Quality Control Use Case for CMIP5

Document History

V0.1	BNL 3/10/10	Initial draft
V0.2	BNL 15/10/10	Minor mods + Added new issues from Martina Stockhause (as comments)

Introduction

This document describes:

- What the CMIP5 quality control (qc) system is expected to achieve,
- the actors involved in creating and using quality control information,
- and how we expect them to interact with the system.

The CMIP5 qc system

The main aim of the CMIP5 qc system is to provide a high level of assurance that CMIP5 data is fit for purpose for scientific analysis. It is **not** designed to provide assurance of scientific quality (although it will support annotation which could include comments as to the presence or absence of specific scientific qualities in the data).

Experience suggests that much data deposited for use in the CMIP5 system may have unexpected glitches and errors which are as much associated with data processing as they are with the modelling process itself. To that end, the qc process is designed with three levels, the first two of which are applied to data and metadata separately (see below for definitions of these terms), the final one of which is applied jointly:

1. qc level 1: objective: to catch obvious errors in data formatting and some limit excedences. See LINK for a full description of qc-lev-1d for data and LINK qc-lev-1m for metadata.
2. qc level 2: subjective: to find obvious errors in content (hemispheric flipping, glitches in timeseries of major variables etc, meaningless metadata etc). See LINK for a full description of qc-lev-2d for data and qc-lev-2m for metadata.
3. qc level 3: combined check for consistency between internal file metadata and external metadata as well as some moderation of the quality of the level 2 checking which as gone on (and if there has been any extra qc information entered by third parties, a review of whether any of that is accurate and/or relevant to the longevity of the data). Data which has passed qc-lev-3 will be eligible for a DOI (see LINK).

Definitions:

- Data: NetCDF files which conform to the CMIP5 conventions
- Metadata: Information entered via the metafor questionnaire (which is persisted as XML files when published from the questionnaire).
- Dataset: CMIP5 publication units are “realm datasets” which consist of “atomic datasets” as defined in the DRS document.
- Collection: an ESG gateway concept, used to organise datasets into common access control domains.

Note that there is internal metadata within the NetCDF files, and inherent in the layout of the files on disk (which is expected to conform to the DRS conventions, see LINK). One of the roles of qc-level-3 is to ensure consistency (where necessary) between that “internal” metadata and the other

(metafor) “external” metadata.

It is expected that the qc levels should impact on data availability. One option for how qc and data availability could interact is depicted in Figure 1 which is designed to indicate the flow of logic associated with qc and access control, not to indicate the physical flow of data.

However, it is recognised that it might be necessary to omit the left-hand branch of this diagram (at least initially) and provide NO access to data via ESG until data has completely passed qc-lev-2m. (This would be necessary if we believe the software implications or the operational implications – at data nodes and/or gateways and/or PCMDI – are too onerous).

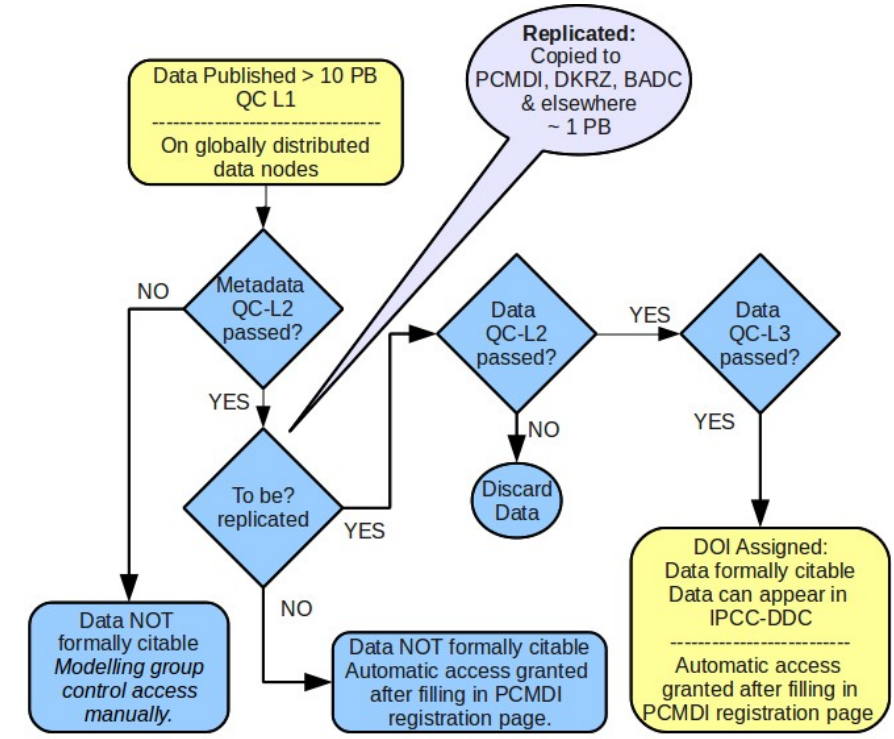

Figure 1: One possible logical relationship between acesss control and qc state (others are discussed in the text since it is possible that this configuration will be too difficult to deliver in a timely manner with existing data centre stafff and/or software development schedules.

Actors

(This section and figures need revising to make DataOwner correspond to an actual IP owner, and to use DataProvider for the ESG data node

People

a) DataOwner: person responsible for managing data on an ESG data node.
b) MetadataOwner: person responsible for managing metadata entry into the Metafor Questionnaire.
c) DataQC: person responsible for carrying out QC level-2-d (second level quality control for data) *for a specific dataset (or set of datasets).*
d) MetadataQC: person responsible for carrying out QC level-2-m (second level quality control for metadata) *for a specific record (or set of records).*
e) DOI-Publisher: person responsible for QC level 3 *for a specific dataset.*
f) User: Anyone interacting with the system, both as a consumer of quality control, and potentially as a person who might make quality control comments for entry into the system.

g) ESGF-DataOwner: person responsible for specific dataset visibility in the ESG system and ensuring the right access control is in place.
h) Replicator: person responsible for moving data
i) PCMDI: person with overall responsibility for access control for the CMIP5 system.

Key system components:

1. ESG data node (multiple instances), including:
    - ESG publisher
    - Replication Tool
2. ESG gateway (multiple instances), including:
    - Metafor Document Ingest System for use in
        - Metadata display (“trackback” pages), and
        - Quality control use in Access Control, and
        - QC display on
            - data pages, and
            - metadata (trackback) pages.
3. QC Entry Tool
4. Metafor Metadata Questionnaire

Interactions

Key Use Cases

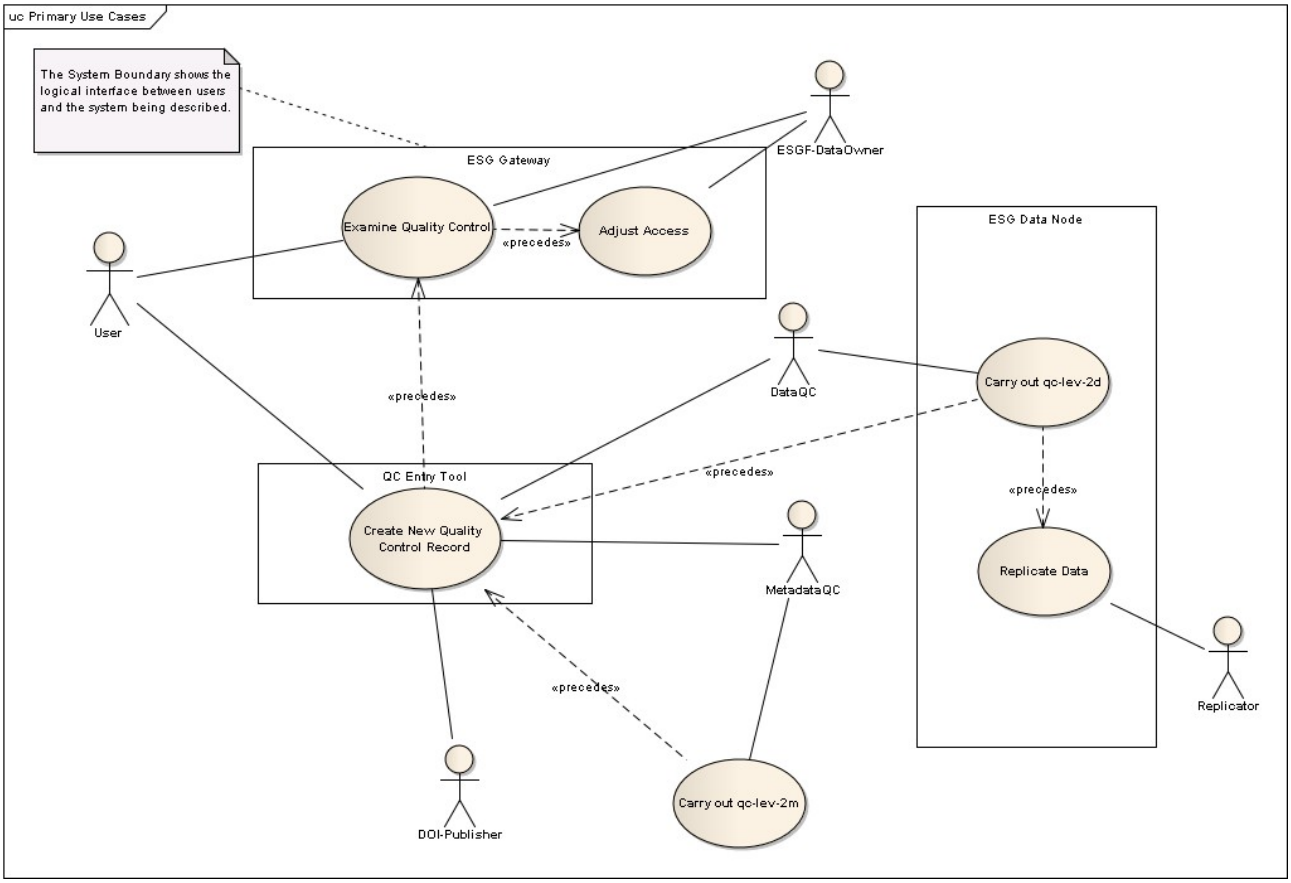

Figure 2: summary of key use cases

Create QC Records

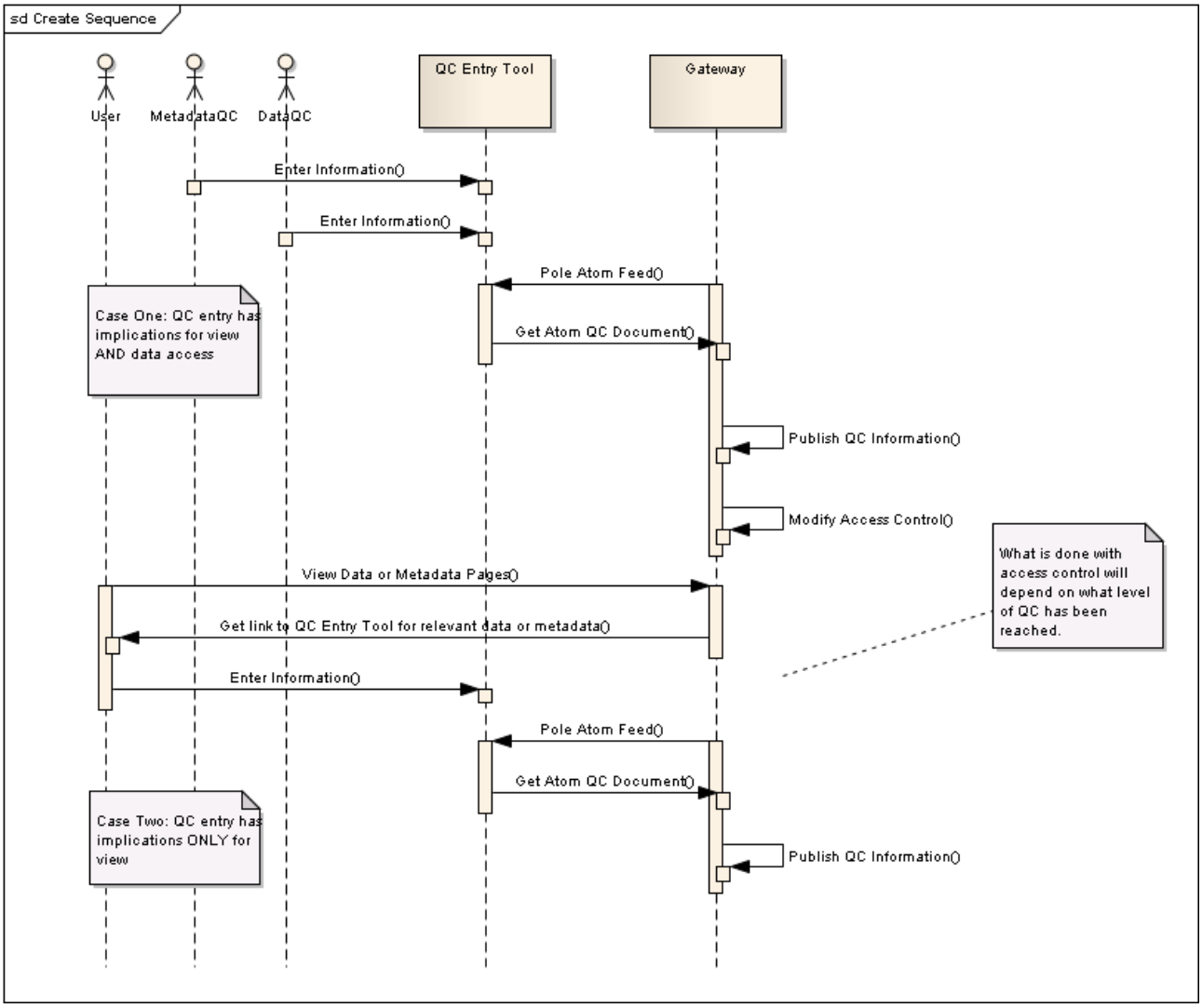

Figure 3: Sequene of events from qc-entry to gateway responses

Issue:

- How do we expect other gateway instances to interact with the qc tool? Directly or via the gateway-to-gateway interactions?

Use Case Story A: problem with data

(need to explain how a qc problem might be found in an atomic dataset, what might happen with the record for the realm dataset, and the consequential acts in ESGF)

Use Case Story B: qc level 3 activities.

The metadata of the finished QC level 3 is more than quality, i.e. more than a CIM quality document. Summary, contact and citations may be changed by the author during STD-DOI publication.

Replication

Assumptions:

- Replication occurs out of band from publishing. That is, datasets being replicated from A to B do not have to be (ESG) published until B wishes to do so.
- qc-lev-2d is needed before data should be replicated from a core data centre.

Implications for Access Control

Assumptions

- data published by the ESG datanode is published to a collection on an ESG gateway
- different collections can have differing access control

Requirements.

- It would need to be possible to move a dataset from one collection to another if, either
  - data was made available with differing access conditions between qc-states (e.g as in Figure 1), or
  - it was possible to ESG publish data before any qc-level-2m, in the event that qc-level-2m was required before ESG made data visible.

(Alternatively, one tries to stop the initial ESG publisher step. I don't see how you can stop someone attempting to publish data on their ESG datanode to a gateway unless the gateway can somehow know that the incoming dataset is expected and has metafor data ready or not. I suspect having that functionality in place might even be harder than achieving the above, or at least the same order of difficulty ...)

Then we need the ability either

- for a ESGF-dataowner to be able to manually make those access transitions after inspecting a quality control record, or
- for the ESG-gateway to parse incoming data documents and do it automatically.

ESG could launch with the former, and deliver the latter in a later version of software.

Appendix: QC in Metafor

This version includes some proposed but not yet accepted changes to the CIM.

(More detail available at http://metafortrac.badc.rl.ac.uk/trac/browser/CIM/branches/dev1.5/quality/quality.xsd)