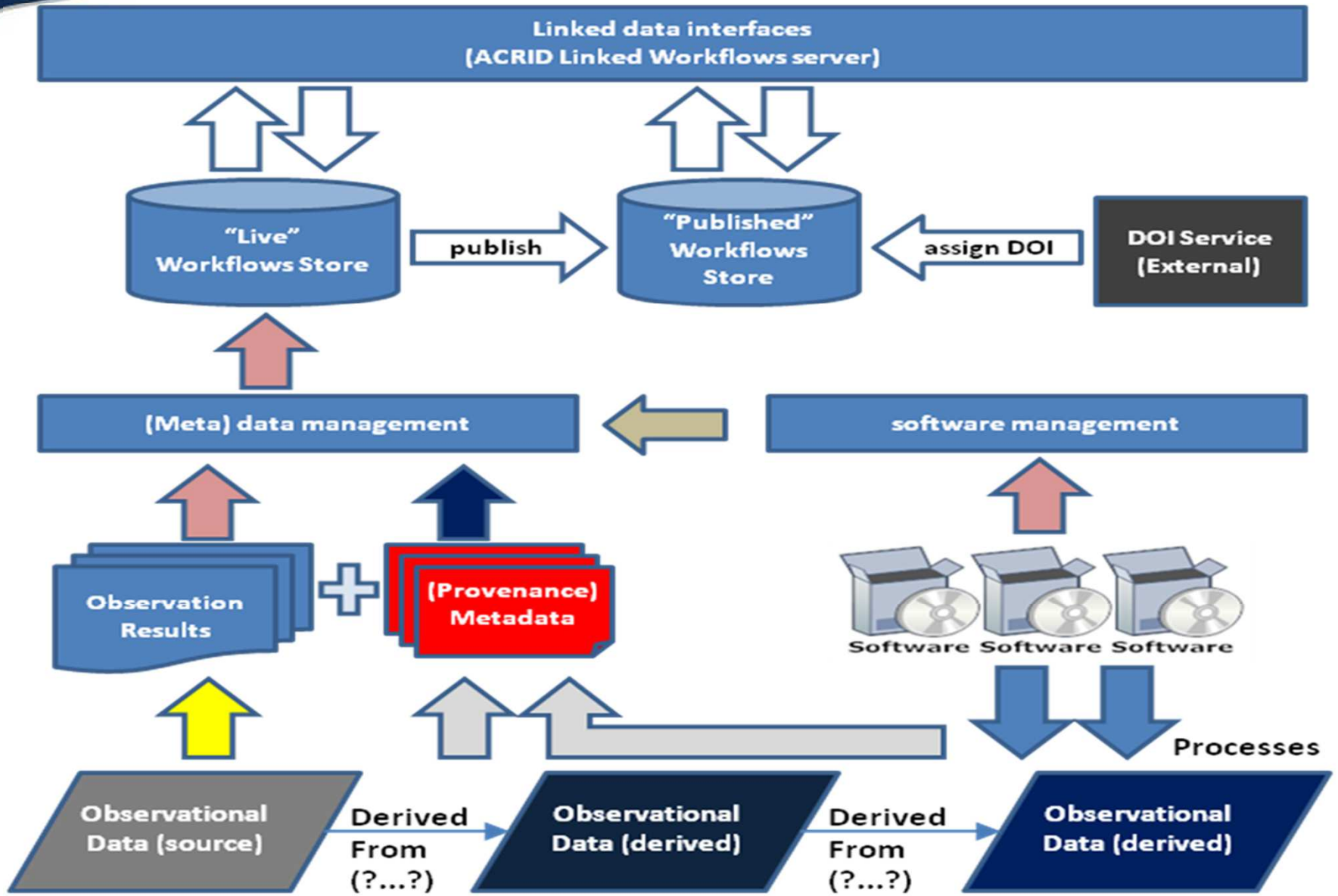
Science & Technology Facilities Council

A Linked Workflow in OAI-ORE



Science & Technology Facilities Council

Infrastructure



Science & Technology Facilities Council

Conclusions

- ***opening up climate research datasets***
  - facilitates greater transparency and traceability of the data life-cycle;
  - enables data accessibility and sharing
  - uses ISO standards and linked-data principles
  - potential to be applicable in wider areas of science
- **But its only the start**
  - the *real* benefit of ACRID may not be realised without *real* consumers
  - Linked-data vs Geospatial data formats needs to be addressed
  - Official Ontology representations of the ISO 19000 models
  - the application of DOIs for publishing linked scientific workflows should be explored further

Questions?

arif.shaon@stfc.ac.uk

ACRID Website:
http://www.cru.uea.ac.uk/cru/projects/acrid/

ACRID Linked Workflows Server:
http://westerly.badc.rl.ac.uk:8080/alws/index.html

Data Citation (Dissemination)

- Open Archives Initiative -Object Reuse and Exchange (OAI-ORE)
- OAI-ORE *defines standards for the description and exchange of aggregations of Web resources*.
- Leverages the RDF and Linked Data concepts.
- Consists of the following notions:
  - **Aggregation (A)**: a set of Web-based Resources.
  - **Aggregated Resource (AR)**: a Resource that is a constituent of an Aggregation.
  - **Resource Map(ReM):** describes an Aggregation.