SCIENCE AND TECHNOLOGY FACILITIES COUNCIL

Advanced Climate Research Infrastructure for Data (ACRID)

Report II: Work Package 2.2 - Information Architecture

Arif Shaon, Spiros Ventouras (STFC) and Jeremy Tandy (UK Met Office)

23 March 2011

Revision History:

Version	Date	Authors	Sections Affected / Comments
0.1	18/02/2011	AS	Outline
0.2	28/02/2011	AS	Main body
0.3	01/03/2011	AS	Numbering and footnote correction

Contents

1 Introduction and Rationale

A well-designed information architecture is the key to the resourceful management and effective sharing of digital information. An important aspect of such information architecture is the use of standardised approaches to representing and exposing information in order to facilitate interoperability with other information systems. In the past few years this particular aspect has been becoming increasingly pertinent for environmental information, especially in Europe where there has been a drive in the form of INSPIRE[1] directive to share geospatial data through interoperable Spatial Data Infrastructures.

In addition, the 2009 investigation of the House of Commons Science and Technology Committee in to the data management practices of the Climate Research Unit (CRU)[2] at the University of East Anglia identified that it is essential to publish scientific datasets as re-usable scientific resources in their own right with verifiable provenance rather than, as traditionally, as an adjunct to the related publications. In general, this would require capturing and publishing the chain of intermediate data results and their associated metadata (including provenance) along with the final research outputs.

In ACRID, we have therefore formulated an information architecture that is intended to improve the current approaches to managing the CRU datasets by facilitating greater transparency and traceability of the data life-cycle, and enabling improved and interoperable data accessibility and sharing through adoption of suitable ISO standards and linked-data[3] principles.

The information architecture presented in this report principally consists of two components: an **Information Model** (Section 2) and a **Data Management Infrastructure** (Section 3) for the CRU datasets. The former is intended to accurately describe the workflows associated with the CRU datasets, thereby enabling re-enactment of the workflows to verify provenance, and the latter defines various important aspects of managing the CRU datasets including capturing versioning information and metadata, interaction with external data centres as well as identifying the software components that need to be developed and deployed.

2 Terms and Definitions

There are many key terms used throughout this report, some of which may appear to be overlapping with similar terms in some specific discipline(s) and might convey unintended

[1] INSPIRE directive was legislated by the European Parliament and the Council of 14 March 2007 in order to develop a pan-European spatial data infrastructure (SDI) to enable efficient discovery and uniform accessibility of environmental data across Europe. The official INSPIRE website is accessible at: http://inspire.jrc.ec.europa.eu/

[2] http://www.cru.uea.ac.uk/

[3] http://linkeddata.org/

meanings. These terms therefore need to have well-defined meanings in the context of this report. This section compiles and defines those terms and presents them in an alphabetical order.

2.1 Application schema

A conceptual schema for data required by one or more applications [1, definition 4.2]

2.2 Feature

An abstraction of real-world phenomena [1, definition 4.11]

2.3 Feature type

A class of **features** having common characteristics [2, definition 4.6]

2.4 Grid-series

A time-series of gridded parameter fields, e.g. a series of datum measurements made at various geographical locations over a certain time period and organised into a series of adjoining geographical bounding box.

2.5 Observation

An 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.

2.6 Observation procedure

A method, algorithm or instrument, or system of these which may be used in making an **observation** [2, definition 4.10]

2.7 Observation result

An estimate of the value of a property determined through a known observation procedure [2, definition 4.13]

2.8 Point-series

A time-series of single datum measurements at a fixed location.

2.9 Sampling feature

A **feature**, such as a station, transect, section or specimen, which is involved in making **observations** concerning a particular application domain**.** [2, definition 4.16]

2.10 Time-Series

A series of values measured at different points of time as the result of an observation.

3 CRU Dataset Workflow Information Model

We have developed an information model using the Model Driven Architecture (MDA)[4] as adopted by INSPIRE, in order to enable detailed and accurate description of the workflows associated with the CRU datasets. From a wider perspective, the information captured by the model could facilitate verification of the provenance and integrity of the CRU datasets by enabling re-enactment of their workflows.

3.1 Characteristics of the CRU datasets

An analysis of the scientific workflows associated with the CRU datasets (Report I) indicates that the workflows typically consist of a chain of intermediate data results and their associated metadata including the processes used (i.e. provenance) to generate the results. The information model generalises this into the following concepts:

3.1.1 Observation

The practice 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 is referred to as ***Observation*** within the geospatial community. The CRU datasets are essentially the outcomes of such observations that primarily fall under two categories: ***raw or source observation dataset*** collected at various land-based climate monitoring stations or cites around the world, and ***publishable observation dataset*** (e.g. CRU TS[5] dataset) that is derived from the source datasets and typically is published and/or used as the basis for publications.

In addition, the CRU datasets essentially consist of time-series data (see 2.10) with varying structures. For example, the source observation datasets for CRUTEM3[6] are essentially a collection of single temperature measurements collected during a designated period of time at fixed locations (i.e. Point-Series), whereas the CRUTEM3 is a time-series of gridded parameter fields and their corresponding values (i.e. Grid-Series) derived from a collection of the CRU source observation datasets.

3.1.2 Process

A process is essentially the action or the set of actions performed to produce the result (i.e. Dataset) of an observation. In practice, a process may be an algorithm, a computation, a manual

[4] A platform-independent modelling technique that uses a common but domain-specific modelling language, such as the Unified Modelling Language (UML). More information about MDA is available at: http://en.wikipedia.org/wiki/Model-driven_architecture

[5] http://badc.nerc.ac.uk/data/cru/

[6] http://www.cru.uea.ac.uk/cru/data/temperature/

observation or calculation. In addition, a process may consist of a sequence of steps, where the outputs of one step may be used as the inputs of another step that succeeds it.

3.1.3 Processor

This is an entity or a set of entities responsible for performing a process in order to produce the result of an observation. In practice, a processor may be a human, computer software or any type of hardware, such as weather observation instrument.

3.2 Review of Existing Related Information Models

We have reviewed a number of existing information models with a view to identify a suitable model for describing the CRU datasets. Of particular note among these models are ***Open Provenance Model (OPM)***, ***ISO 19156 Observations and Measurements (O&M) model*** and ***Climate Science Modelling Language (CSML)***.

3.2.1 Open Provenance Model

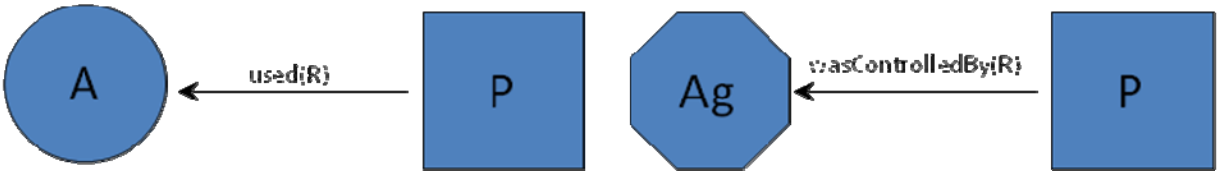

Figure 1: the main concepts of the Open Provenance Model [3]

In general, OPM is intended to enable digital representation of provenance for any object, whether it is digital or physical. To do this, the OPM defines the following three principle notions:

1. **Artifact (A)**: a digital or physical object in an immutable state
2. **Process (P):** "Action or series of actions performed on or caused by artifacts, and resulting in new artifacts" [3].
3. **Agent (Ag)**: an entity or a set of entities responsible for conducting or contributes towards the conducting a process.

Additionally, OPM defines a number of auxiliary concepts to represent the relationships between the aforementioned three main concepts (Figure 1).

A comparison of the OPM concepts with the main concepts (Section 3.1) of the workflows associated with the CRU datasets indicates a close parallel between these concepts. Conceptually, the OPM Artifact, Process and Agent concepts are analogous to the CRU Dataset, Process and Processor concepts respectively. However, as mentioned before, the OPM concepts are very generic and abstract; therefore, they would need to be specialised substantially to

accurately capture the geospatial aspects of the CRU workflow concepts. In addition, OPM is not commonly used within the geospatial community; hence it may not be compatible with other existing information systems and models.

3.2.2 ISO 19156 Observations and Measurements (O&M) Model

The ISO 19156 O&M model [2] defines a conceptual schema for describing environmental observations (Section 2.1) and the features involved in the sampling associated with such observations. This conceptual schema could also be used to exchange information describing observation acts and their results between communities.

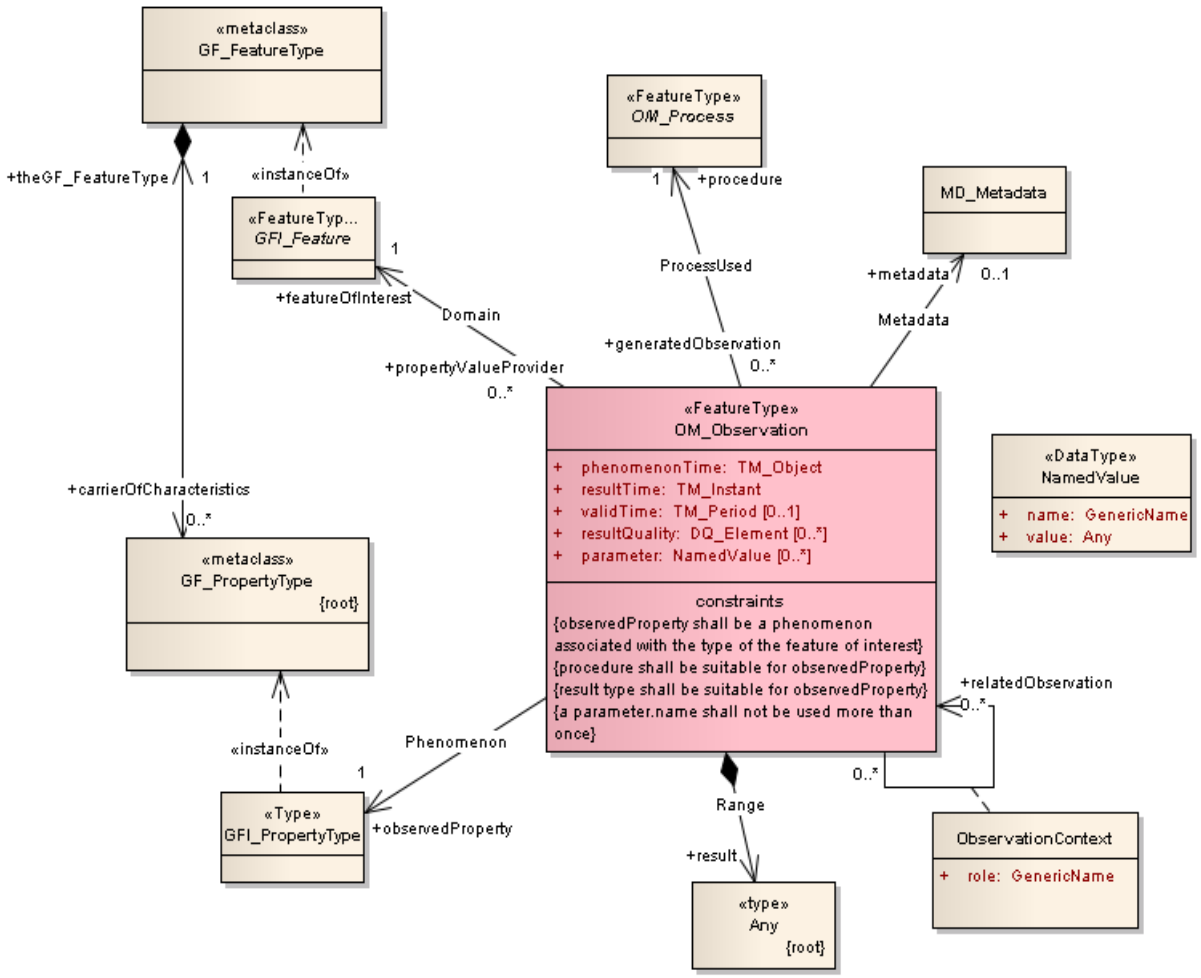

Figure 2: The Core ISO O&M Observation Type

In summary, the O&M conceptual schema defines a number of UML classes (Figure 2) for representing different aspects of an observation, such as *OM_Observation* for representing an observation and *OM_Process* for describing the procedure used to generate the result of an

observation. The schema also uses concepts from other related ISO standards including the ISO 19115 metadata model [4] to represent additional information about an observation, such as the feature of interest and its property associated with an observation [2].

In contrast with the OPM, the ISO O&M Model is specifically designed for describing environmental observations, such as the ones represented by the CRU datasets. However, in common with the OPM, the ISO O&M model too is intended to be generic, albeit offering a few example specialised observation types, such as Temporal Coverage Observation [2]. Therefore, specialisation of the main ISO O&M classes, such as *OM_Observation* and *OM_Process* (Figure 2) would be needed to capture the distinct characteristics (e.g. gridded time-series data, links to publications etc.) of the CRU observations.

3.2.3 Climate Science Modelling Language (CSML)

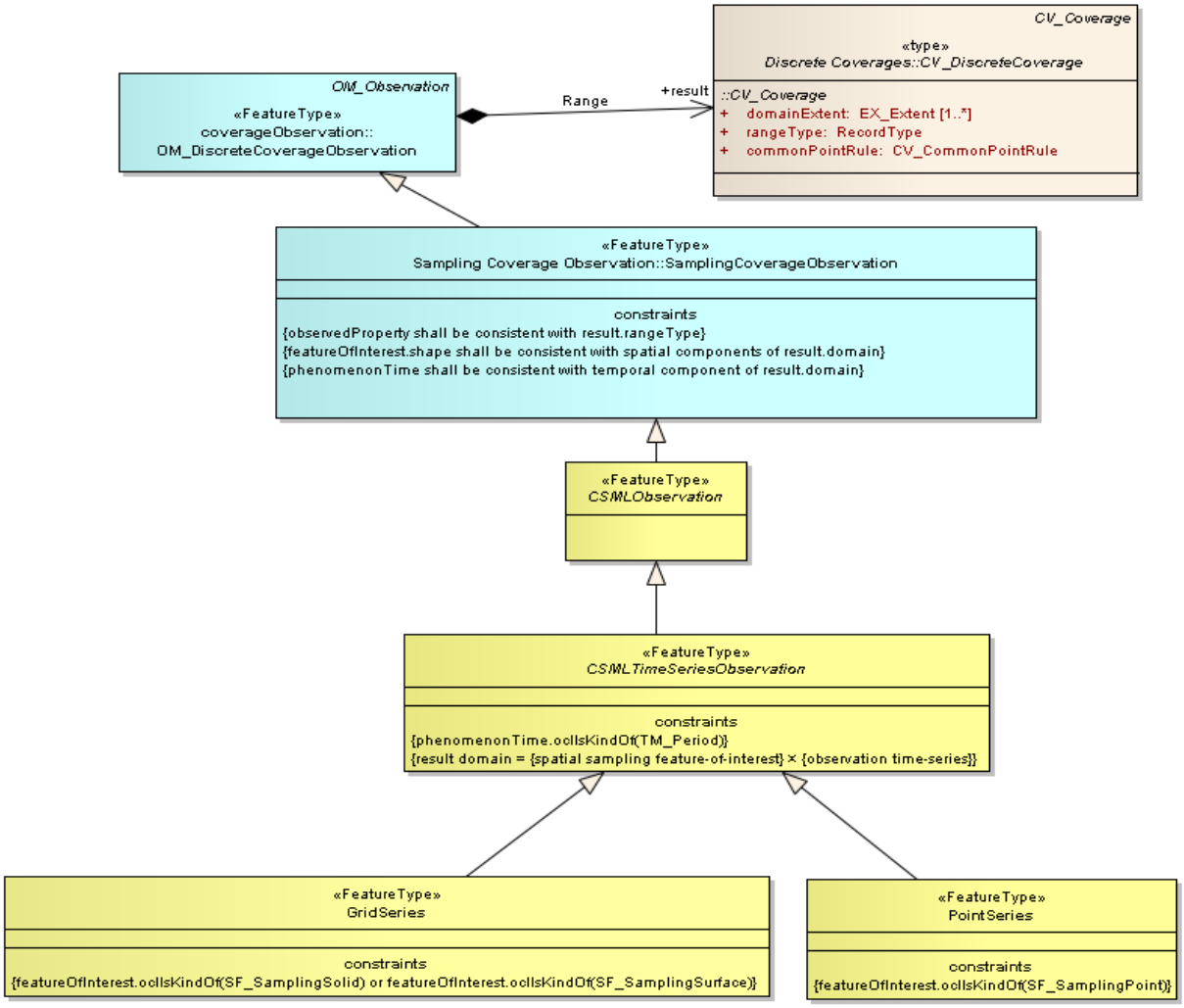

Figure 3: CSML GridSeries and PointSeries Observation Types

The Climate Science Modelling Language (CSML)[7] was originally developed as part of the NERC Data Grid (NDG)[8] project in the UK as an application schema (see 2.1) of Geography Markup Language (GML)[9] to describe complex feature types for the atmospheric and oceanographic domain. More pertinently, CSML defines specific feature types (e.g. *PointSeriesFeature*, *GridSeriesFeature* etc.) for representing time-series data. In addition, CSML has recently been aligned with ISO O&M Model by combining the O&M concept of 'sampling features' (see 2.9) together with an observation result based on the ISO 19123 coverage model [5] (Figure 3). Furthermore, CSML has a growing user community lead by the British Atmospheric Data Centre (BADC)[10] in terms of developing and providing tools and software support for understanding and manipulating datasets encoded in CSML.

Considering the fact that CSML is effectively an application schema of the ISO O&M model specialised for representing time-series datasets, it would be a perfect fit for the CRU datasets. Of course, further but trivial specialisations of CSML would be needed to enable detailed and structured descriptions of the CRU observations. But, the resultant model for the CRU datasets would generally be interoperable with both CSML and the ISO O&M model. The former would enable existing tools that support CSML to also understand the CRU model, thereby facilitating processing and manipulating the CRU datasets. The latter (i.e. interoperability with the ISO O&M model) on the other hand, would facilitate the CRU datasets to be shared with a wider Geospatial community, potentially through a global Spatial Data Infrastructure, such as the INSPIRE SDI.

Therefore, in view of the aforementioned advantageous features of CSML, the CRU information model has been developed as an application schema of the ISO O&M Model, with the observation related concepts derived from the *CSMLTimeSeriesObservation* classes (Figure 3). The UML classes of the CRU Information model are described in the following section.

[7] http://csml.badc.rl.ac.uk/

[8] http://proj.badc.rl.ac.uk/ndg

[9] http://www.opengeospatial.org/standards/gml

[10] British Atmospheric Data Centre – http://badc.nerc.ac.uk/home/index.html

3.3 UML Conceptual Schema for the CRU Observations

3.3.1 Packaging

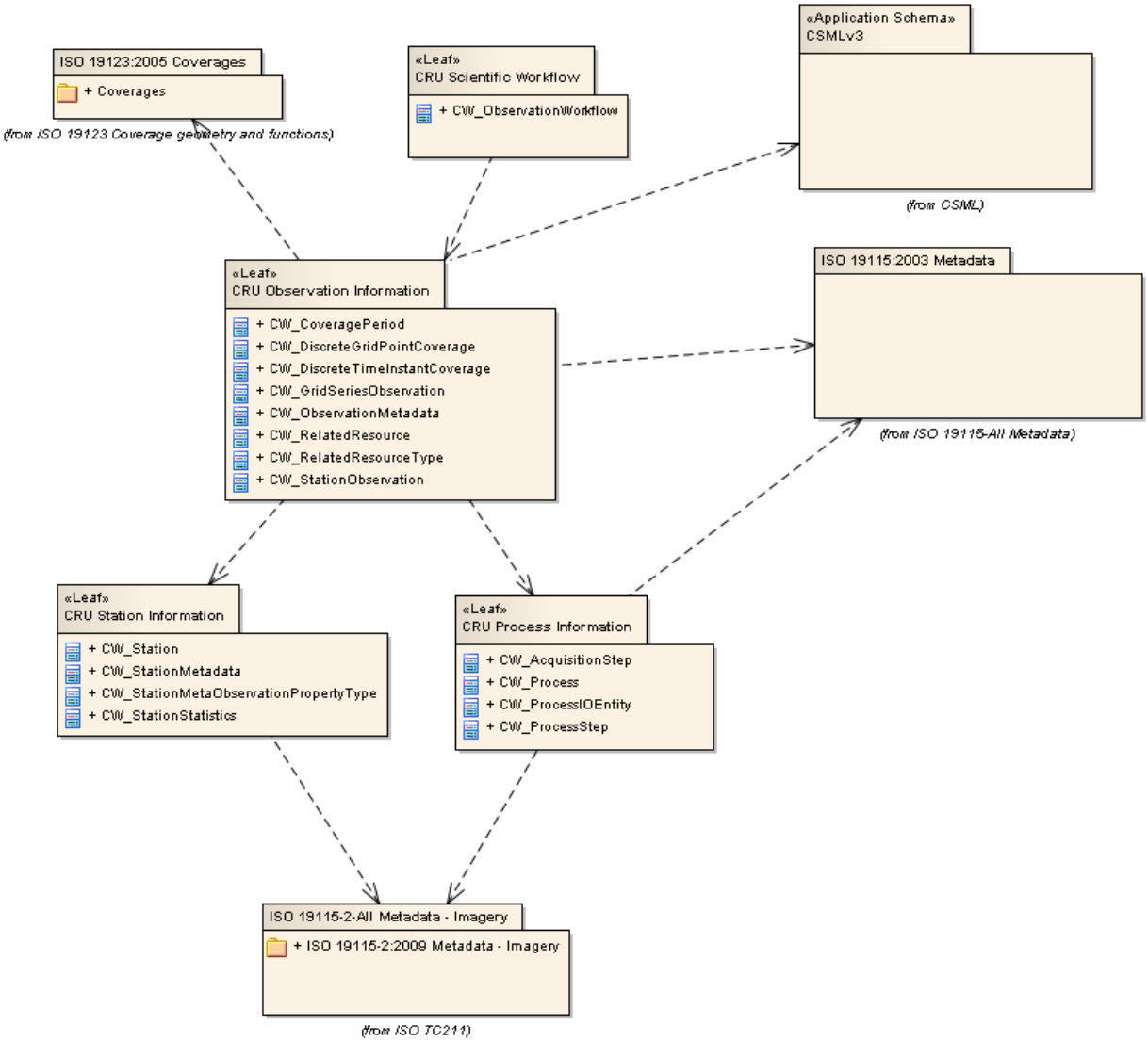

Figure 4: Inter-package dependencies of the CRU Conceptual Schema

The CRU application schema is organized into four leaf packages: Observation Information, Process Information, Station Information, and Workflow Information, with dependencies on several other packages from geographic information International Standards. The inter-package dependencies are shown in Figure 4. Please note that in keeping with the ISO UML class naming convention, the UML class names in the CRU conceptual schema contain a two-letter prefix "CW", which is an abbreviation of "CRU Workflow".

3.3.2 CRU Observation Classes

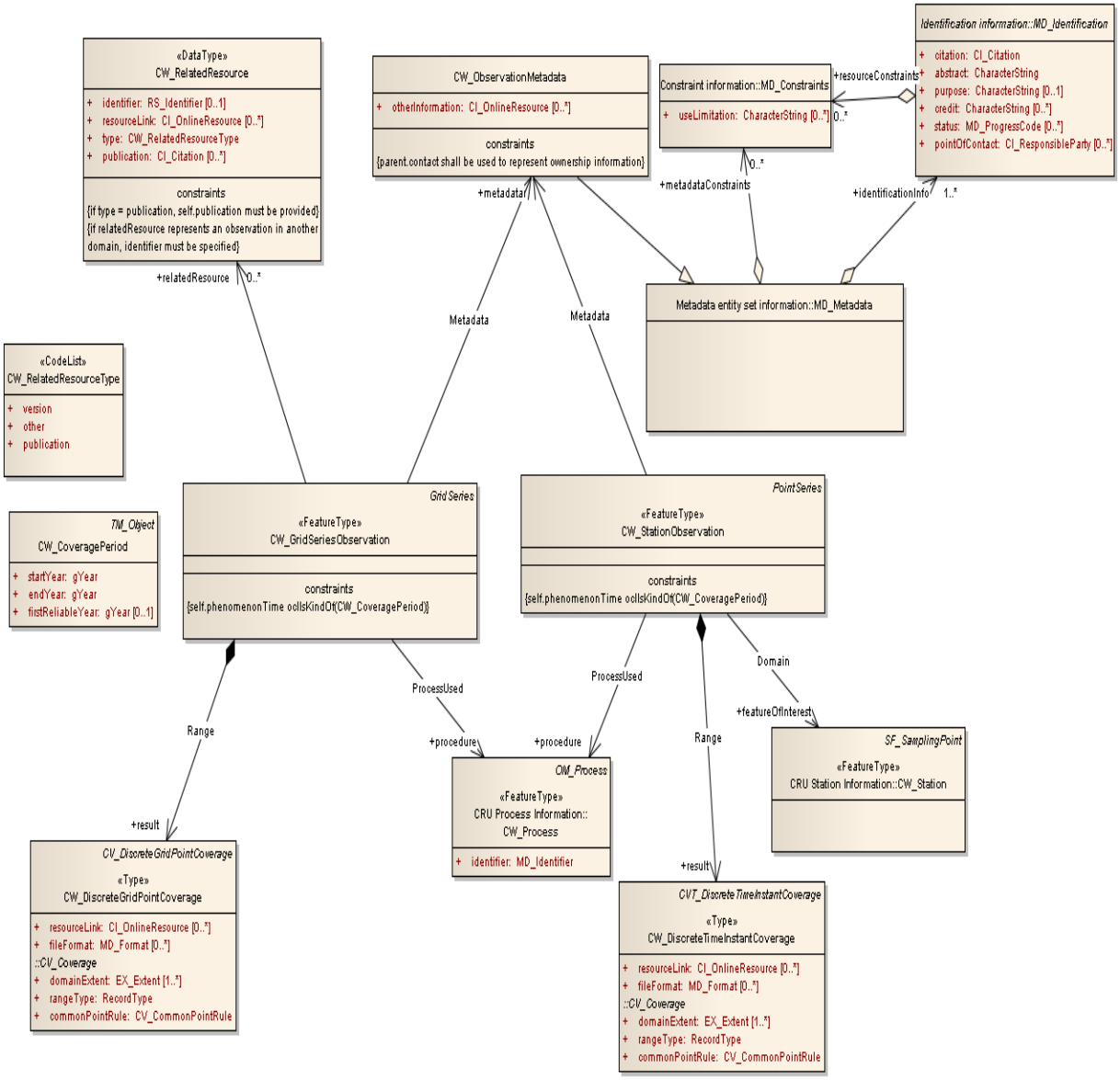

Figure 5: CRU Observation Types

3.3.2.1 CW_GridSeriesObservation

The class *CW_GridSeriesObservation* is intended to represent the gridded observational datasets that are usually published and/or used as publication materials by CRU. In the schema, it extends the CSML class *GridSeries*, which is a specialisation of the class *CSMLTimeSeriesObservation* that is essentially of type *OM_DiscreteCoverageObservation* from the ISO O&M model. The CSML class *GridSeries* represents the observations that result into grid-series data results (see

2.4). For example, CRUTEM3 consists of a series of gridded datum measurements and their corresponding time points.

The class *CW_GridSeriesObservation* adds a number of new properties and constraints, and overrides some of the existing properties of CSML *GridSeries* class (Figure 5) to provide accurate and detailed representation of the CRU grid-series datasets including their references to the related publications and external datasets (Annex A.1).

3.3.2.2 CW_RelatedResource

This class is intended to represent a reference or a link to a publication or an observation or any other resource that is of relevance to a CRU grid-series observation (represented by the class *CW_GridSeriesObservation*) but exists externally to it. For example, a scientific paper describing the CRUTEM3 or one of its variants, e.g. CRUTEM3v (a "variance adjusted" version of CRUTEM3) may be represented as an instance of *CW_RelatedResource* that could be linked to a *CW_GridSeriesObservation* object representing CRUTEM3. The main advantage of this approach is facilitating resource discovery; searching for CRUTEM3, e.g. in a linked-data browser could also discover and fetch CRUTEM3v and other related resources.

The properties and constraints associated with the class *CW_RelatedReource* are outlined in Annex A.2.

3.3.2.3 CW_DiscreteGridPointCoverage

The class *CW_DiscreteGridPointCoverage* represents the result (e.g. CRUTEM3) of a CRU Grid-series observation. This class extends the ISO 19123:2005 class *CV_DiscreteGridPointCoverage* [5] by adding suitable properties for representing the observation result as a web accessible resource (CRU datasets usually are accessible from their website) as well as indicating the file format (e.g. NetCDF) type of that resource.

3.3.2.4 CW_StationObservation

This class represents a raw or source observation (Section 3.1) conducted at a climate monitoring station. The results of these types of observations are used by CRU to produce publishable observations, such as the CRU grid-series observation. The class *CW_StationObservation* specialises the CSML *PointSeries* class (essentially an *OM_DiscreteObservation* from the ISO O&M model) that is specifically defined to describe observations yielding point-series data (Figure 5).

The class *CW_StationObservation* adds a number of new properties and constraints, and overrides some of the existing properties of CSML *PointSeries* class (Figure 5) to provide accurate and detailed representation of the CRU station datasets (Annex A.4).

3.3.2.5 CW_DiscreteTimeInstantCoverage

The class *CW_DiscreteTimeInstantCoverage* represents the result of a CRU station (point-series) observation. This class extends the ISO O&M class *CVT_DiscreteTimeInstantCoverage*, a specialization of *CV_DiscreteCoverage* [5], by adding suitable properties for representing CRU station observation datasets as web accessible resources as well as enabling recording the associated file formats (e.g. typically ASCII).

3.3.2.6 CW_ObservationMetadata

The class *CW_ObservationMetadata* is an extension of the class *MD_Metadata* from the ISO 19115:2003 Metadata model [4]. In particular, this class uses the "contact" and "identificationInfo" properties of the class *MD_Metadata* to describe ownership and constraints (e.g. for use and access) related information associated with a CRU observation amongst other information (Figure 5).

3.3.2.7 CW_CoveragePeriod

This class is a specialisation of the ISO 19108: 2006 class *TM_Object* [6] to represent the phenomenon time of a CRU Observation, i.e. the time period to which the observation result applies. This class adds a number of suitable properties (Annex A.6) to represent coverage periods applicable to the CRU datasets.

3.3.3 CRU Process Information

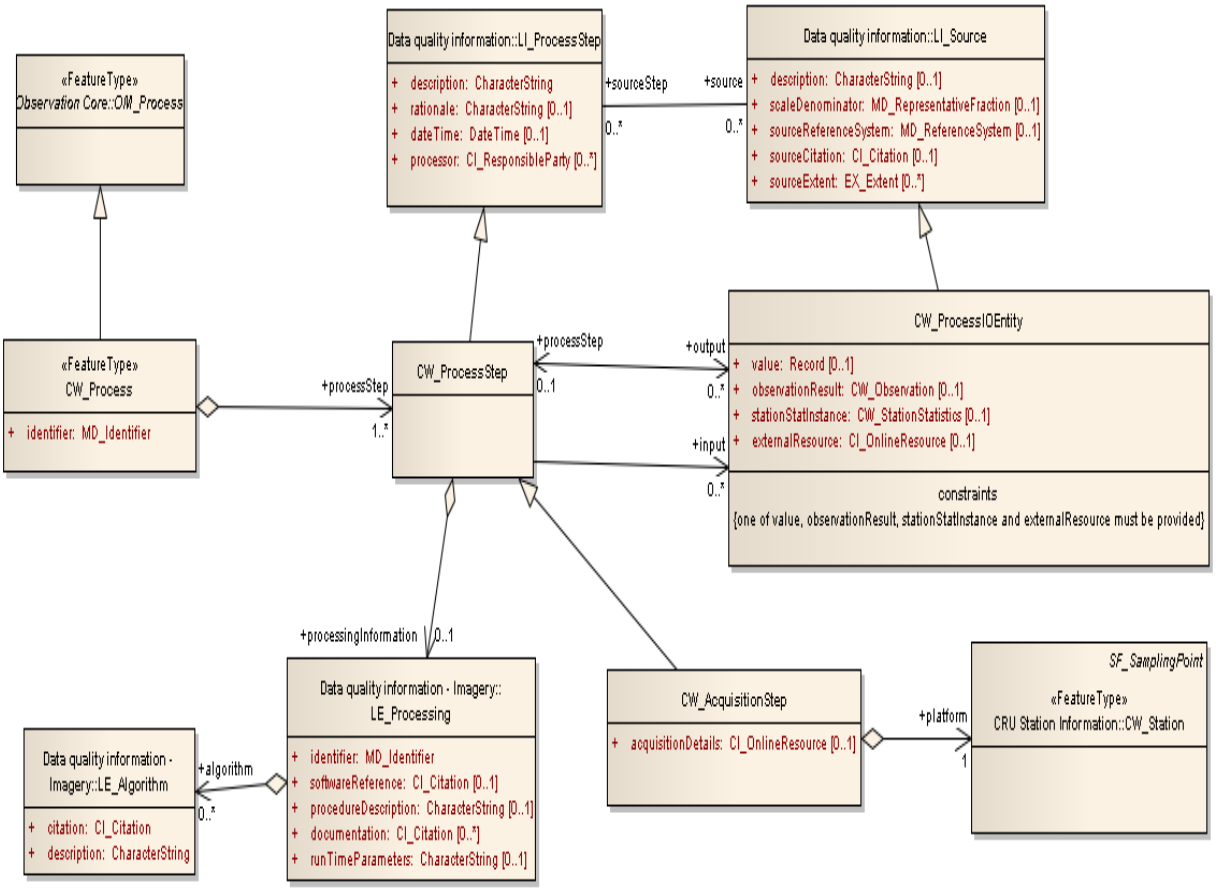

Figure 6: CRU Process Classes

3.3.3.1 CW_Process

The class *CW_Process* specialises the core O&M class *OM_Process* to describe various aspects the process associated a CRU observation. In particular, it adds information about the different steps in a process through the property "processStep" (Figure 6). The properties of the class *CW_Process* are outlined in Annex B.1.

3.3.3.2 CW_ProcessStep

The class *CW_Process* extends the ISO 19115:2003 class *LI_ProcessStep* [4] to capture detailed information about a step involved in a process associated with a CRU observation. It adds a number of new properties to provide a comprehensive description of the various aspects of a process step including inputs and outputs, algorithm employed and **processor** information (3.1.3), such as software used and its parameters. Details of these properties are provided in Annex B.2.

3.3.3.3 CW_ProcessIOEntity

This class is a specialised *LI_Source* class from the ISO 19115 metadata model [4]. It is intended to represent the inputs and outputs that are associated with a process step (i.e. represented by an instance of *CW_ProcessStep* class). In general, it captures the results of CRU observations as the inputs or outputs of a process while also providing scope for describing any interim data results yielded and any external resources used as inputs by a process step. More information about the properties and constraints associated this class is provided in Annex B.3.

3.3.3.4 CW_AcquisitionStep

This is a specialised *CW_ProcessStep* specifically intended for representing the process step for acquiring a CRU Station dataset (i.e. source observation). The distinguishing feature of this class is the "platform" property, which describes the station information (e.g. physical location, ownership) associated a CRU station observation. In practice, CRU usually have no or very limited information about the actual process used for conducting a source observation and acquiring the resulting dataset from a climate monitoring station as there are usually external organisations responsible for such activities. But the class *CW_AcquisitionStep* could be used, possibly with further specialisation, to represent such information, if it becomes available to CRU.

3.3.4 CRU Station Information

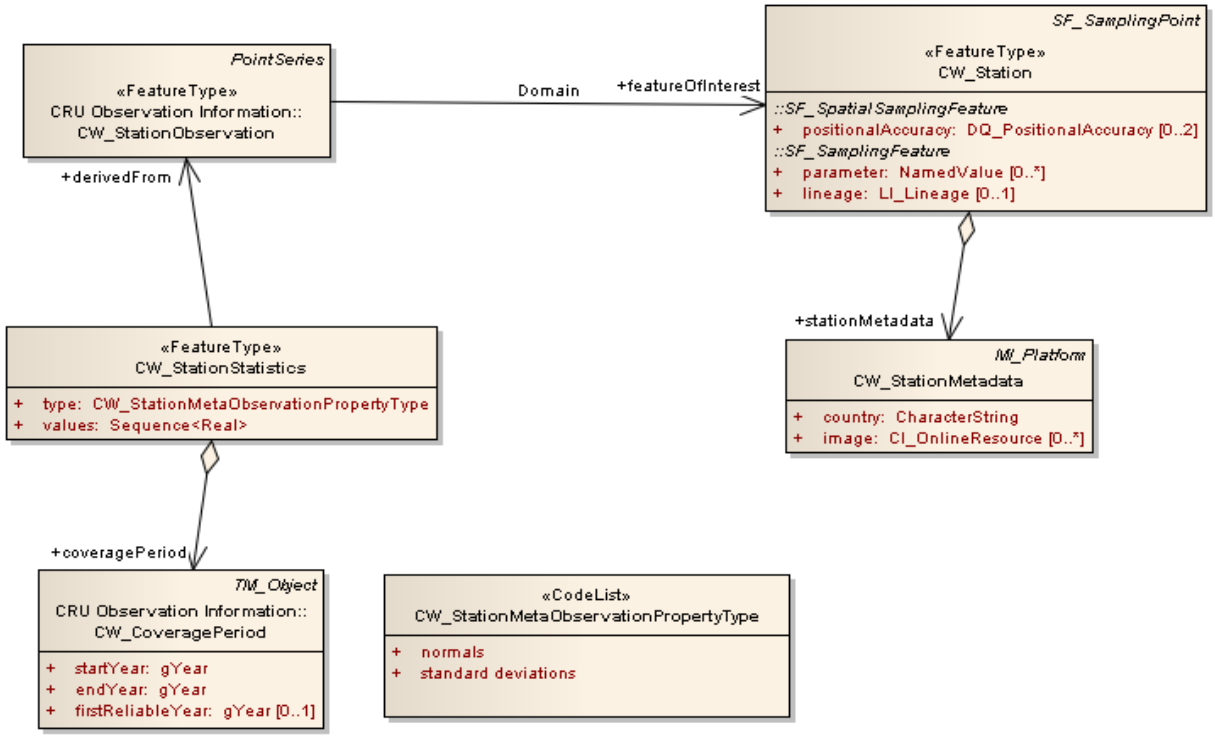

Figure 7: CRU Station classes

3.3.4.1 CW_Station

This class describes a climate monitoring station as an ISO O&M *SamplingFeature* and based on the definition provided by the World Meteorological Organization (WMO). By definition, the class *CW_Station* is a specialization of the ISO O&M class *SF_SamplingPoint*, which defines a number of properties to describe the geospatial aspects (e.g. geographical location) of the climate monitoring stations at which the CRU source observation datasets are collected. The class *CW_Station* defines additional properties to represent other general information about a station, such as ownership information, web page, image etc. More information about the *CW_Station* properties is provided in Annex C.1.

3.3.4.2 CW_StationMetadata

This is an extension of the ISO 19115-2:2009 class *MI_Platform* [7], mainly intended for describing the non-geospatial aspects of a Climate Monitoring Station, such as station identifier and ownership information.

3.3.4.3 CW_StationStatistics

The class *CW_StationStatistics* is specifically intended to represents the two types of interim statistical entities, *Normals* and *Standard Deviations* that are derived from the CRU source observation datasets as part of the process associated with the CRUTEM3 dataset. These interim statistical entities are not of primary interest, though they are usually persisted after the production of CRUTEM3. The properties of the class *CW_StationStatistics* are described in Annex C.3.

3.3.5 CRU Scientific Workflow Information

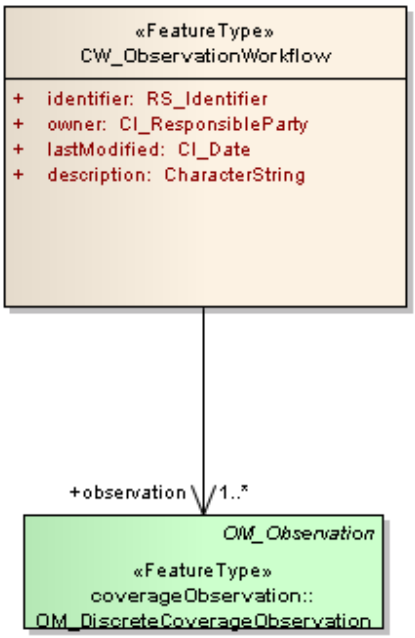

Figure 8: CRU Workflow Classes

3.3.5.1 CW_ObservationWorkflow

The class *CW_ObservationWorkflow* is effectively a wrapper class that encapsulates the CRU observation and process instances to provide a coherent and structure view of the workflow associated with a CRU observation dataset. By definition, this class can be used to encapsulate an instance of the ISO O&M *OM_Observation* class or any of its subclasses, such as the CRU and CSML observation classes. Therefore, it provides flexibility in terms of defining new observation types for the CRU datasets, if necessary. Furthermore, it defines a number of properties to record additional metadata about a CRU workflow. These properties are listed in Annex D.1.

4 CRU Data Management Infrastructure

We have analysed CRU's current approaches to managing their datasets and identified the need for more effective mechanisms for capturing information about their dataset workflows as well as managing different versions of their datasets. We have therefore designed a data management infrastructure that is mainly intended to incorporate the ability to accurately and efficiently capture information about CRU's dataset workflows (represented by the CRU Workflow Information Model – Section 3) into their current data management practices. In addition, this infrastructure is designed to articulate and improve some of the existing data management related processes of CRU, such as data versioning and publishing.

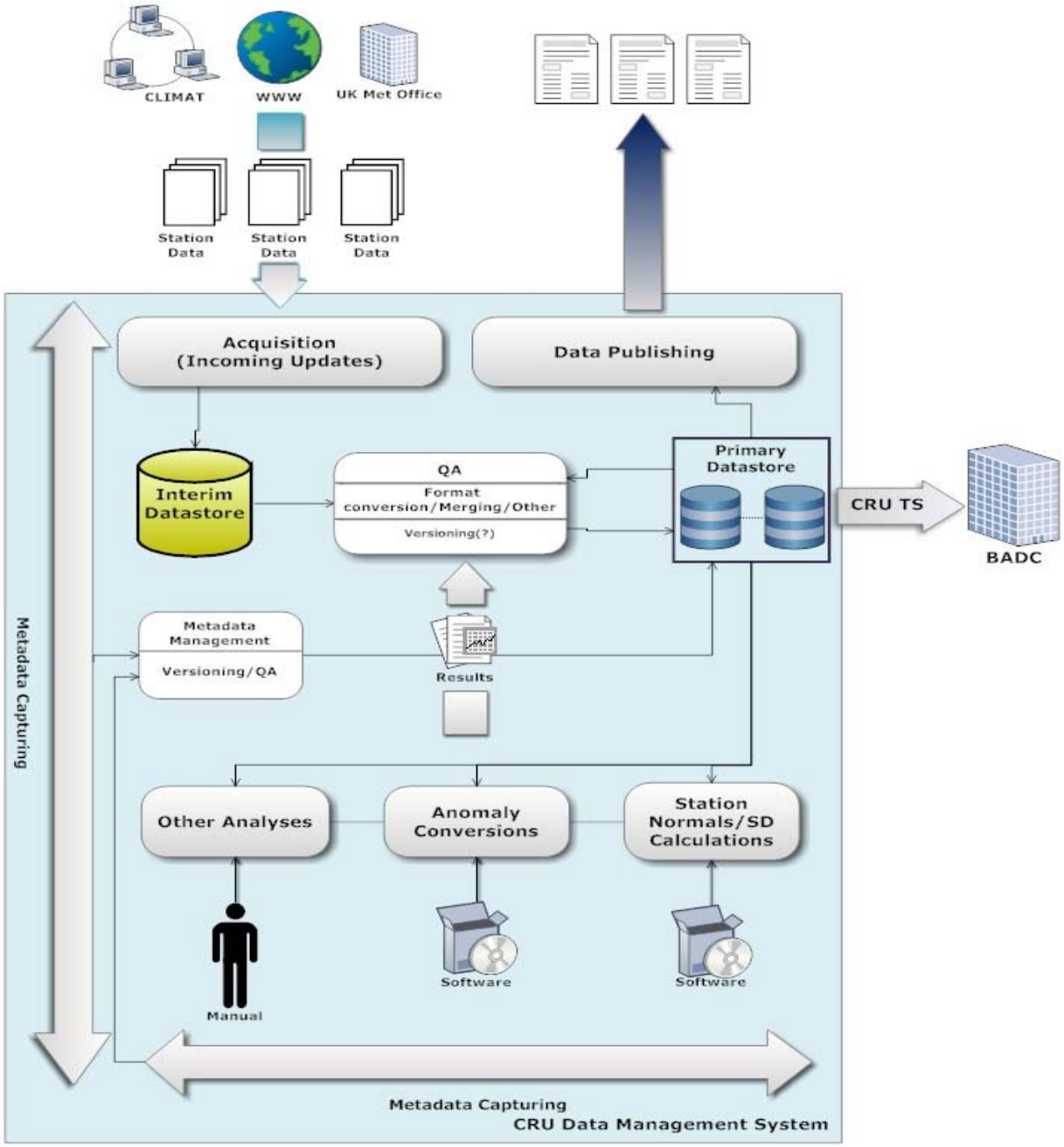

Figure 9: A Data management infrastructure for CRU

The following subsections describe the main components of this data management infrastructure as illustrated in Figure 9.

4.1 Data Acquisition

CRU receives source or raw weather observation datasets (Section 3.1) from a range of sources including the UK Met Office[11]and the CLIMAT[12] network, the WMO's global network for the transmission of meteorological data from various weather stations around the globe. These raw datasets are initially stored as received into an **Interim Datastore** before being subjected to quality assurance and harmonisation including format conversions (to supported formats) and merging with existing datasets, in cases where the newly received datasets are updates to the existing datasets. After being quality assured and harmonised, the source datasets are ingested into a primary (file-based) data store.

Of particular note here is the process for making updates to the existing datasets. The current practice is to replace the existing data with new or merged data; the old data is not retained. This could potentially make reversing to a previous version of a dataset (if needed e.g. for disaster recovery, integrity verification etc.) extremely difficult; it would require re-acquiring the appropriate version of the source data, provided it is still available, and then re-processing it.

As a long-term resolution to the aforementioned issue, the data management infrastructure presented in this report proposes incorporation of versioning of CRU's source observation datasets into the data quality assurance and harmonisation process, through a suitable version management system, such as **Subversion**[13].

4.2 Data Analysis

After quality assurance and harmonisation, CRU's source observation datasets are used as the inputs of a wide range of analyses and processes to produce publishable observation datasets (Section 3.2). These publishable datasets are also stored in a file-based **Primary Datastore** (except for CRU TS, which is submitted to BADC for storage and management). As mentioned before, the analytical processes for creating the publishable datasets typically generate various interim data results (e.g. normals and standard deviations for station data), some of which are persisted and stored, also in a file-based database.

As with CRU's source observation datasets, interim changes to the publishable datasets are not retained; only the major versions are persisted. For example, CRUTEM2, the immediate ancestor of CRUTEM3 is still persisted and accessible, though the interim versions of CRUTEM3 are not.

We therefore propose employment of an efficient change management system (e.g. Subversion or CVS) for managing different versions, both major and interim, of CRU's publishable datasets.

[11] http://www.metoffice.gov.uk/

[12] http://en.wikipedia.org/wiki/CLIMAT

[13] http://subversion.tigris.org/

4.3 Metadata Capturing

In order to ensure effective use of the information model presented earlier in this report (in terms of enabling greater transparency of the workflows associated with the CRU datasets), information represented by the model would need to be captured at various important stages of the lifecycles of the datasets, quality assured, versioned and finally stored for dissemination. While the mechanisms employed for capturing metadata could vary between different stages of the data life-cycle, they should ideally be fully or semi automated wherever possible. For example, it should not be difficult to record in an automated fashion, the inputs, outputs and other related parameters of a software run that is a part of an analytical process conduced on a dataset. However, as mentioned before, the current data management practices of CRU lacks such efficient mechanisms for capturing important information about the processes (as mentioned above) executed on their datasets.

Therefore, as illustrated in Figure 9, the data management infrastructure proposes employment of efficient mechanisms for capturing metadata about the important stages of the life-cycles of the CRU datasets, quality assurance and versioning of the metadata captured and finally storing it, ideally in a medium (e.g. relational database, or RDF triple store) that is suitable for efficient querying and dissemination of the metadata.

4.4 Data Publishing

Traditionally, CRU have published their climate datasets in the form of journal publications with the actual datasets made available as web-accessible resources on their website. However, this practice has been identified (by the House of Commons investigation) as inefficient for providing sufficient transparency of the life-cycles of CRU's datasets, and thereby enabling traceability and verification of their provenance and integrity. This is because it is not always possible to detail in a structured and re-usable manner, the entire workflow associated with a dataset within the limited scope of a journal paper.

Therefore, we propose publishing the complete workflows associated with the CRU datasets as linked-data to allow effective sharing and querying as well as providing greater transparency into CRU's scientific workflows. We have already developed a linked-data server as the outcome of another project (GeoTOD[14]) which will be customised and configured for serving up CRU's scientific dataset workflows as linked-data.

[14] Geospatial Transformation with OGSA-DAI - http://geotod.sourceforge.net/

5 Conclusions and Future Work

In the current phase of the project, we have developed an information architecture consisting of an information model and a data management infrastructure for CRU with a view to improve their current approaches to managing and sharing their weather observation datasets. The next phases of the project will focus on producing an RDF ontology representation of the CRU information model and developing tools necessary to capture, store and expose information represented by the model. This will be followed by the deployment of these tools and other software components that are proposed by the data management infrastructure presented earlier in this report. These activities may dictate further changes to be made to both the information model and data management infrastructure.

References

[1] ISO 19101:2002 - Geographic information -- Reference model

[2] ISO 19156:2010 - Geographic information — Observations and measurements

[3] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. and Van den Bussche, J. (2010) The Open Provenance Model core specification (v1.1). *Future Generation Computer Systems*, last accessed 26 February 11 - http://eprints.ecs.soton.ac.uk/id/eprint/21449

[4] ISO 19115:2003 Geographic information – Metadata

[5] ISO 19123:2005 - Geographic information -- Schema for coverage geometry and functions

[6] ISO 19108: 2006 – Temporal Schema

[7] ISO 19115-2:2009 Metadata – Imagery

[8] ISO 19107:2003 Spatial Schema

Annex A: Properties of the CRU Observation Classes

A.1: Properties of CW_GridSeriesObservation Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
procedure	The process used to produce the result associated with a CRU grid-series observation.	Mandatory	1	An instance of *CW_Process* (3.3.3.1)
Result	The outcome/result of a CRU grid-series observation.	Mandatory	1	An instance of *CW_DiscreteGridPointCoverage* (3.3.2.3)
relatedResource	The reference or link to a publication, an external observation (e.g. a previous version of the observation being described) or any type of resource that is of relevance to a CRU grid-series observation.	Optional	*	An instance of *CW_RelatedResource*
Metadata	Additional information about a CRU grid-series observation including its identifier, purpose and ownership information.	Mandatory	1	An instance of *CW_ObservationMetadata* class (3.2.3.6)
phenomenonTime	Inherited from the ISO O&M class *OM_Observation*, this	Mandatory	1	An instance of *CW_CoveragePerio*

	property describes the time that a CRU grid-series observation result applies to the property of the feature-of-interest [2].			*d* class (3.2.3.7)

A.2: Properties of CW_RelatedResource Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
identifier	An identifier to the external resource	Mandatory if the external resource is an observation else optional	1	An instance of *RS_Identifier* (ISO 19115:2003)
resourceLink	A web link to the external resource	Optional	*	An instance of *CI_OnlineResource* (ISO 19115:2003)
type	A text indicating the type of the external resource (e.g. publication, version, etc.).	Mandatory	*	A value from the CodeList *CW_RelatedResourceType* (Figure 5)
publication	a description of the external resource which is a publication	Mandatory if type is “publication”	1	An instance of the class *CI_Citation* (ISO 19115:2003)

A.3: Properties of CW_DiscreteGridPointCoverage Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
resourceLink	A link to the grid-series observation result accessible on the web.	Optional	*	An instance of *CI_OnlineResource* (ISO 19115:2003)
fileFormat	A text indicating the file format (e.g. NetCDF, ASCII) of the web-accessible grid-series data.	Optional	*	An instance of *MD_Format* (ISO 19115:2003)

A.4: Properties of CW_StationObservation Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
procedure	The process used to produce the result associated with a CRU Station (i.e. point-series) observation.	Mandatory	1	An instance of *CW_Process* (3.3.3.1)
result	The outcome/result of a CRU Station observation.	Mandatory	1	An instance of *CW_DiscreteTimeInstantCoverage* (3.3.2.5)
featureOfInterest	Describes the station information (e.g. physical location, ownership) associated a CRU station observation	Mandatory	1	An instance of *CW_Station*

Metadata	Additional information about a CRU station observation including its identifier, purpose and ownership information.	Mandatory	1	An instance of *CW_ObservationMetadata* (3.2.3.6)
phenomenonTime	Inherited from the ISO O&M class *OM_Observation*, this property describes the time that a CRU station observation result applies to the property of the feature-of-interest [2].	Mandatory	1	An instance of *CW_CoveragePeriod* class (3.2.3.7)

A.5: Properties of CW_DiscreteTimeInstantCoverage Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
resourceLink	A link to the point-series station observation result accessible on the web.	Optional	*	An instance of *CI_OnlineResource* (ISO 19115:2003)
fileFormat	A text indicating the file format (typically ASCII) of the web-accessible point-series station data.	Optional	*	An instance of *MD_Format* (ISO 19115:2003)

A.6: Properties of CW_CoveragePeriod Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
firstYear	The first Gregorian calendar year to which an observation result applies	Mandatory	1	XML Schema type: gYear
endYear	The final Gregorian calendar year to which an observation result applies	Mandatory	1	XML Schema type: gYear
firstReliableYear	This only applies to the CRU station datasets. It represents the first Gregorian calendar year before which the observation values collected are deemed problematic, and hence are discarded.	Optional	1	XML Schema type: gYear

Annex B: Properties of the CRU Process Classes

B.1: Properties of the class CW_Process

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
processStep	A step associated with a process. A process may have one or more steps.	Mandatory	*	An instance of *CW_ProcessStep* (3.3.3.2)
Identifier	A unique identifier (applicable within the relevant domain) for the process.	Mandatory	1	An instance of *MD_Identifier (ISO 19115:2010)*

B.2 Properties of the class CW_ProcessStep

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
Input	Information about an input to a process step	Optional	*	An instance of *CW_ProcessIOEntity* (3.3.3.3)
Output	Information about an output produced by a process step	Optional	*	An instance of *CW_ProcessIOEntity* (3.3.3.3)
processingInformation	Information about the algorithm employed, the processor, e.g. software used to run the algorithm, details of person who invoked the software etc. associated	Optional	*	An instance of *LE_Processing* (ISO 19115-2:2009 [7])

	with a process step.			

B.3 Properties of the CW_ProcessIOEntity Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
observationResult	Result of a CRU observation represented as an input/output of a process step	Mandatory if the input/output is an observation result	1	An instance of *OM_DiscreteCoverageObservation* (ISO O&M)
stationStatistics	Special interim data results derived from CRU Station observation data for producing CRU publishable datasets, i.e. they are typically interim inputs or outputs of a process step (only applies to CRUTEM3 datasets)	Mandatory if the input/output is of type "*CW_StationStatistics*"	1	An instance of *CW_StationStatistics* (3.3.4.3)
Value	Represents an interim input or output of a process step that is neither a observation result nor a CRU station related information.	Mandatory if the input/output is an interim resource	1	An instance of *Record* (ISO 19115-2:2009)
externalResource	Represents an external resource that is used as the input of a process	Mandatory if the input is an external	1	An instance of *CI_OnlineResource* (ISO 19115:2003)

	step.	resource.		

B.4 Properties of the class CW_ AcquisitionStep

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
platform	Information about the location (i.e. station) at which the observation was made.	Mandatory	1	An instance of *CW_Station* (3.3.4.1)
acquisitionDetails	Reference to additional information about the data acquisition process that is available on the web.	Optional	1	An instance of *CI_OnlineResource* (ISO 19115:2003)

Annex C: Properties of the CRU Station Classes

C.1: Properties of the CW_Station Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
Geometry (inherited from the parent class, *SF_SamplingPoint*)	Geographical location (e.g. coordinates, coordinate reference system, etc.) of a climate monitoring station.	Mandatory	1	An instance of *GM_Point* (ISO 19107:2003 Spatial Schema [8])
stationMetadata	Non-geospatial information about a climate monitoring station, such as station identifier and ownership information.	Mandatory	1	An instance of *CW_StationMetadata* (3.3.4.2)

C.2: Properties of the CW_StationMetadata Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
Country	Country of the station	Optional	1	Character String
Image	The reference to a web-accessible picture of the station	Optional	*	An instance of *CI_OnlineResource* (ISO 19115:2003)

C.3: Properties of the CW_StationStatistics Class

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Association/ Constraints
Type	e.g. Normals or Standard Deviation	Mandatory	1	A value from the <<CodeList>>*CW_StationMetaObservationPropertyType* (Figure 7)
Values	The calculated values, which are typically a sequence Real numbers	Mandatory	1	A sequence of Real
coveragePeriod	The time period to which the normals or standard deviation applies.	Mandatory	1	An instance of *CW_CoveragePeriod* class (3.2.3.7)
derivedFrom	A link back to the station observation from which the normals or standard deviations were derived.	Mandatory	1	Association with an instance of *CW_StationObservation* (3.3.2.4)

Annex D: Properties of the CRU Workflow Classes

D.1: Properties of the class CW_ObservationWorkflow

Name/Role Name	Definition	Obligation/ Condition	Maximum Occurrence	Data Type/Constraints
Identifier	A unique identifier for a CRU workflow	Mandatory	1	An instance of *RS_Identifier* (ISO 19115:2003)
Owner	Ownership information about a CRU workflow	Mandatory	1	An instance of *CI_OnlineResource* (ISO 19115:2003)
lastModified	The date and time for the last updates made to the workflow. Useful for keeping track of changes made.	Mandatory	1	An instance of *CI_Date* (ISO 19115:2003)
Description	A brief textual description of the workflow	Mandatory	1	Character String
Observation	The CRU observations that a workflow is composed of.	Mandatory	*	Associates an instance of the O&M *OM_Observation* or any of its subclasses.