Advanced Climate Research Infrastructure for Data (ACRID)

Arif Shaon (STFC), Tim Osborn (UEA), Sarah Callaghan (STFC), Colin Harpham (UEA) and Bryan Lawrence (STFC)

UEA
University of East Anglia

CRU

Science & Technology
Facilities Council

Project Background

- The UEA CRU is recognised as a world-leading centre for the analysis and provision of climate datasets.
- Recommendations arising from inquiries into the 2009 hacking of emails from CRU are:
  - publish the scientific workflows associated with the CRU climate research datasets
  - the published workflows should include, wherever possible, both raw and processed data, with processing methods and codes
  - ACRID aims to implement these recommendations

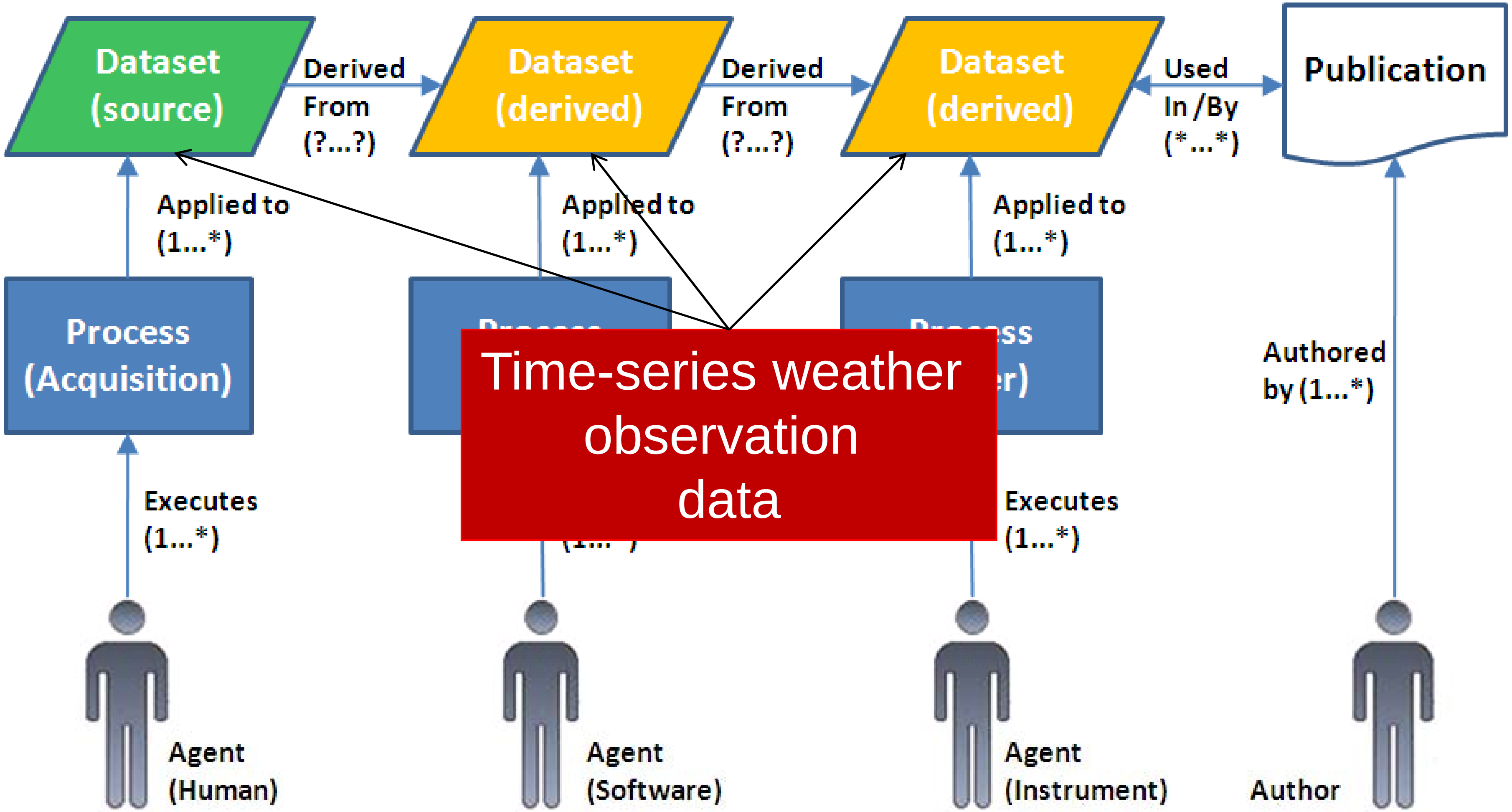
Project Deliverables

- **Information architecture**:
  - develop an information model to describe some of the scientific data workflows in climate research (completed)
  - deploy infrastructure to capture relevant metadata for climate research data, software, and workflows
- **Data citation**: develop a ‘linked-data’ approach to publishing and citing climate research data
- **Prototype** our approach using four high-profile climate research datasets: CRUTEM, CRU TS, tree-ring chronologies, and HadCET

CRU Information Architecture (Information Model)

- ➢ The information model is key!
  - – need a model that captures key aspects of data workflow
  - – enables re-enactment of the workflows
  - – facilitates traceability of the provenance of published data
- ➢ Issues
  - – Information model itself!
  - – Dynamic/evolving data
  - – Data subsets and versioning

CRU Information Model (Workflow Analysis)



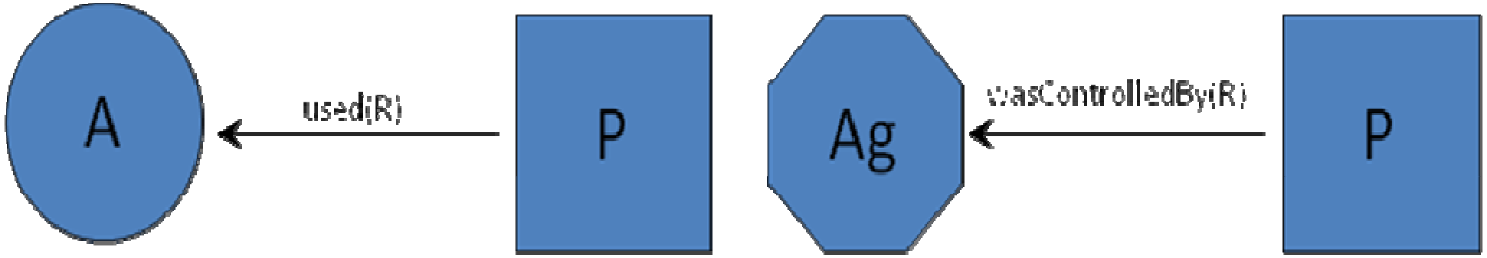
Time-series weather observation data

CRU Information Model (Existing Models)

- The Open Provenance Model
- ISO 19156 Observations and Measurements (O&M) Model
- Climate Science Modelling Language (CSML)

The Open Provenance Model (1)

- A widely-adopted generic model that
  - enables digital representation of the provenance information about any digital or physical object.
  - enables exchange of provenance information between computer systems.
- OPM consists of three different "Nodes" (notions): Artefact (A), Process (P) and Agent(Ag).

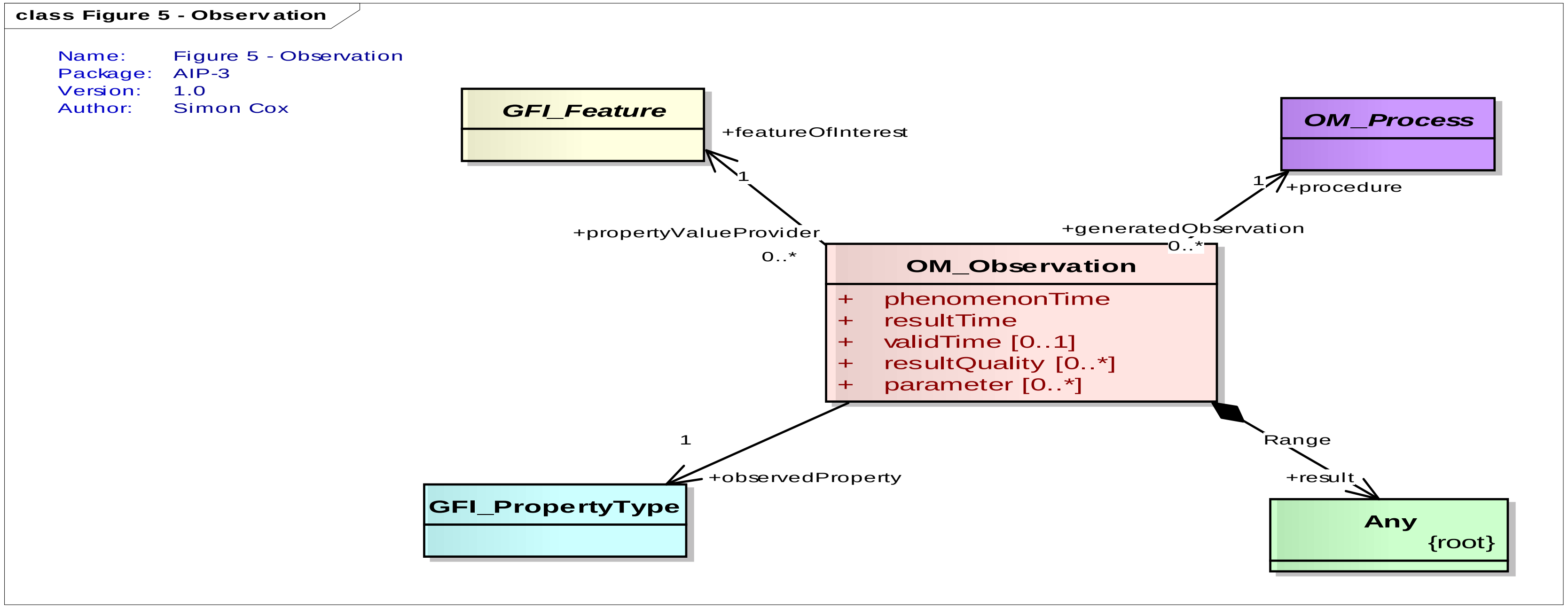


The Open Provenance Model (2)

- A close parallel between the CRU and OPM concepts
- BUT OPM is too generic; significant specialisations would be needed to describe CRU workflows
- Not widely used within the Geospatial community; may not be interoperable with existing tools and systems.

ISO 19156 O&M Model (1)

- Defines a conceptual schema for
  - describing environmental observations
  - the features involved in the sampling associated with such observations.



ISO 19156 O&M Model (2)

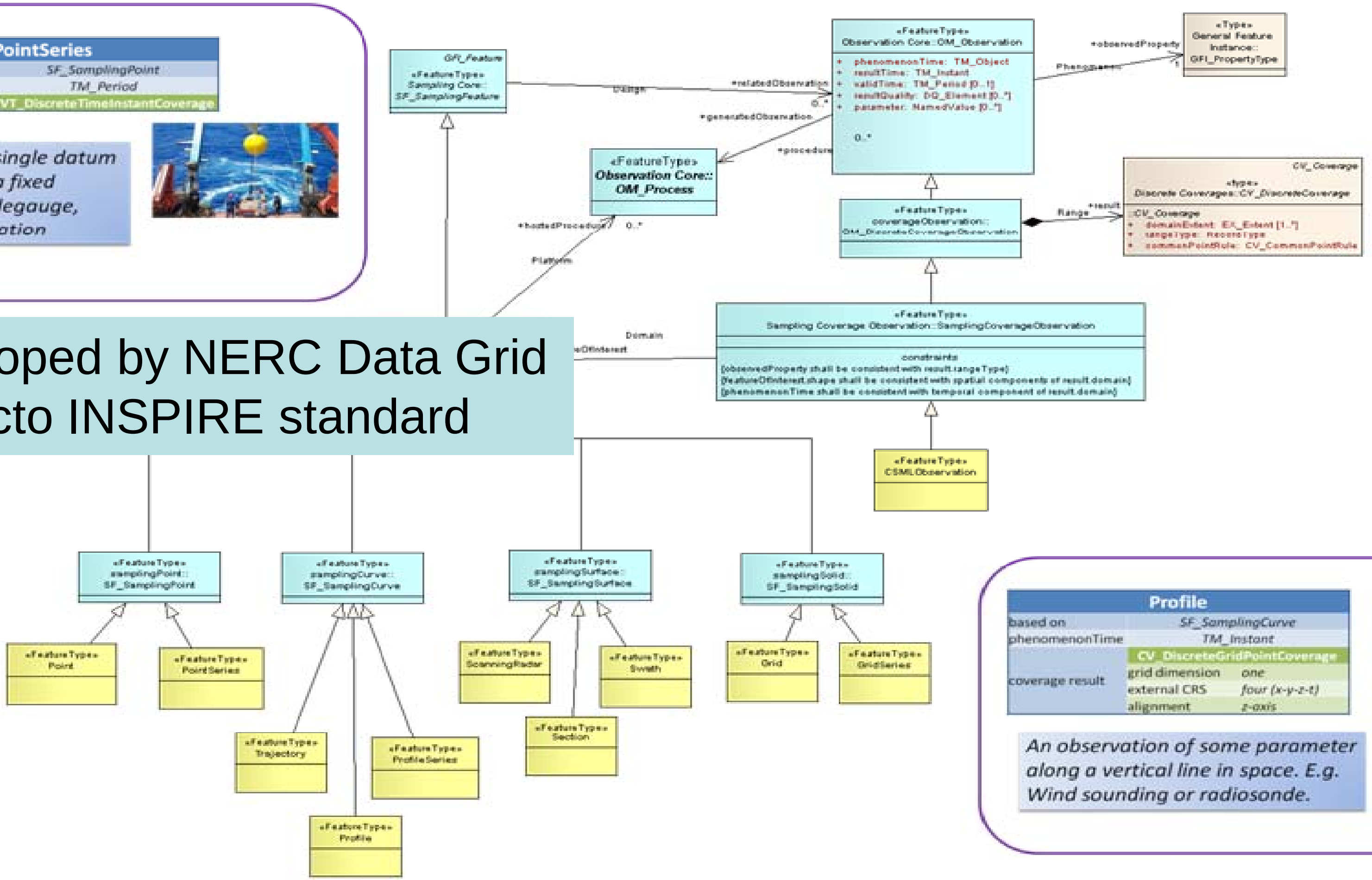
- ISO O&M Model is specifically designed for describing environmental observations, such as the ones represented by the CRU datasets.
- However, O&M too is an abstract model; specialisation of the main O&M classes, such as *OM_Observation* and *OM_Process* would be needed to capture the distinct characteristics (e.g. gridded time-series data, links to publications etc.) of the CRU observations.

Science & Technology
Facilities Council

Climate Science Modelling Language (1)

PointSeries	
based on	*SF_SamplingPoint*
phenomenonTime	*TM_Period*
coverage result	CVT_DiscreteTimeInstantCoverage

A time-series of single datum observations at a fixed location. E.g. Tidegauge, buoy, weather station



- Developed by NERC Data Grid
- De-facto INSPIRE standard

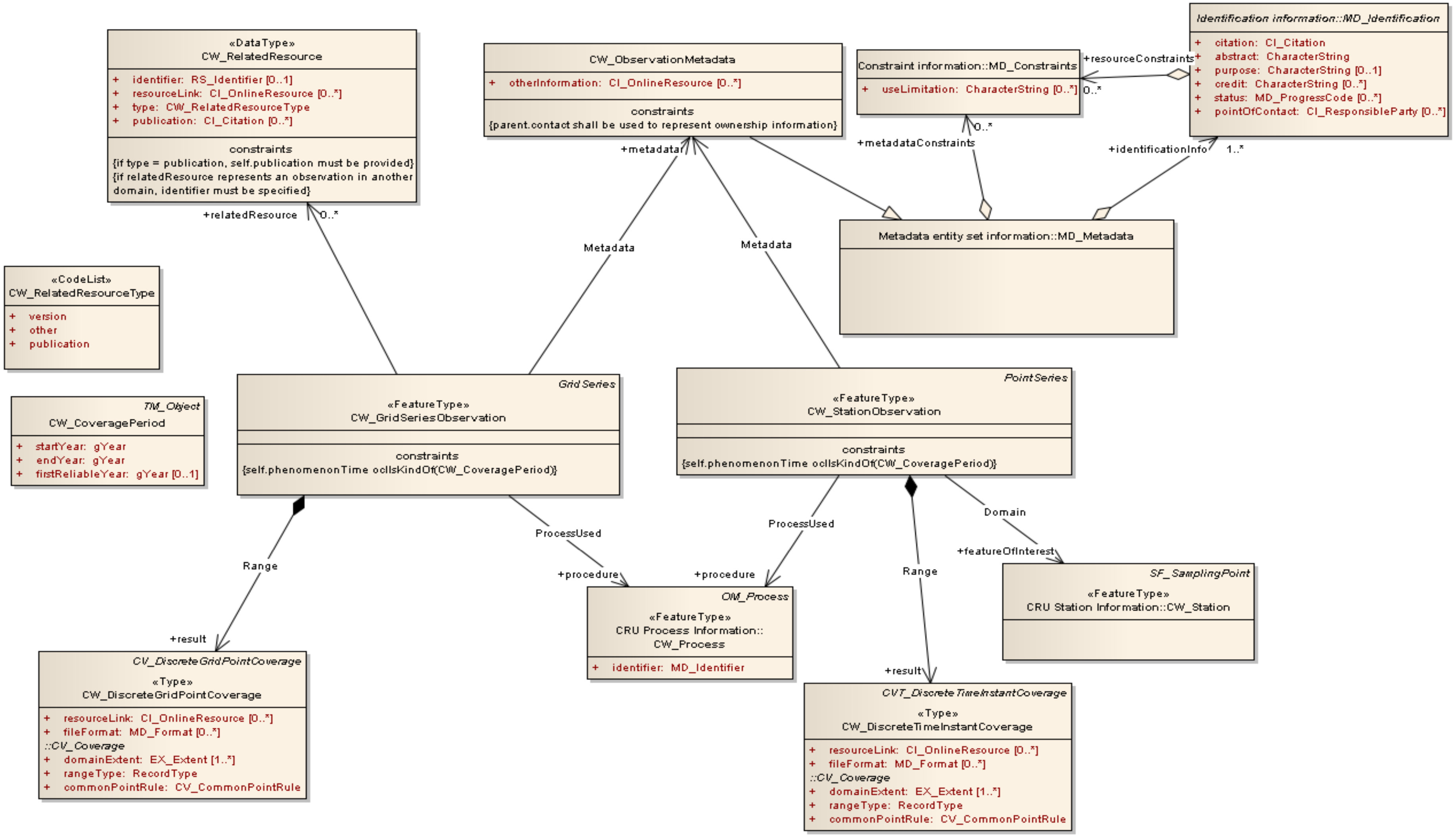
Profile		
based on	*SF_SamplingCurve*	
phenomenonTime	*TM_Instant*	
coverage result	CV_DiscreteGridPointCoverage	
	grid dimension	one
	external CRS	*four (x-y-z-t)*
	alignment	*z-axis*

An observation of some parameter along a vertical line in space. E.g. Wind sounding or radiosonde.

Climate Science Modelling Language (2)

- CSML (v3.0) is an application schema of the ISO O&M model specialised for representing time-series datasets - a perfect fit for the CRU datasets.
  - further but trivial specialisations of CSML would be needed to describe the CRU observations.
- The resultant model would generally be interoperable with both CSML and the ISO O&M model. This would:
  - enable support for existing tools
  - facilitate data sharing, potentially through the INSPIRE SDI

CRU Information Model – Overview (1)



«DataType»
CW_RelatedResource
+ identifier: RS_Identifier [0..1]
+ resourceLink: CI_OnlineResource [0..*]
+ type: CW_RelatedResourceType
+ publication: CI_Citation [0..*]
constraints
{if type = publication, self.publication must be provided}
{if relatedResource represents an observation in another domain, identifier must be specified}

CW_ObservationMetadata
+ otherInformation: CI_OnlineResource [0..*]
constraints
{parent.contact shall be used to represent ownership information}

Constraint information::MD_Constraints
+ useLimitation: CharacterString [0..*]

Identification information::MD_Identification
+ citation: CI_Citation
+ abstract: CharacterString
+ purpose: CharacterString [0..1]
+ credit: CharacterString [0..*]
+ status: MD_ProgressCode [0..*]
+ pointOfContact: CI_ResponsibleParty [0..*]

Metadata entity set information::MD_Metadata

«CodeList»
CW_RelatedResourceType
+ version
+ other
+ publication

TM_Object
CW_CoveragePeriod
+ startYear: gYear
+ endYear: gYear
+ firstReliableYear: gYear [0..1]

Grid Series
«FeatureType»
CW_GridSeriesObservation
constraints
{self.phenomenonTime ocllsKindOf(CW_CoveragePeriod)}

PointSeries
«FeatureType»
CW_StationObservation
constraints
{self.phenomenonTime ocllsKindOf(CW_CoveragePeriod)}

OM_Process
«FeatureType»
CRU Process Information::
CW_Process
+ identifier: MD_Identifier

SF_SamplingPoint
«FeatureType»
CRU Station Information::CW_Station

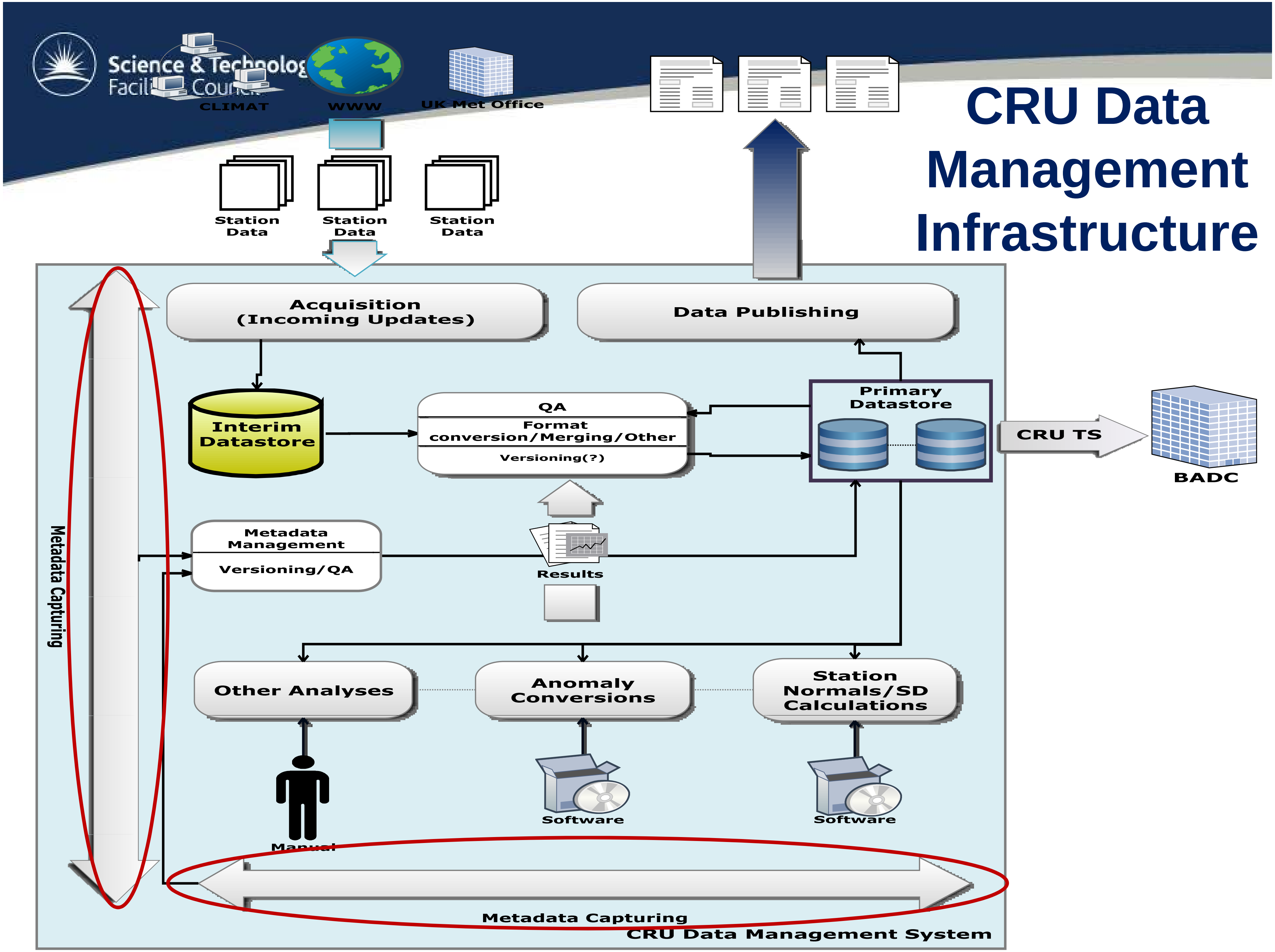
CV_DiscreteGridPointCoverage
«Type»
CW_DiscreteGridPointCoverage
+ resourceLink: CI_OnlineResource [0..*]
+ fileFormat: MD_Format [0..*]
::CV_Coverage
+ domainExtent: EX_Extent [1..*]
+ rangeType: RecordType
+ commonPointRule: CV_CommonPointRule

CVT_DiscreteTimeInstantCoverage
«Type»
CW_DiscreteTimeInstantCoverage
+ resourceLink: CI_OnlineResource [0..*]
+ fileFormat: MD_Format [0..*]
::CV_Coverage
+ domainExtent: EX_Extent [1..*]
+ rangeType: RecordType
+ commonPointRule: CV_CommonPointRule

CRU Information Model – Overview (2)

- Developed as
  - an application schema of the ISO O&M Model
  - with the observation related concepts derived from the CSML *TimeSeriesObservation* classes
- Available in three representation formats: UML, GML schema and RDF Ontology.

CRU Data Management Infrastructure



CLIMAT

WWW

UK Met Office

Station Data

Station Data

Station Data

Acquisition (Incoming Updates)

Data Publishing

Interim Datastore

QA

Format conversion/Merging/Other

Versioning(?)

Primary Datastore

CRU TS

BADC

Metadata Management

Versioning/QA

Results

Metadata Capturing

Other Analyses

Anomaly Conversions

Station Normals/SD Calculations

Manual

Software

Software

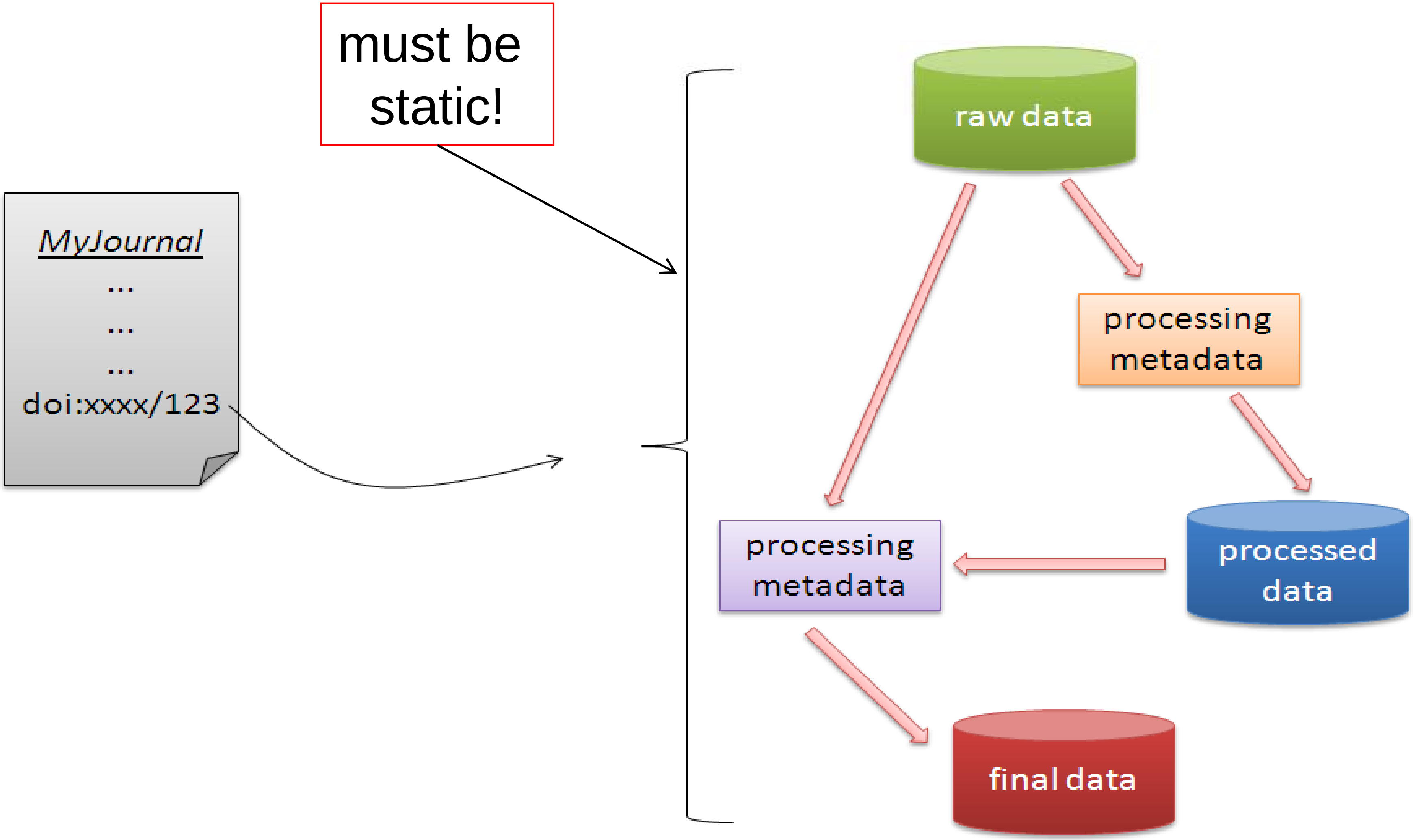
Metadata Capturing

CRU Data Management System

CRU Data Citation (Identification)

- Digital Object Identifier (DOI) for dataset identification
- JISC 14/09 refers to *DataCite* initiative
  - International consortium, incl. British Library, assigning DOIs to datasets
  - Earth System Science Data journal: provides DOIs for data publications
- The main DOI "question" – *What does a DOI point to?*
  - We propose "linked-data".

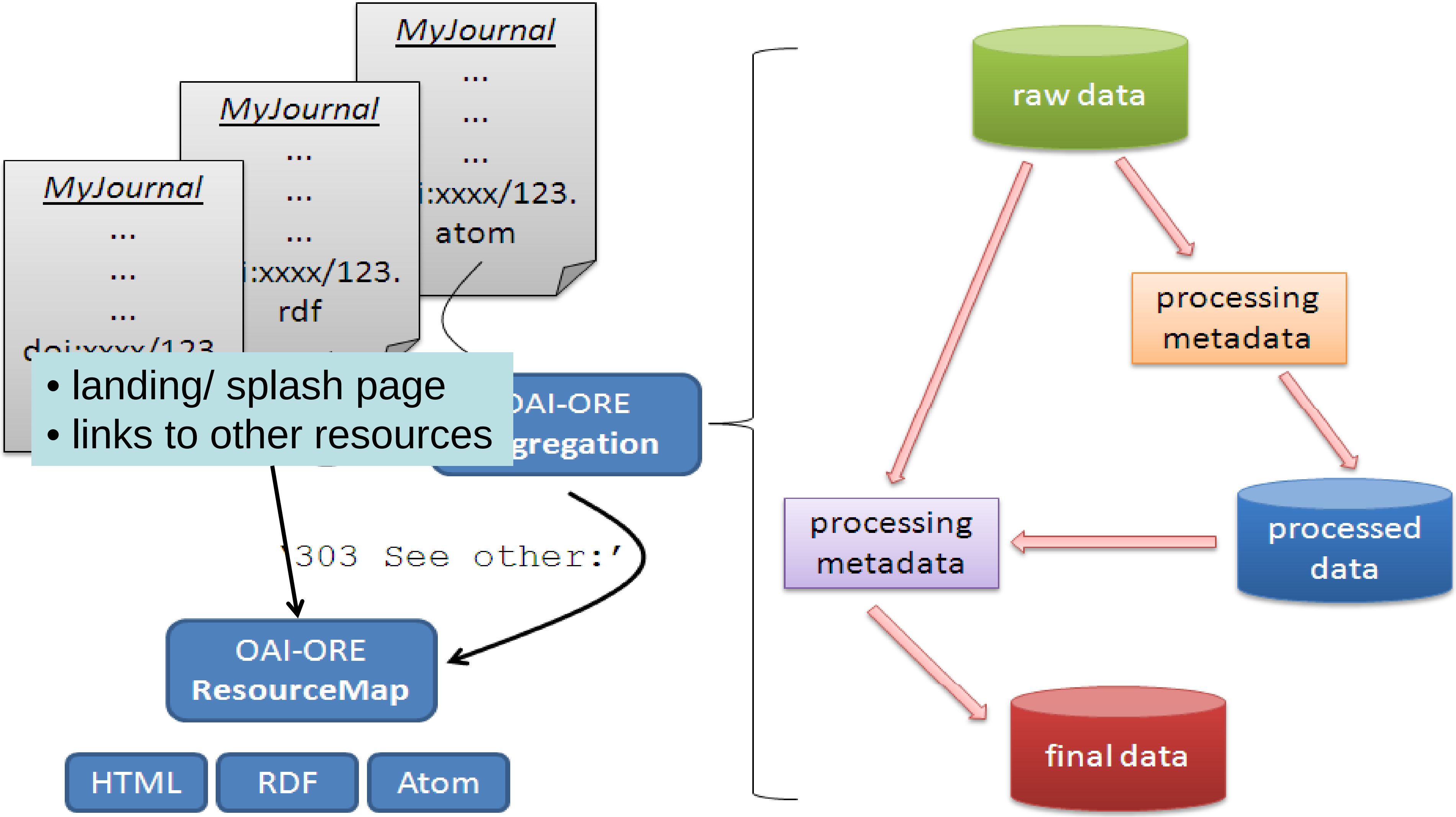
CRU Linked-data



Open Archives Initiative - Object Reuse and Exchange

- 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.
  - **Proxy (P):** used in an assertion specific to an Aggregated Resource (e.g. relationship with another aggregated resource) in the context of a specific Aggregation.

CRU Linked-data in OAI-ORE



- landing/ splash page
- links to other resources

CRU Prototype Linked-data Server

- Use GeoTOD linked-data server
  - Developed by STFC for an STFC/OMII UK funded project
  - Implements of the UK Cabinet Office (2009) “Designing URI Sets for the Public Sector” guidelines.
- GeoTOD will be configured to serve up the workflows for CRUTEM3, CRU TS 3.0, CRU Tree-ring chronologies and HADCET datasets as linked-data

Summary

- ➢ We have developed an information architecture consisting of an information model and a data management infrastructure for CRU
- ➢ The outcomes of ACRID should
  - – improve CRU's current approaches to managing and sharing their weather observation datasets.
  - – facilitate greater transparency and traceability of the data life-cycle
  - – enable improved and interoperable data accessibility and sharing through adoption of suitable ISO standards and linked-data principles

Acknowledgements

- Spiros Ventouras (STFC)
- Jeremy Tandy (UK Met Office)
- Andrew Woolf (Bureau of Meteorology, Australia)

References

- OAI-ORE: http://www.openarchives.org/ore/1.0/datamodel.html
- OMP: http://openprovenance.org
- Datacite: http://www.datacite.org
- DOI: http://www.doi.org
- GeoTOD: http://sourceforge.net/projects/geotod/

Questions?

arif.shaon@stfc.ac.uk

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