7th Framework Programme
ENV.2010.4.1.2-2
Integrating new data visualisation approaches of earth Systems into GEOSS development

Project Nr: 265178

QUAlity aware VIualisation for the Global Earth Observation system of systems

Deliverable D6.1
Best practice document for quality encodings

Version 3.1

Due date of deliverable: 31/05/2012
Actual submission date: 15/08/2012
<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>D6.1 Best practice document on quality encodings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Creator</strong></td>
<td>LB_AST</td>
</tr>
<tr>
<td><strong>Editor</strong></td>
<td>LB_AST, ST_FRAUN and JM_CREAF</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Documentation for the use of the producer and user quality models and associated schema implementations</td>
</tr>
<tr>
<td><strong>Publisher</strong></td>
<td>GeoViQua Consortium</td>
</tr>
<tr>
<td><strong>Contributors</strong></td>
<td>GeoViQua Partners</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td>Text</td>
</tr>
<tr>
<td><strong>Format</strong></td>
<td>MS-Word</td>
</tr>
<tr>
<td><strong>Language</strong></td>
<td>EN-GB</td>
</tr>
<tr>
<td><strong>Creation date</strong></td>
<td>16/03/2012</td>
</tr>
<tr>
<td><strong>Version number</strong></td>
<td>3.1</td>
</tr>
<tr>
<td><strong>Version date</strong></td>
<td>15/08/2012</td>
</tr>
<tr>
<td><strong>Last modified by</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Rights</strong></td>
<td>Copyright © 2012, GeoViQua Consortium</td>
</tr>
<tr>
<td><strong>Dissemination level</strong></td>
<td>CO (confidential, only for members of the consortium)</td>
</tr>
<tr>
<td></td>
<td>X PU (public)</td>
</tr>
<tr>
<td></td>
<td>PP (restricted to other programme participants)</td>
</tr>
<tr>
<td></td>
<td>RE (restricted to a group specified by the consortium)</td>
</tr>
<tr>
<td>When restricted, access granted to:</td>
<td></td>
</tr>
<tr>
<td><strong>Nature</strong></td>
<td>X R (report)</td>
</tr>
<tr>
<td></td>
<td>P (prototype)</td>
</tr>
<tr>
<td></td>
<td>D (demonstrator)</td>
</tr>
<tr>
<td></td>
<td>O (other)</td>
</tr>
<tr>
<td><strong>Review status</strong></td>
<td>X Draft</td>
</tr>
<tr>
<td>Where applicable:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>WP leader accepted</td>
</tr>
<tr>
<td></td>
<td>Accepted by the PTB</td>
</tr>
<tr>
<td></td>
<td>PMB quality controlled</td>
</tr>
<tr>
<td></td>
<td>Accepted by the PTB as public document</td>
</tr>
<tr>
<td></td>
<td>Coordinator accepted</td>
</tr>
<tr>
<td><strong>Action requested</strong></td>
<td>to be revised by all GeoViQua partners</td>
</tr>
<tr>
<td></td>
<td>X for approval of the WP leader</td>
</tr>
<tr>
<td></td>
<td>X for approval of the PMB</td>
</tr>
<tr>
<td></td>
<td>X for approval of the Project Coordinator</td>
</tr>
<tr>
<td></td>
<td>X for approval of the PTB</td>
</tr>
<tr>
<td><strong>Requested deadline</strong></td>
<td>31/5/2012</td>
</tr>
<tr>
<td>Version</td>
<td>Date</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>0.0</td>
<td>16-03-2012</td>
</tr>
<tr>
<td>1.0</td>
<td>21-03-2012</td>
</tr>
<tr>
<td>2.0</td>
<td>03-05-2012</td>
</tr>
<tr>
<td>2.1</td>
<td>18-05-2012</td>
</tr>
<tr>
<td>2.2</td>
<td>28-05-2012</td>
</tr>
<tr>
<td>2.3</td>
<td>03-06-2012</td>
</tr>
<tr>
<td>2.4</td>
<td>11-06-2012</td>
</tr>
<tr>
<td>2.5</td>
<td>26-06-2012</td>
</tr>
<tr>
<td>2.6</td>
<td>28-06-2012</td>
</tr>
<tr>
<td>3.0</td>
<td>07-08-2012</td>
</tr>
<tr>
<td>3.1</td>
<td>15-08-2012</td>
</tr>
</tbody>
</table>

Copyright © 2012, GeoViQua Consortium

The GeoViQua Consortium grants third parties the right to use and distribute all or parts of this document, provided that the GeoViQua project and the document are properly referenced. THIS DOCUMENT IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS DOCUMENT, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
Table of Contents

1. Introduction ........................................................................................................................................... 7
   1.1 Introduction to the producer quality model ........................................................................... 7
   1.2 Introduction to the user quality model ............................................................................... 2
   1.3 Accessing the UML definitions for these conceptual models ........................................ 3

2. Relevant ISO standards referenced and reused .................................................................................. 4
   2.1 19115 / 19139 – Metadata ........................................................................................................ 4
   2.2 19157 – Data Quality ............................................................................................................. 6

3. Design principles adopted ................................................................................................................. 8

4. Producer quality model ....................................................................................................................... 11
   4.1 Publications .......................................................................................................................... 11
   4.2 Traceability ........................................................................................................................... 14
   4.3 Reference datasets used for evaluation ............................................................................... 15
   4.4 Producer soft knowledge: Discovered issues .................................................................... 17
   4.5 Populating DQ_Result elements using UncertML ............................................................ 19
      4.5.1 Quantitative Accuracy .................................................................................................. 20
      4.5.2 Thematic accuracy ........................................................................................................ 22

5. User quality model ............................................................................................................................... 24
   5.1 Classes of the conceptual model ............................................................................................ 24
      5.1.1 Feedback foci and qualifiers ....................................................................................... 25
      5.1.2 Feedback target roles .................................................................................................. 27
   5.2 Tasks .......................................................................................................................................... 28
      5.2.1 Encoding feedback catalogue answers ........................................................................ 28
      5.2.2 Submitting feedback ..................................................................................................... 29
   5.3 Opportunities in the implementation of the user / feedback model ......................................... 29
      5.3.1 New elements ................................................................................................................ 29
         5.3.1.1 Require domain qualifier ..................................................................................... 29
      5.3.2 Default value recommendations for the user /feedback model .................................. 29
         5.3.2.1 Domain ................................................................................................................... 29
         5.3.2.2 Role ......................................................................................................................... 29
   5.4 Application domain ontology ..................................................................................................... 30

6. Using the GeoVIQua schemas ........................................................................................................... 30
   6.1 Dataset and unique resource identifiers ................................................................................... 31
      6.1.1 Recommended practice for recording dataset identifiers ........................................... 31
      6.1.2 Identifiers - uniqueness and canonicalization ............................................................... 32
      6.1.3 Identifiers - Granularity ................................................................................................. 32
   6.2 Relationship of the GeoVIQua information model to the QIAP protocol ............................. 32
      6.2.1 QIAP: a protocol which uses metadata quality information ....................................... 32
6.2.2 QIAP concepts versus GeoViQua user and producer model ........................................ 33
6.2.3 User model ‘discovered issue’ feedback vs QIAP ‘quality event’ report ............................. 34
6.2.4 Producer model ‘discovered issue’ statement vs QIAP ‘quality issue’ report ............... 35
6.2.4.1 Embed ‘discovered issues’ within data quality elements .................................. 36
6.2.4.2 Exploit the user model to supply information on discovered issues .................. 36

6.3 Use cases: producer quality model .............................................................................. 36
6.3.1 Citing a publication within a metadata document .................................................. 37
6.3.2 Producer supplies ‘soft knowledge’ or advice on issues discovered with the dataset which cannot be easily encoded elsewhere. .................................................. 37
6.3.3 Recording the traceability of a quality statement .................................................. 38
6.3.4 Citing one or more datasets used as reference for a quality evaluation .................. 38
6.3.5 Providing full statistical information on the results of quality assessments ............. 39

6.4 Use cases: user model .................................................................................................. 41
6.4.1 A rating based on a dataset as a whole ................................................................... 41
6.4.2 A rating based on a spatial or temporal subset of the data ..................................... 41
6.4.3 A domain-specific rating in the context of weather forecasting ............................... 41
6.4.4 Justification for a rating .......................................................................................... 41
6.4.5 Provide general feedback on a complete dataset .................................................. 42
6.4.6 Provide general feedback on a part of the data ...................................................... 42
6.4.7 Provide general feedback on a metadata record ..................................................... 42
6.4.8 Provide general feedback on a specific metadata record elements ....................... 42
6.4.9 Add a domain-specific comment to a data set ...................................................... 42
6.4.10 Start a thematic discussion on a dataset ............................................................... 43
6.4.11 Report external feedback related to a dataset ..................................................... 43
6.4.12 Report external feedback related to a part of a dataset ...................................... 43
6.4.13 Report publication which cites a dataset ............................................................ 43
6.4.14 Add primary target ............................................................................................. 43
6.4.15 Add secondary target ......................................................................................... 43
6.4.16 Add supplementary target .................................................................................. 44
6.4.17 Search for comments pertaining to a domain ...................................................... 44

7. Full User stories ............................................................................................................. 44

7.1 Supplying a comment on a problem with a dataset .................................................... 44
7.2 Adding a secondary target .......................................................................................... 45
7.3 Adding a supplementary target .................................................................................. 46
7.4 Citing a publication which uses specific data ........................................................... 46
7.5 Reporting intercomparison of datasets, and fitness-for-purpose assessment ............ 47
7.6 Constructing a ‘GEOLabel’ for a dataset ................................................................... 48
7.7 Providing general feedback on a metadata record .................................................... 49
7.8 Providing general feedback on a specific metadata record elements ....................... 51
7.9 Searching for a domain-specific rating on a resource .............................................. 51
7.10 Searching for information on problems with a dataset for a specific space/time window ........................................................................................................... 52
7.11 Searching for individual products which are cited in more than 15 unique journal articles ........................................................................................................ 52
7.12 Finding datasets which meet user-specified quality thresholds ............................ 52

Appendix 1: example metadata document based on the GLC 2000 dataset ..................... 54
Figures

Figure 1: The GeoViQua producer quality model ................................................................. 7
Figure 2: The GeoViQua user quality model .................................................................. 2
Figure 3: The new version of MD_Identifier (Source: https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Identifiers), with highlighting of new attributes added .................................................................. 5
Figure 4: MD_AssociatedResource .................................................................................. 5
Figure 5: New components of the DQ_Element in 19157 .................................................. 6
Figure 6: ISO 19157 mechanism for standalone reporting .................................................. 7
Figure 7: ISO 19157 'metaquality' additions .................................................................. 7
Figure 8: Relationship between the type and the nillable 'PropertyType' in OGC schemas ...... 10
Figure 9: Locations where a publication can or must be supplied in the GeoViQua producer model ................................................................. 11
Figure 10: The GVQ_Publication class, which extends CI_Citation .................................. 12
Figure 11: A publication element in a metadata document .................................................. 13
Figure 12: The 'Traceability' element ........................................................................ 14
Figure 13: How metaquality is linked to quality in ISO 19157/GeoViQua ..................... 15
Figure 14: The new GeoViQua elements which permit reference datasets to be recorded .... 16
Figure 15: An evaluation element with a reference dataset .................................................. 16
Figure 16: GVQ_DiscoveredIssue type ....................................................................... 17
Figure 17: A 'discoveredissue' element within a data quality item .................................. 18
Figure 18: The existing 19157 schema elements relating to quality measurement reporting ................................................................. 19
Figure 19: Concrete types available for encoding a DQ_Result ........................................ 20
Figure 20: A FeedbackCollection contains any number of FeedbackItems, which have targets referencing GEOSS resources ........................................... 25
Figure 21: Data-centric feedback focus refines the section of a target to which the feedback applies ................................................................. 26
Figure 22: The domain, tags and reply-to qualifiers on a feedback item ............................. 27
Figure 23: The 'subject' qualifier of a feedback item ......................................................... 27
Figure 24: Primary, secondary and supplementary targets on a feedback item ................ 28
Figure 25: Components of a QIAP Quality Event Report .................................................. 34
Figure 26: Components of a QIAP Quality Issue Report .................................................... 35
Figure 27: Components of a QIAP Quality Issue Report .................................................... 35
Figure 28: Publication information from a GeoViQua producer metadata document .......... 37
Figure 29: Producer 'soft knowledge' from a GeoViQua producer metadata document .......... 37
Figure 30: Traceability information from a GeoViQua producer metadata document .......... 38
Figure 31: Reference dataset information from a GeoViQua producer metadata document ................................................................. 39
Figure 32: An UncertML ConfusionMatrix from a GeoViQua producer metadata document in HTML rendering ....................................................... 40
Figure 33: A screen shot of the NOAA / GeoViQua completeness evaluation rubric for metadata ................................................................. 50
1. Introduction
This document describes the GeoViQua data encodings for producer and user quality information. We explain the relevance of new schema elements and give examples of good practice for their use.

1.1 Introduction to the producer quality model

The producer quality model is shown in Figure 1, and is more fully discussed in section 4.
The producer quality model introduces elements to record qualitative and quantitative quality information, and to identify resources (i.e., datasets) in order to relate metadata in hierarchical or other ways. The model extends ISO 19115\textsuperscript{1}, 19115-2\textsuperscript{2} and 19157\textsuperscript{3}, adding means to report publications, discovered issues, reference datasets used for quality evaluation, traceability, and statistical summaries of quantified uncertainty.

1.2 Introduction to the user quality model

The user quality model is shown in Figure 2. It re-uses some ISO quality and metadata elements, and elements of the producer model, but is far less strictly bound to existing ISO schemas.

![Figure 2: The GeoViQua user quality model.](http://schemas.geoviqua.org/GVQ/3.1.0/UML/Feedbackmodel.jpg)

The root element of the schema is the FeedbackCollection, which can hold zero or more FeedbackItems. An empty FeedbackCollection indicates, for example, a search for feedback where no relevant data was found.

A ‘FeedbackItem’ may 4 'own' or 'hold' for example any combination of:

- (optional) rating;
- one or more user comments;
- one or more reports of usage (including reports of any discovered issues);
- citation of one or more publications;
- a text description of the feedback subject;
- text tags which might assist with topic-based search and linking;
- one or more quality overrides.

It must ‘own’:
- one or more sets of details on the focus of the feedback;
- a reference to at least one target of the feedback (a dataset, resource, etc.);
- one user information describing the submitter of the feedback.

These components of user feedback are discussed in more detail in section 5.

1.3 Accessing the UML definitions for these conceptual models.

Both the producer and user quality model have been formulated as UML diagrams which, as well as being viewable as online images (see above) may also be accessed using Enterprise Architect5. A file containing both models, as well as all supporting schemas, may be downloaded from the following URL.

http://schemas.geoviqua.org/GVQ/3.1.0/UML/ISO_19115-1-2_and_19157_Metadata_and_GeoViQua_v3_1.EAP

This UML defines all cardinality, extensions and relationships to existing ISO standards.

---

4 The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 (http://www.apps.ietf.org/rfc/rfc2119.html)

5 A free version of Enterprise Architect with reading capabilities can be downloaded from http://www.sparxsystems.com/uml_tool_guide/enterprise_architect_modeling_tool/ea_lite.htm
2. Relevant ISO standards referenced and reused.

The producer model in particular is designed to extend the existing practices of metadata generation, and thus wherever possible, widely-accepted ISO elements used by INSPIRE and national institutions are employed. All of the schema documents discussed here can be accessed at http://schemas.geoviqua.org/

2.1 19115 / 19139 – Metadata

The 19115 metadata standard is extended in order to exploit existing and accepted conceptual models for citation, instrumentation, spatial and temporal extent, contact details, spatial data representations and other elements of the geospatial domain of discourse.

The schema documents referenced for most metadata elements other than data quality are the 19139 schema implementations of the 19115 specification. The 2007 versions hosted at schemas.opengis.net may be used, and the 2007 version is bundled with the schemas supplied as part of deliverable 6.1, for local reference.

However, there are several updates to the 19115 specification which have not yet been implemented in the available 19139 schema documents. These have therefore been added within the GeoViQua project as part of a separately-referenced schema document, in the anticipation that these updates will ultimately be incorporated into the published schema implementations.

The specific updates concerned are as follows, and can be accessed at http://schemas.geoviqua.org/ISO/19139/20120707_GVQ/19115_updates.xsd

1. ISO 19115 includes a MD_Identifier object for associating identifiers with various objects in the metadata record\(^6\). The RS_Identifier extension of MD_Identifier (also part of ISO 19115) included several important additions (see Figure 3) because of the increased importance of namespaces in the XML and Web environments and because of ambiguity in how the namespace is actually defined in the CI_Citation associated with a MD_Identifier. These

additions are to be incorporated into the `MD_Identifier`, but are not yet available in published XML Schema documents (XSDs). The importance and correct use of unique dataset identifiers is further discussed in section 6.1.

```xml
<<DataType>>
MD_Identifier
+ authority [0..1] : CI_Citation
+ code : CharacterString
+ codeSpace [0..1] : CharacterString
+ version [0..1] : CharacterString
+ description [0..1] : CharacterString
```

Figure 3: The new version of `MD_Identifier` (Source: https://geo-ide.noaa.gov/wiki/index.php?title=ISO_Identifiers), with highlighting of new attributes added.

2. The `MD_AssociatedResource` element of 19115 is ideal for giving details of any reference dataset which has been used as the benchmark for a quality assessment. It allows the dataset to be identified using an identifier, and the means in which it was used to be described using ISO codes (see Figure 4). This is also unavailable in published schema implementations and has been locally implemented. The use of the element is demonstrated in Sections 4.3 and 7.5.

Some of the modifications or extension to ISO19115 has been submitted to the ISO19115-1 Metadata – Fundamentals (the new revised version of the Metadata that is under the review process. The modifications submitted and results of this interaction have been explained in the GeoViQua deliverable D1.7.
2.2 19157 – Data Quality

This standard is not approved yet but is in the final stages of the process. For our development purposes, we gained access to the advanced drafts. Only small changes are expected at this stage. The harmonised 19157 standard for data quality information has introduced several changes:

- Separating attributes of the DQ_Element into three distinct objects: measure (optional), evaluation (optional) and result (required). This is illustrated in Figure 5, and is designed to solve problems with achieving appropriate repetition in the original schema.

![Figure 5: New components of the DQ_Element in 19157](image_url)

- Allowing a standalone report to be referenced from the DQ_DataQuality element, with a short abstract describing its content. This allows supplementary information outside the conceptual scope of the framework to be accessed. A standalone report is encoded as a new object, DQ_StandaloneReportInformation, and is referenced as a ‘report’ element of the DQ_DataQuality element, in the same manner that a normal metadata element would be referenced (see Figure 6). This allows linkage between ISO metadata and data quality information which does not fit the conceptual model of the available ISO schemas.
Introducing a metaquality element (see Figure 7) to record information about the representativity, homogeneity and confidence levels of the quality data itself. This allows users to assess the rigour and relevance of the assessments which generated the quality statement, and could thus be extremely important in assessing fitness-for-purpose.

Since 19157 has not yet been approved, the schema is not yet available in the form of published schema documents, so for practical implementation, as with the 19115 updates, a local implementation was produced. This is accessible at: http://schemas.geoviqua.org/ISO/19157/20120707_GVQ/19157_DataQuality.xsd

This follows the specification in all regards, with the expectation that in future, a reference to a published schema document may simply be swapped in its place.
3. Design principles adopted

This schema imports existing 19115 schema elements where appropriate, by referencing the 2007 implementation of ISO 19139, at
http://schemas.opengis.net/iso/19139/20070417/
Elements of 19115 are replaced by specific GeoViQua documents as follows:

Top-level schema

- GeoViQua_PQM_UQM.xsd
  Defines many of the elements of the producer and user quality models, including:
  - the GVQ_Metadata root element used by a producer to generate metadata,
  - the GVQ_FeedbackCollection root element which may be used to hold the results of a request to the feedback server.

The relationship between these two elements which allows them to be combined in a single document is explained in section 6.
All GeoViQua elements are encapsulated in the namespace

http://www.geoviqua.org/QualityInformationModel/3.1

Imported schemata

- ISO 19139 implementation of the 19115 standard

- 19157_DataQuality.xsd
  imported in place of the 19115 ‘dataQuality.xsd’ document (see section 2.2).

- 19115_updates.xsd
  imported to allow use of unavailable 19115 elements (see section 2.1).

- uncertml.xsd
  imported to allow use of UncertML elements to describe quantified uncertainty (see section 4.5)

- GeoViQua_DataQuality.xsd
  Contains all the remaining GeoViQua-specific elements for the producer model – often these consist of extensions to ISO 19157.
Throughout the new documents, the OGC approach of allowing a ‘nilReason’ attribute for missing element values has been adopted. In practice, this leads to elements looking as follows:

```xml
<gmd:CI_ResponsibleParty>
  <gmd:organisationName>
    <gco:CharacterString>European Petroleum Survey Group</gco:CharacterString>
  </gmd:organisationName>
  <gmd:contactInfo>
    <gmd:CI_Contact>
      <gmd:onlineResource>
        <gmd:CI_OnlineResource>
          <gmd:URL>http://www.epsg-registry.org</gmd:URL>
        </gmd:CI_OnlineResource>
      </gmd:onlineResource>
    </gmd:CI_Contact>
  </gmd:contactInfo>
  <gmd:nilReason>missing</gmd:nilReason>
</gmd:CI_ResponsibleParty>
```

OGC-recommended values for ‘nilReason’ are as follows:

- **inapplicable**
  - there is no value
- **missing**
  - the correct value is not readily available to the sender of this data. Furthermore, a correct value may not exist
- **template**
  - the value will be available later
- **unknown**
  - the correct value is not known to, and not computable by, the sender of this data. However, a correct value probably exists
- **withheld**
  - the value is not divulged

This approach may change depending on the ultimate implementation style of the 19157 data quality schema. Several possible approaches for nillable or voidable stereotypes exist, and it is likely that we will follow the approach recommended by INSPIRE. It should also be possible to relax application of the stereotype for elements which are truly optional.
The mechanism to allow this behaviour consists of the declaration of a wrapper ‘ElementName_PropertyType’ for each element type, containing a reference to the named element type, plus the gco:nilReason attribute. This ‘x_PropertyType’ is then referenced as the type for schema elements, as shown above in Figure 8.
4. Producer quality model

A full metadata document which conforms to the producer schema is shown in Appendix 1, and may also be viewed at
http://schemas.geoviqua.org/GVQ/3.1.0/example_documents/GLC_2000_GVQ_raw.xml

A version of the same document, styled using XSLT, may be viewed at

Screenshots from the styled version are used in section 6.3 to illustrate some of the benefits of the producer quality model.

4.1 Publications

Publications (e.g., journal articles, technical reports) may be added to a number of quality elements within the metadata document. In each case, an existing DQ_ or MD_ element is extended to allow a ‘referenceDoc’ element to be added (see Figure 9). The resulting new objects are GVQ_Lineage, GVQ_DataIdentification and GVQ_Usage. DQ_Evaluation already has a ‘referenceDoc’, and since GVQ_Publication is substitutable for CI_Citation, it may also be employed here.

![Figure 9: Locations where a publication can or must be supplied in the GeoViQua producer model.](image)

Each publication extends a CI_Citation record with some GeoViQua-specific elements, including codes describing the purpose and the medium of publication (see Figure 10). At
the stage when metadata is produced, these publications are likely to be related to the Cal/Val process or quality assurance procedures.

Figure 10: The GVQ_Publication class, which extends CI_Citation

Note that the GVQ_Publication has a ‘target’ which should uniquely identify the dataset to which this reference relates. This allows the formalised publication element to be reused independently if necessary, for example by users submitting references to distributed databases and catalogues after the release of a dataset and its associated metadata. The use of publication elements as citations in the feedback model is more fully discussed in section 5.

An example of a publication element can be viewed online at http://schemas.geoviqua.org/GVQ/3.1.0/example_documents/Publication_example.xml

A compressed summary is shown in Figure 11.

Unique identifiers for a publication, such as ISBN, ISSN and DOI (digital object identifier) may be encoded using character strings. The ISSN and ISBN derive from the original 19115 standard, but the DOI, which can be very useful for retrieving online resources, has been added in the GeoViQua conceptual model. The ways in which the DOI may be practically used to retrieve resources are discussed later in this section.

Another addition to the existing citation is a GVQ_PublicationCategoryCode, which allows the nature of a wide variety of resources to be recorded, including traditional reports and journals as well as wikis and online videos. In cases where the publication is available online, its URL can be recorded as part of a CI_OnlineResource element.

Finally, the GVQ_PublicationPurposeCode allows specification of the relevance of a publication to the resource on which it is being cited. This allows us to specify, for example, whether the publication documents evaluation, derivation, description or comparison of the dataset. This code may be used particularly in the user context to search for publications which give a particular insight into a dataset.
Figure 11: A publication element in a metadata document.

In the example above, a DOI is supplied as a simple character string which may be parsed and used by a client to retrieve the resource using their own chosen DOI resolver. In the ‘onlineResource’ element is a direct URL link to the published document at the journal homepage. This makes it very easy for a client to pull together online resources which can be of great value in actually using the metadata.

If the metadata producer wishes to recommend that a specific DOI resolver should be used to retrieve the publication, they could instead specify the combination of resolver URL and DOI as the content of the ‘onlineResource’, as follows:

http://dx.doi.org/10.1109/TGRS.2006.864370

which resolves to:

http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1645273
We recommend that if this strategy is used, the DOI string should still be supplied in a ‘doi’ element, to allow the use of alternative resolvers if necessary.

To do this work we took into consideration ISO 690, particularly GVQ_PublicationCategoryCode contains all categories specified in this standard,"

4.2 Traceability

A fourth element has been added to the ‘metaquality’ concrete types described in section 2.2, to allow the lineage of a data quality assessment to be recorded, along with its representativity and coverage. This element contains a full LI_Lineage or GVQ_Lineage element as its ‘trace’ component, allowing process steps to be recorded for the sampling and evaluation involved in generating the quality statement (see Figure 12). If the substitutable GVQ_Lineage is used instead of LI_Lineage, then one or more reference documents may also be cited in support of the traceability statement.

![Figure 12: The ‘Traceability’ element.](image)

Metaquality elements are created as reports belonging to the DQ_DataQuality element of the metadata document. They are therefore found at the same level as the reports which they reference using a ‘relatedElement’ attribute. A compressed example is
shown below in Figure 13. This example uses XML Linking Language (xlink)\(^7\) to reference the id of the relevant data quality report.

\[\text{Figure 13: How metaoquality is linked to quality in ISO 19157/GeoViQua.}\]

4.3 Reference datasets used for evaluation

An important element of data quality assurance is the verification of selected values against an accepted independent calibration or validation dataset. Currently, however, the ISO metadata standards allow no systematic way to record the identity of this data, other than describing its origin as part of a free text entry or a non-standard document describing the cal/val process. We have therefore extended the ‘dataEvaluation’ section of the 19157 schema to allow any data evaluation record to report the identity of the reference dataset, and the way in which it was used.

Figure 14 shows how the extension was practically achieved: an initial attempt to extend DQ_DataEvaluation directly caused problems with substitutability in older schemas, and so each concrete type was extended individually. An example of one of these extended elements is shown in Figure 15. The publication element is not shown in full, but may be inspected in full online at


The citation element is compressed to save space, but, as can be seen from the namespace prefix, it uses the updated MD_Identifier in which a codespace may be specified.

\(^7\) [http://www.w3.org/TR/xlink11/](http://www.w3.org/TR/xlink11/)
Figure 14: The new GeoViQua elements which permit reference datasets to be recorded.

Figure 15: An evaluation element with a reference dataset.
4.4 Producer soft knowledge: Discovered issues

As well as legal and security constraints, it is now possible to add one or more discovered issue (e.g., a problem which the producer has identified during generation of a dataset) to an extended version of the DQ_DataQuality element of a metadata document (Figure 16).

![Diagram of GVQ_DiscoveredIssue type]

The GVQ_DiscoveredIssue type is a standalone class which offers the option of supplying a reference to a corrected dataset, suggestions for workarounds, alternative datasets, and a free text description of the identified problem. It also allows reference to be made to expected dates when a problem will be fixed, and to the location of datasets in which the problem has been fixed.

Note that the GVQ_DiscoveredIssue contains a required ‘target’ element which uniquely identifies the resource to which it refers. Fixed resources and alternative datasets are encoded more fully using MD_DataIdentification elements, which contain contact information as well as unique resource identifiers. The recommended use of MD_Identifier elements is discussed in section 6.1.

When discovered issues are embedded within a metadata document in this way, the ‘target’ identifier becomes somewhat redundant since, according to good practice, that identifier should already be recorded elsewhere in the document (for example, in CI_Citation > identifier). However, this approach aims to make the GVQ_DiscoveredIssue element re-usable within the user quality model, in the form of isolated records in databases and catalogues which record user feedback and expert judgements. The possible redundancy within a producer quality document can be addressed by using xlink for internal document referencing, as shown in the example data quality element containing a discovered issue (see Figure 17). As with all the other examples in this section, this may be more fully viewed in Appendix 1, as part of an example metadata document, or online at [http://schemas.geoviqua.org/GVQ/3.1.0/example_documents/GLC_2000_GVQ_raw.xml](http://schemas.geoviqua.org/GVQ/3.1.0/example_documents/GLC_2000_GVQ_raw.xml)
Figure 17: A ‘discoveredIssue’ element within a data quality item.
4.5 Populating DQ_Result elements using UncertML

A data quality statement which conforms to 19115/19139 or the proposed 19157 standard contains one or more ‘report’ elements where numerical information such as quantified accuracy may be recorded. These all extend the abstract DQ_Element, which can contain one or more ‘result’ DQ_Result elements (as shown in Figure 18).

![Diagram](image.png)

**Figure 18:** The existing 19157 schema elements relating to quality measurement reporting.

The current and proposed schemas do not restrict the values that can be used as the results of these elements, and so for the proposed GeoViQua updates, no changes are visible on the UML diagram. However, in each of these locations, UncertML may be used to provide more information about the nature of the values supplied, by offering encodings for probabilistic statistical information.

This fulfils the interoperability requirements identified by a number of geospatial data users, including the National System for Geospatial Intelligence Metadata Foundation (NMF). Section 7.5 of their document on quality metadata specifically notes that the

---

flexibility of the conceptual **Record** type, which may contain an indexed list of highly heterogeneous data structures and types, is actually a problem for effective exchange of quantitative data. Their recommendation is that strictly constrained data types should be used in order to permit the expression of semantics which allow the automated generation and use of quantitative data. In the referenced document, SWE Common entities are proposed as constrained categories for the encoding of quantitative information within an ISO 19115 **Record** element. The semantic information which conveys the nature of the Record (e.g., ‘percentile’, ‘standard deviation’ and ‘covariance’ may then be retrieved from publicly available registries which define specific quantitative measure types, such as the OGC online registry, NASA’s Semantic Web for Earth and Environmental Terminology (SWEET) ontology, or the Marine Metadata Interoperability (MMI) registry. Thus one may enrich a SWE Common **Quantity** with a reference to a URN which defines the meaning of that value; e.g., [http://sweet.jpl.nasa.gov/2.0/mathStatistics.owl#Percentile](http://sweet.jpl.nasa.gov/2.0/mathStatistics.owl#Percentile).

UncertML, as another publicly-available dictionary of statistical concepts and measures, has the potential to be used in exactly the same way, with the added benefit that it already defines XML encodings (as well as JSON and netCDF representations) for those concepts and measures, allowing quantitative data to be consistently represented, validated and automatically parsed from traditional XML metadata documents with no change to the underlying ISO schema. We demonstrate several examples of how this may be achieved in the following sections.

### 4.5.1 Quantitative Accuracy

The 19157 standard retains the ‘**DQ_Result**’ components of a data quality element, within which are embedded the actual results of that specific quality assessment (see Figure 19)

![Diagram](image)  
**Figure 19:** Concrete types available for encoding a **DQ_Result**.
The element within which UncertML may be used is the DQ_QuantitativeResult. For instance, consider a traditional DQ_QuantitativeResult which is the ‘result’ of a DQ_AbsoluteExternalPositionalAccuracy element (defined in ISO 19115 and retained in 19157). The element conveys the vertical accuracy of a DEM, as follows.

```
<gmd:result>
  <gmd:DQ_QuantitativeResult>
    <gmd:valueType>
      <gco:RecordType xlink:href="http://somesite.org/vert_acc">Value for vertical DEM accuracy</gco:RecordType>
    </gmd:valueType>
    <gmd:valueUnit xlink:href="urn:ogc:def:uom:OGC:1.0:metre"/>
    <gmd:value>
      <gco:Record>3.5</gco:Record>
    </gmd:value>
  </gmd:DQ_QuantitativeResult>
</gmd:result>
```

This description is fairly rich in information, but is deficient in the statistical information we can obtain about the uncertainty itself. Was the spread of error at the sampled locations genuinely Gaussian? Was there any bias in the error? The ‘3.5’ value here probably represents two standard deviations around a Gaussian mean, but even this information is not apparent from the above example.

There are two straightforward ways in which UncertML may be used to provide more information. The first, and the least invasive, is to supply the UncertML dictionary URI as the valueType, but leave the value as a simple number, as follows: Assuming that the errors from which the RMSE was calculated have a mean of 0 (i.e., no bias), this gives the same information as the first example. However, this time we know that the value given is a variance, (and, assuming that the original value of 3.5 did indeed represent 2 standard deviations, then this new value will be 3.0625)

```
<gmd:result>
  <gmd:DQ_QuantitativeResult>
    <gmd:valueType>
      <gco:RecordType xlink:href="http://www.uncertml.org/statistics/variance">Value for vertical DEM accuracy</gco:RecordType>
    </gmd:valueType>
    <gmd:valueUnit xlink:href="urn:ogc:def:uom:OGC:1.0:metre"/>
    <gmd:value>
      <gco:Record>3.0625</gco:Record>
    </gmd:value>
  </gmd:DQ_QuantitativeResult>
</gmd:result>
```
A more informative approach, but one which requires a client to be able to parse UncertML, is to use the Uncertainty element as the ‘value’ of the element, and to supply the UncertML dictionary URI as the ‘valueType’, as follows:

```
<gmd:result>
  <gmd:DQ_QuantitativeResult>
    <gmd:valueType>
      <gco:RecordType xlink:href="http://www.uncertml.org/distributions/normal"> Value for vertical DEM accuracy </gco:RecordType>
    </gmd:valueType>
    <gmd:valueUnit xlink:href="urn:ogc:def:uom:OGC:1.0:metre"/>
    <gmd:value>
      <gco:Record>
        <un:NormalDistribution>
          <un:mean>0</un:mean>
          <un:variance>3.0625</un:variance>
        </un:NormalDistribution>
      </gco:Record>
    </gmd:value>
  </gmd:DQ_QuantitativeResult>
</gmd:result>
```

In the above example, the distribution of values from which the vertical accuracy was calculated is given in the form of a NormalDistribution element. This information is exactly the same as that in the original example, but this time there is an explicit recognition that the samples were considered to acceptably fit a Normal distribution with mean 0 (i.e., no bias was observed) and a publicly-accessible dictionary definition of the distribution is referenced.

UncertML allows more detail still to be encoded where it is available. If the sampled error values did not fit a Normal distribution, then the user may choose an alternative to encode the information, and to convey any bias or skewness in the error. Alternatively, the user may encode an ExceedanceProbability or CredibleInterval element which tells us that, based on the evidence, there is a 99% chance that a recorded elevation falls within a certain range. For even more detail, UncertML can record the variability of the samples without oversimplifying their distribution – either by encoding a Histogram or set of Quantiles which summarises the sample values, or by encoding the raw sample values as a SystematicSample element.

### 4.5.2 Thematic accuracy

The usual means of recording thematic accuracy (e.g., categorical labels for a dataset of vector polygons or classified raster pixels) is through a confusion matrix with an associated Kappa statistic. Historically, this information has been recorded in free text or through
hyperlinks to external documents recording the matrix values. This type of information must by definition be read by a human, so no numerical information can be automatically extracted. The UncertML standard defines a ConfusionMatrix encoding element which means that a client capable of parsing UncertML could extract numerical information instead. This potentially allows the use of those confusion counts to generate likelihoods for specific misclassifications, and carry out modelling and simulation in automated workflows.

```xml
<gmd:result>
  <gmd:DQ_QuantitativeResult>
    <gmd:valueType>
      <gco:RecordType xlink:href="http://www.uncertml.org/statistics/confusion-matrix">Confusion matrix recording counts from ground truthing</gco:RecordType>
    </gmd:valueType>
    <gmd:valueUnit xlink:href="urn:ogc:def:uom:OGC:1.0:metre"/>
    <gmd:value>
      <gco:Record>
        <un:ConfusionMatrix xmlns:un="http://www.uncertml.org/2.0">
          <un:sourceCategories>
            Water Forest Wheat
          </un:sourceCategories>
          <un:targetCategories>
            Water Forest Wheat
          </un:targetCategories>
          <un:counts>42 2 3 4 52 1 7 8 29</un:counts>
        </un:ConfusionMatrix>
      </gco:Record>
    </gmd:value>
  </gmd:DQ_QuantitativeResult>
</gmd:result>

A confusion matrix within a ThematicClassificationCorrectness element (defined in ISO 19115 and retained in 19157) might look like the example above.
5. User quality model

The user model is generally intended for user feedback, but also for producer feedback that is changing more rapidly than official metadata for which more quality control is implied.

The model is intended to connect GEOSS datasets and services (referred to as resources) to user generated data. System-centric information, i.e. information about users such as accounts and authentication or GEOSS resources are therefore not represented directly, but are linked in at pre-defined junction points. These junction points are the user information and the target. A real implementation is expected to enrich these elements as appropriate, but not necessarily in a directly observable way.

While the model sometimes refers to ISO191xx schema elements, it is not intended to be seen as metadata or directly related to it. The ISO elements are included to ease interoperability in cases where a producer decides to include user-suggested data into the official metadata he or she maintains, or other cases where a common dataset is advantageous.

5.1 Classes of the conceptual model

The user model’s main elements are the:

GVQ_FeedbackCollection containing one or more:

- GVQ_FeedbackItems, individual feedback items containing:
  - Mandatory information on the role of the user recording the feedback (e.g., ‘Commercial data producer’)
  - GVQ_UserInformation, the user information. (This may optionally contain information about the various roles in which the user has interacted with the resource on which they are commenting).
  - Information qualifying the focus of the feedback, such as its subject, application domain, whether it is a reply to other feedback and keyword tags
  - References to one or more targets (GVQ_FeedbackTarget); uniquely identified GEOSS resources or arbitrary subsets of those resources, qualified by the role they play within the feedback item.
  - Other optional information such as a rating, comments, a quality override which may be considered as superseding producer quality information, a report of usage or a citation – a reference to a publication.
The user information contains self-asserted information about the user which may be used to (assist in) qualifying the feedback s/he produces. The model contains this class as a clear handover point to the implementation’s user management, and to specify what an implementation should know about its users.

The feedback item (GVQ_FeedbackItem) is an instance of user feedback; for example, a comment on a data set backed by a report. Each item is associated with one or more targets, which define the context of the discussion in the item. Further, an item can be qualified e.g. with a subject or an application domain in which it applies. These qualifications may be used to identify relevant feedback.

The target (GVQ_FeedbackTarget) points to a data set, resource, or something else which is pre-existent in the domain of discourse, and potentially also describes a sub-set of the resource to allow to narrow the discussion context as appropriate. The target as contained in the model is geared towards GEOSS resources.

These top-level elements of the model are illustrated in Figure 20.

Figure 20: A FeedbackCollection contains any number of FeedbackItems, which have targets referencing GEOSS resources.

5.1.1 Feedback foci and qualifiers

The feedback model relies on qualifiers and foci to narrow the focus of an item. This is intended to increase discoverability by enabling a user to submit and discover focused feedback. Some qualifiers are designed to be machine-comparable for this reason.

The different types of focus and qualifier are as follows:

- **Data-centric focus** (GVQ_DataFocus) This qualifies the target of the feedback. It defines spatial or temporal range over which the feedback applies, or any other specific focus (supplied as free text) which may include spectral bands, themes etc. The absence of a data-centric focus for a target implies that the feedback applies over the entire resource. The relevant part of the model is shown in Figure 21.
Figure 21: Data-centric feedback focus refines the section of a target to which the feedback applies.

- **Domain qualifier**
  A domain qualifier applies to the feedback item, and specifies the application domain or theme which generated this feedback (for example ‘hydrology’). A URI should be supplied which maps to an agreed ontology such as the GEMET\(^9\) thesaurus. A feedback item may have 0 to many domain foci (see Figure 22).

- **Feedback Response qualifier (reply-to)**
  This qualifies a feedback item as pertaining to other feedback, which should be specified by a unique identifier. This states the feedback to be an answer to some other feedback in the system (see Figure 22).

- **Tag qualifier**
  The tag qualifier associates tags, parts of a so-called *folksonomy*, to the feedback item (see Figure 22). Such free-form memes are often aggregated in “tag clouds”, allowing a community to reach an informal understanding of these tags. The precision of such features is debateable, but in social networks they are often considered essential.

\(^9\) http://www.eionet.europa.eu/gemet
Subject qualifier
This is just a subject. A subject is used to qualify feedback that may not be sufficiently qualified by other means. It just contains a title/subject text and can be regarded as a sort of free-form qualification of input (see Figure 23).

Figure 22: The domain, tags and reply-to qualifiers on a feedback item.

5.1.2 Feedback target roles
The targets of an item describe its context, guided by one of the three possible roles primary, secondary and supplementary. They define how a target is relevant to an item. Targets are intended to be comparable to each other, resulting in a comparison result such as “identical”, “overlapping” or “disjoint”, helping to establish an order of how relevant feedback is to a circumscribed issue.

A primary target points to the data set or resource the feedback is about.

Figure 23: The 'subject' qualifier of a feedback item.
A **secondary target** points to datasets to which the feedback does not directly relate, but which are relevant in such a way that a user searching for feedback on these resources is likely to also be interested in this feedback.

A **supplementary target** adds additional references, for example, another region in another data set with similar problems. It is used to formally model references that somehow are related to the feedback item at hand, but does not imply that the feedback is relevant for the referenced subject. Giving an example or pointing to data merely suspected to be relevant should be modeled as a supplementary target.

As described, roles qualify the relevance of an item to a target. When trying to discover relevant feedback for a target, a possible way to deal with this could be ordering feedback linked by a primary role before any feedback which “is” secondary and then supplementary.

All three types of target are illustrated in Figure 24.

![Figure 24: Primary, secondary and supplementary targets on a feedback item](image)

5.2 Tasks

This section discusses issues specific to the application of the user model to perform certain tasks. Since GeoViQua will implement some kind of feedback catalogue or infrastructure, the tasks are derived from this setting. Implementations of the user quality model are expected in the Task 5.3 of the workpackage WP5 of the GeoViQua project.

5.2.1 Encoding feedback catalogue answers

A catalogue answer is best encoded in a **FeedbackCollection** XML element (or an equivalent in another encoding such as JSON). All feedback items relevant in an answer can be placed in this element.
5.2.2 Submitting feedback

Submission should also be based on the FeedbackCollection (or analogous elements of other encodings). On submission, applicable constraints should be checked by a server/catalogue service application.

5.3 Opportunities in the implementation of the user / feedback model

This section discusses constraints that (may) make sense within GeoViQua, but are not enforced at the conceptual model or XSD level. It also discusses options in the model which can allow the generation of tools to ease the production of metadata and user feedback.

5.3.1 New elements

5.3.1.1 Require domain qualifier

We could require one or more domain qualifier to be added to each feedback item. With user accounts and sensible defaults this might lead to better qualified feedback.

5.3.2 Default value recommendations for the user /feedback model

Sensible use of default values will be very important in encouraging use of the models through smarter clients which can harvest at least some of the required information to help populate a metadata document or a feedback item. Automated usage reports are one example.

5.3.2.1 Domain

The first user domain, if known, may be advertised as default in the UI for generating feedback. Thus, a default domain qualifier would be added which matches the user’s background if s/he does not intervene.

5.3.2.2 Role

The user may have more than one user role, e.g. s/he may be a producer and a consumer of geodata. Thus, one user may want to speak “officially” as a producer at some time (e.g. endorsing other feedback) but not another. Consumers of the feedback may also be interested in filtering to view feedback from only a subset of users (e.g., commercial data producers). Therefore, each FeedbackItem contains a single mandatory role code which the user should specify. However, since there usually is an unproblematic default code (most users will just have one role anyway) this could be offered as default in a UI. At the very least, the user information for a user could be recorded in a persistent manner and re-used, allowing the user to choose from the subset of roles which they have specified in their user information, or to edit an existing user information record when they work in a new domain, without having to re-enter contact details.
5.4 Application domain ontology

GEMET concepts seem a good candidate for a domain ontology, which would be used to properly identify relevant feedback. This is demonstrated in section 6.4.3, and the value of such domain information is illustrated with a search use case in section 7.9.

6. Using the GeoViQua schemas

The root element of the producer metadata document should be a GVQ_Metadata element.

The root element of user-submitted data should be a GVQ_FeedbackCollection containing one or more GVQ_FeedbackItem elements.

When producer and user quality information are to be aggregated using a brokering approach, the document returned may take the form of a GVQ_Metadata element containing 0 or more GVQ_FeedbackCollection elements named ‘userFeedback’. The design which permits this is shown in Figure 25.

![Diagram showing the structure of GeoViQua schemas](image)

**Figure 25:** A GVQ_Metadata element may contain 0 or more GVQ_FeedbackCollection elements.

Because GVQ_Metadata is substitutable for MD_Metadata, this permits a collection of producer and user metadata to be returned as the content of a CSW catalogue response.
6.1 Dataset and unique resource identifiers

Some sort of unique identifier for each dataset / resource will be crucial in connecting producer and user quality information from a variety of distributed sources. This approach, which is effectively a form of runtime matching similar to that which supports LinkedData\(^\text{10}\), is fundamental to the construction of dynamic metrics such as the GEOLabel\(^\text{11}\). It is also vital if the hierarchical relationship of datasets is to be recorded and understood, allowing a dataset to inherit quality information from its ‘parent’. However, beyond the provision of many ‘target’ identifiers in the models presented here, we cannot enforce the management, maintenance, or resolving of such identifiers: this is an issue which really devolves to GEOSS.

6.1.1 Recommended practice for recording dataset identifiers

Many GeoViQua quality elements contain unique identifiers for datasets, encoded as MD_Identifier elements. The origin and maintenance of these unique identifiers is outside the scope of GeoViQua, but since it is a problem that affects GeoViQua in the ability to relate the quality elements to datasets, GeoViQua is stimulating the debate about metadata and data identifiers in GEOSS as an AIPS technical activity\(^\text{12}\). It is assumed that there will be some option available composed of a code / codespace combination. In order to be compliant with INSPIRE’s URI mappings, it is recommended that this information is recorded in an MD_Identifier element using the codespace and code elements, as shown below:

```
<updated19115:MD_Identifier>
  <gmd:codeSpace>
    <gco:CharacterString>
      uncertgeo.com
    </gco:CharacterString>
  </gmd:codeSpace>
  <gco:CharacterString>
    56666ershvtyrf456
  </gco:CharacterString>
</updated19115:MD_Identifier>
```

\(^\text{10}\) http://linkeddata.org/

\(^\text{11}\) The GEOLabel is a sub-activity of the GEO former ST-09-02 group (currently included in ID03). They aim is to develop a GEO Label for GEOSS that is related to the scientific relevance, quality, acceptance and societal needs for activities in support of GEOSS as an attractive incentive for involvement of the Science and Technology communities. Complete a practical definition of the GEOLabel is also a task in the GeoViQua project.

\(^\text{12}\) http://twiki.geoviqua.org/twiki/bin/view/AIPS/UUIDTopic
6.1.2 Identifiers - uniqueness and canonicalization

The target is not really intended to be defined within the model. Potentially the target should solve the problem of creating referenceable hierarchies and could be shared with other models, i.e. from a feedback perspective it merely establishes the context that really lives outside in GEOSS.

It has been decided that, for the purpose of modelling feedback, the existence of a globally unique identifier for the referenced resources can be postulated.

As far as we could determine, that is not the case within GEOSS. For prototyping and evaluation that may be acceptable, but in future we will need to make use of a service of some kind which is capable of resolving dataset identifiers.

6.1.3 Identifiers - Granularity

Not every feedback target can be assumed to have an identifier, because GEOSS datasets are of mixed granularity. As a result, targets need to be able to discern concepts that are accumulated at the granularity that can be identified using well-defined identifiers. In other words, it may well be necessary to specify how resources which are known to exist can be identified in a “synthetic” manner to allow feedback submission and discovery.

For example, a dataset may be organized into layers which do not have a GEOSS identifier. To be able to target such a layer, the targets should be equipped with a locally unique (i.e. within the dataset) layer identifier (e.g. a name) and a parent target which contains the actual GEOSS identifier (possibly canonicalized).

6.2 Relationship of the GeoViQua information model to the QIAP protocol

6.2.1 QIAP: a protocol which uses metadata quality information

The European Space Agency (ESA) has observed the need for standardizing the electronic exchange of information on detailed product quality and corrective actions for earth observation data. An initiative to support this has been the definition of the Quality Information Action Protocol. The goal of QIAP is to standardize how and in what form information about quality issues is sent between the different actors that are involved, and how this information should be treated. This is strongly related to the use of a user and producer model. Within GeoViQua we have investigated whether elements of the QIAP can be reused within for these models as follows:

- First, we investigated the knowledge on how quality information can be shared by producers and users. Elements used for QIAP have been investigated and this knowledge was tested as potential input for the user and producer model.
• Secondly we checked whether the formats of QIAP and the communication used for the models match. This would mean that tooling that uses the QIAP might extract the information it requires from the metadata available in GEOSS.

6.2.2 QIAP concepts versus GeoViQua user and producer model

QIAP focuses on the notification and storing of issues detected in data. At the start of a quality issue is the person or software tool; the user that discovers the anomaly. This can be reported as a 'quality event report'. Information about the issue will have to be sent to the organisation that is responsible for the data products (e.g. the data provider, the data producer). The producer will then investigate and confirm the anomaly. When the anomaly is confirmed, users may need to be informed of the anomaly and may be provided information on corrective actions that can be performed as a workaround solution. The producer will create a 'quality issue report' for this. In an ideal situation an anomaly gets resolved by correcting the problem, and the data and / or quality metadata is updated. If this is not possible, workaround solutions should be defined.

Therefore for a protocol like QIAP, two main communication channels are used: the communication between the user discovering an anomaly and producer and the communication between the producer and all potentially affected users. The GeoViQua producer and user model both contain the GVQ_DiscoveredIssue element (section 4.4) which also can be used for this communication.

Within QIAP there is a clear distinction between the information a producer provides on a found issue (as a quality event report) and what a user provides (quality issue). This is because the user can often indicate a problem with the data, but does not always have the knowledge or possibility to analyse the problem.

There is also a difference in the use of the discovered issue. The information (feedback) provided by the user will be queried by users who want to view quality information on a certain dataset. Also the producer will check if issues have been found on his data. This information will probably be inspected manually. The quality information provided by the producer is considered to contain information on how to handle the data, in other words it should define the action that end users should undertake when using the data. It can be seen as a bug tracker for data and could be used by tools to automatically filter data.

The following subsections show two mappings:

• *User model input to QIAP quality event.*
• *Producer model to QIAP quality issue.*
6.2.3 User model ‘discovered issue’ feedback vs QIAP ‘quality event’ report

A Quality Event Report is defined with the following classes and fields and the instances of a field (see Figure 26).

![Figure 26: Components of a QIAP Quality Event Report.](image)

The feedback model can be mapped to the event report in the following way. A quality event report consists of an originator with at least one quality event. The GVQ_FeedbackItem created in the user model will contain the user information and this user can add a GVQ_UsageReport containing a GVQ_DiscoveredIssue. The creation date is filled in and the mission and instrument can be provided in the tags.

From the GVQ_DiscoveredIssue elements, the knownProblem (description) and workaround (Proposed Resolution) and optionally the referenceDoc will be filled in, the latter providing more information than is available in the QIAP event Report. The affected product can be indicated with the target Identifier. To describe the product references and affected values of the product, the targets of the feedback item need to be filled in.

It is thus possible to capture the Quality Event Report inside a GVQ_FeedbackItem.
The mapping can be captured in the following table:

<table>
<thead>
<tr>
<th>User Model</th>
<th>QIAP Quality Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>GVQ_FeedbackItem/user/</td>
<td>Originator/Organisation</td>
</tr>
<tr>
<td>GVQ_UserInformation (including (userRole)</td>
<td>Quality Event/Mission or Quality Event/Instrument</td>
</tr>
<tr>
<td>GVQ_FeedbackItem/secondaryTarget/</td>
<td>Quality Event/Description, Quality Event/Proposed Resolution</td>
</tr>
<tr>
<td>GVQ_FeedbackTarget/dataFocus/</td>
<td>Quality Event/Instrument</td>
</tr>
<tr>
<td>GVQ_DataFocus/otherFocus or</td>
<td>Quality Event/Affected Products/Product (reference) and Quality Event/Affected</td>
</tr>
<tr>
<td>GVQ_FeedbackItem/tags</td>
<td>Products/Affected Value</td>
</tr>
<tr>
<td>GVQ_UsageReport with</td>
<td></td>
</tr>
<tr>
<td>GVQ_DiscoveredIssue/knownProblem and</td>
<td></td>
</tr>
<tr>
<td>GVQ_DiscoveredIssue/workAround</td>
<td></td>
</tr>
<tr>
<td>GVQ_FeedbackItem/primaryTarget/</td>
<td></td>
</tr>
<tr>
<td>GVQ_FeedbackTarget</td>
<td></td>
</tr>
</tbody>
</table>

6.2.4 Producer model ‘discovered issue’ statement vs QIAP ‘quality issue’ report

The components of a QIAP Quality Issue report are illustrated below in Figure 27.

![Diagram of QIAP Quality Issue Report](image)

**Figure 27**: Components of a QIAP Quality Issue Report.
For the Quality Issue report in the producer model, again the GVQ_DiscoveredIssue element is available. Alternative datasets can be specified, as can a link to a resource which is fixed/corrected. This would help to inform users that they should not use certain data, or steer them to use other datasets/ or parts of datasets. Automatic filtering of data in QIAP is not supported with the current design. Note that from the user information provided with the feedback item it should be clear that it has been created by a trusted user, namely the producer of the data. The userRole defined in the GVQ_UserInformation element will indicate that the GVQ_FeedbackItem was created by a producer.

In this case the mapping can be carried out in two alternative ways. The target field of the GVQ_DiscoveredIssue element simply specifies the identifier of a resource, and does not permit the specification of a focus — i.e., a spatial or temporal extent within that resource. Therefore the GVQ_DiscoveredIssue should be contained within a more informative element. The following two strategies are available to map the QIAP into the GeoViQua models, and that should be used depending on context:

6.2.4.1 Embed ‘discovered issues’ within data quality elements.
The producer model, like the original ISO model, allows the scope of a data quality element to be defined, and therefore the first option for a producer who discovers a problem at the time of data production is to generate a data quality element within the metadata document which specifically holds one or more GVQ_DiscoveredIssue elements, and which uses DQ_Scope to specify the subset to which the report applies.

6.2.4.2 Exploit the user model to supply information on discovered issues
Alternatively, if a producer discovers an issue after resource metadata has been published and disseminated, that producer could make use of the user model feedback system to provide the problem’s causes and resolutions. This strategy is also suitable where additional information about the focus of the problem needs to be supplied — for example, a restriction to specific spectral bands. This information may be supplied within the otherFocus attribute.

6.3 Use cases: producer quality model
Example use cases for the producer model have been detailed earlier in the text, and are referenced accordingly in this section. The figures in this section show a HTML rendering of the information in an example producer metadata document. The HTML rendering was generated using an XSLT 2.0 stylesheet, and is available online at:
6.3.1 Citing a publication within a metadata document

This is described in section 4.1. The documents cited may fall into several categories – description of the data, documentations of uses of the data, evidence of the data lineage and documentation of discovered problems in the dataset. Publications may be of any type, including *wikis* and online videos as well as traditional reports and journal articles.

The GeoViQua Publication model allows a ‘DOI’ and an ‘onlineResource’ to be specified; meaning that direct access to the resource can be encoded in the metadata and presented to a user with appropriate styling. Figure 28 shows an example.

![Figure 28: Publication information from a GeoViQua producer metadata document.](image)

6.3.2 Producer supplies ‘soft knowledge’ or advice on issues discovered with the dataset which cannot be easily encoded elsewhere.

One requirement identified from initial user and producer surveys was for producers to be able to supply ‘soft knowledge’ about their data – i.e., recommendations, warnings and tips which do not fit into the current conceptual model of quality as available in the ISO metadata model. As described in section 4.4, such ‘soft knowledge’ can be encoded as a GVQ_DiscoveredIssue which gives the opportunity to specify alternative datasets, and give details about the nature of the problem. An example is shown below in Figure 29.

![Figure 29: Producer ‘soft knowledge’ from a GeoViQua producer metadata document.](image)
6.3.3 Recording the traceability of a quality statement

As described in section 0, this can be achieved using an extended DQ_MetaQuality element. An example of the result is shown below in Figure 30.

![Figure 30: Traceability information from a GeoViQua producer metadata document.](image)

6.3.4 Citing one or more datasets used as reference for a quality evaluation

As described in section 4.3, this can be done by giving an identifier and details for the relevant dataset. Since an MD_DataIdentification is used, the details given can be substantial if required. An example is shown in Figure 31, with the summarised information about the reference dataset highlighted in red.
6.3.5 Providing full statistical information on the results of quality assessments.

As described in section 4.5, a wide variety of statistical metrics may be encoded in UncertML. Figure 32 shows an example of a confusion matrix which has been encoded in UncertML using the ConfusionMatrix type.
An UncertML ConfusionMatrix from a GeoViQua producer metadata document in HTML rendering.
6.4 Use cases: user model

RATINGS

6.4.1 A rating based on a dataset as a whole

- User creates one GVQ_FeedbackItem. In this and all subsequent cases, it is assumed that the user is not generating raw XML/JSON, but is generating these entities by filling out a form which allows them to select the dataset from a list or search, and to select /enter other values as appropriate using an easy and interactive interface.

- The user supplies a rating (numerical score between 1 and 5)

- The primary target of the GVQ_FeedbackItem which is generated will refer to the data set in question, using a unique identifier (MD_Identifier).

- User information (as for all subsequent cases) may be attached to the GVQ_FeedbackItem by accessing further parts of the form, to give varying amounts of detail about the user’s identification, expertise and domain.

6.4.2 A rating based on a spatial or temporal subset of the data

This works as described for 6.4.1, except that a data-centric focus is added to the primary target reference. The user would supply this by selecting a spatial and / or temporal region, and specifying any specific spectral bands to which the feedback applies. This combination describes the part of the dataset the feedback is about, by giving a bounding area and/or time frame. If more than one ‘hotspot’ area of interest/concern is being referred to, further focus elements may be added.

6.4.3 A domain-specific rating in the context of weather forecasting

This works as described for the rating above, except that the domainURN of the item is <http://www.eionet.europa.eu/gemet/concept/9278> - a “weather forecasting” domain focus.

6.4.4 Justification for a rating

This involves creating a GVQ_FeedbackItem containing:

- One rating (this contains a score and a text justification)
As in the examples above, the GVQ_FeedbackItem should point to the data set in question as its primary target.

**GENERAL FEEDBACK**

6.4.5 Provide general feedback on a complete dataset

- User creates one GVQ_FeedbackItem
- The item contains a comment.
- The primaryTarget of the GVQ_FeedbackItem refers to the data set in question, using a unique identifier (MD_Identifier).

6.4.6 Provide general feedback on a part of the data

- As for 6.4.2, one of more data-centric foci should be added to the primary target reference to describe the part of the dataset to which the feedback refers.

6.4.7 Provide general feedback on a metadata record

- User creates one GVQ_FeedbackItem
- The item contains a comment.
- The primaryTarget of the GVQ_FeedbackItem refers to the data set in question, using a unique identifier (MD_Identifier).

6.4.8 Provide general feedback on a specific metadata record elements

Again, the user creates one GVQ_FeedbackItem containing a comment, with the metadata document as its primaryTarget. However, this time sufficient detail must be given in the comment to identify the section of the metadata document to which the feedback refers. For an XML document, this may be achieved using xpath.

6.4.9 Add a domain-specific comment to a data set

A domainURN which maps to a conceptual dictionary description of the application domain should be supplied, as shown in 6.4.3.
6.4.10 Start a thematic discussion on a dataset
This could work like a message board opened by creating a GVQ_FeedbackItem with a subject, but no comment or other content, giving the dataset as primaryTarget. Any replies to this item are seen as posts on the subject.

EXTERNAL FEEDBACK

6.4.11 Report external feedback related to a dataset
- User creates one GVQ_FeedbackItem
- The item contains an externalFeedback element, which locates the external feedback using a resourceURL, and specifies a mime-type which can be used to retrieve the external feedback.
- The primaryTarget of the GVQ_FeedbackItem refers to the data set in question, using a unique identifier (MD_Identifier).

6.4.12 Report external feedback related to a part of a dataset
This works as for 6.4.12, except that one or more data-centric foci should be added to the primaryTarget as described in 6.4.2.

6.4.13 Report publication which cites a dataset
- User creates one GVQ_FeedbackItem
- The item contains a citation element, which locates the publication using details such as ISBN, ISSN, doi, onlineResource URL, and standard citation information such as author and title.
- The primaryTarget of the GVQ_FeedbackItem refers to the data set in question, using a unique identifier (MD_Identifier).

ADDITION PRIMARY, SECONDARY AND SUPPLEMENTARY TARGETS

6.4.14 Add primary target
A GVQ_FeedbackItem may be directly applicable to more than one resource. If this is the case, then more than one primaryTarget may be added, each with its own unique identifier.

6.4.15 Add secondary target
A GVQ_FeedbackItem on resource A may be not directly applicable to resource B, but it may be very relevant, such that anyone searching for feedback on resource B would benefit from the option of seeing and using the feedback on resource A. An example might
be a piece of feedback on an individual Landsat tile which is relevant to all Landsat users, but does not directly apply to the whole Landsat product. In this case, resource B (the generic Landsat dataset) should be added as a secondaryTarget while resource A (the specific Landsat tile) remains the primaryTarget.

6.4.16 Add supplementary target
A GVQ_FeedbackItem which references a resource only for illustrative purposes should use supplementaryTarget for that. An example might be a GVQ_FeedbackItem on resource A explaining an error found to be in A. Anticipating the difficulty of discovering and treating the error, the user adds a supplementaryTarget to a dataset B which also had the type of error in an earlier version but it got fixed later by its producer. The producer might want to follow this link to B to help assess the error. It is, however, unlikely that anyone interested in B will want to know about A, and thus a supplementaryTarget is the right choice.

SEARCH

6.4.17 Search for comments pertaining to a domain
This would work by first finding a URN within the applicable ontology, e.g. GEMET concepts. Then, all Domains which map to the URN are identified (we might want to include related domains, using some semantic tools) and the GVQ_FeedbackItems which have the matching domainURNs are searched.

7. Full User stories

7.1 Supplying a comment on a problem with a dataset

User A wants to comment on the striping artefacts within a specific Landsat tile.

- Using a form in the GEOSS catalogue, they create a GVQ_FeedbackItem with a primaryTarget pointing at the unique identifier of that Landsat tile, and add a subject which is a simple comment or keyword 'Landsat TM Striping problems'. They also add a secondaryTarget to the target record, which points at this whole Landsat series, to indicate that this feedback will be relevant to people interested in that secondary target.

- They add a GVQ_UsageReport to the item with a reportAspect of 'Problem'.

- They supply user information (at minimum, this will contain a CI_Role of 'User')

- In the usage report they put some text (possibly minimal) in the usageDescription to say what they did with the data.

- They also add to the usage report a GVQ_DiscoveredIssue with some text describing the knownProblem (the striping, and its nature). They can specify workarounds and alternative datasets if they like. Here, they have to give the
dataset identifier for the Landsat tile again, but this can be copied from the primaryTarget information they have already supplied.

- They can add other things if they like, including information about the user’s expertise, application domain and role.

This feedback item would be of interest to someone using this specific tile, but also to someone using Landsat in general. It could be found by searching for GVQ_DiscoveredIssue reports with targets relating to this family of identifiers (Landsat). This means that the user (i.e., ultimately the Broker) needs to have access to a set of identifiers and a set of recorded hierarchical relationships between them. However, there is also an option of finding the feedback through a keyword search on the subject, which names Landsat TM as the resource. Text recorded in this subject field can be searched for relevant records (e.g., using the keywords 'Landsat TM')

### 7.2 Adding a secondary target

User B is an agricultural land use researcher who wants to provide domain specific comments on a dataset X that covers the South Wales region for the whole of June 2012 and contains crop codes. Crop codes presence is not very common in the datasets any longer and therefore user B wants to emphasise that X contains all crop codes for the region. As additional information, user B wants to include links to several other datasets that contain crop codes, thinking that this might be of use to other agricultural land use researchers or professionals.

- User B creates a GVQ_FeedbackItem with a primaryTarget pointing at the unique identifier of the dataset X, and adds a subject which is a simple comment or keyword 'Agricultural - crop codes for South Wales region'.
- User B adds a secondaryTarget pointing at the unique identifier of a dataset Y (which also contains crop codes), to indicate that this secondary dataset may also be of interest.
- User B adds another secondaryTarget pointing at the unique identifier of a dataset Z that also contains crop codes but this time for May 2012. The user can add several more secondaryTargets if he/she thinks that they may be relevant.
- User B then adds a GVQ_UsageReport to the item which has three reportAspects: 'FitnessForPurpose', 'Usage' and 'Alternatives'.
- User B supplies some user information about him/herself (at minimum, this will contain a CI_Role of 'User')
- In the usage report he/she puts some text (possibly minimal) in the usageDescription to say what he/she did with the data.
- User B adds a text comment to the feedback item which stresses that dataset X contains a complete set of crop codes of agricultural land use.
7.3 Adding a supplementary target
User C is a small producer of oceanographic datasets who uses low level satellite and aircraft imagery and processes it to higher level products. User C knows that one of the sensors he uses works better in the middle of the range rather than at the edges, although he has no “hard” proof of it. To warn potential users he wants to provide a comment describing a potential problem with dataset X that he produced. He also wants to provide supplementary links to other datasets that were derived using data collected with that sensor.

- **User C creates a GVQ_FeedbackItem with a primaryTarget pointing at the unique identifier of the dataset X, and adds a subject which is a simple comment or keyword ‘Sensor range issues’. A parent identifier may be automatically added to the target record, which points at this whole data series he produced. User C does not have to provide this parent information, as it should be automatically looked up within GEOSS.**
- **User C then adds a supplementaryTarget, to indicate that this feedback is also relevant to other datasets he produced using this sensor. Users who will be interested in this feedback may not directly be interested in the supplementary dataset, but providing links to one or more datasets with a similar problem could emphasise the problem with the sensor’s range, and give a useful example.**
- **Then he adds a GVQ_UsageReport to the item with a reportAspect of ‘Problem’.**
- **User C then supplies some user information about himself (at minimum, this will contain a CI_Role of ‘User’) specifying that he is the producer of this data. Perhaps in this context, an obligatory rolecode and expertiseLevel would be useful.**
- **In the usage report he puts some text (possibly minimal) in the usageDescription to say that this dataset was not directly used by him, but he is the producer.**
- **User C also adds to the usage report a GVQ_DiscoveredIssue with some text describing the knownProblem (the problem with the sensor range).**
- **He can add other things if he likes, including information about his expertise, application domain and role (producer).**

7.4 Citing a publication which uses specific data
User D wants to report a publication which uses global MODIS data from a certain time window for cropland mapping.

- **She creates a GVQ_FeedbackItem with a primaryTarget pointing at the identifier of the global product, and adds to the target a GVQ_Focus which is datacentric - it has a temporal extent in it which describes the temporal footprint of**
the data she used. She also adds a domainURN to the GVQ_FeedbackItem which reports (by mapping to a GEMET or other concept) that she focussed on cropland mapping.

- She supplies some user information about herself (at minimum, this will contain a CI_Role of 'User')
- She adds a GVQ_UsageReport to the item whose reportAspect is 'Usage'.
- In the usage report she puts an abstract in the usageDescription to say what she did with the data.
- She adds a citation to the GVQ_FeedbackItem, with the publication information in it.
- She can add other things if she likes, as above. A client that allows users to report publications could force them to add some contextual information.

This feedback would be of interest to someone looking for publications which cite MODIS data (possibly for specific time periods), or which relate to cropland mapping (possibly for specific time periods).

The same issues of search by identifier or keyword arise as for the case above, and are common to practically all useful searches. We haven't explicitly modelled a database of identifiers and their parent-child relationships in the feedback model because we can't promise or hope to maintain this in GeoViQua. What we have done is add many opportunities for sensible identifiers to be used which can later be resolved by software clients.

7.5 Reporting intercomparison of datasets, and fitness-for-purpose assessment

User institution E generates an updated contour map for a small area from a very high-quality DEM and compares it to the national cadastral product. They publish their product with a metadata document that conforms to the producer model, but while validating their data they find some serious frequency spike artefacts in the national product.

- They create a GVQ_FeedbackItem with a primaryTarget pointing at the national contour product, and add to that target a datacentric focus containing a spatial extent which describes the spatial footprint of the area they considered. They may add other descriptive text or domainURN information about the field in which they are working.
- They supply some user information about themself (at minimum, this will contain a CI_Role of 'User'). It can contain plenty of other information, and perhaps in this context, an obligatory userRole and expertiseLevel would be useful.
- They add a GVQ_UsageReport to the item which has three reportAspects: 'FitnessForPurpose', 'Alternatives' and 'Problem'.
In the GVQ_UsageReport they put some information in the usageDescription to say what they did with the data.

They also add to the usage report a DiscoveredIssue with some text describing the knownProblem (the frequency spikes). As an alternative dataset they cite their new contour data.

They add to the feedback item a GVQ_QualityOverride with fuller numerical and statistical information on the issue they discovered. It's a GVQ_DataQuality, so it has a reference to their data as the MD_AssociatedResource in the context of crossReference, and a reference to a fuller publication on the subject.

This feedback would be of interest to people looking for alternatives to the national contour product (for which they can find the identifier) within this region. They can specifically search by extent and by reports with an aspect code of 'Alternatives'. They can decide whether they trust this by looking at the reported evaluation, figures, reading a publication if one is cited, and looking at the role and contact details of the user that supplied the feedback.

7.6 Constructing a ‘GEOLabel’ for a dataset

User F wishes to see summary information for a dataset which corresponds to some aspects of the proposed GEOLabel (see footnote 11 for the definition for this concept). To do this, the producer AND user metadata for the dataset and related datasets can be queried at several levels. This querying will be the task of the GeoViQua Broker, querying one or more feedback catalogues.

Level 1 (basic information, which can be displayed as a simple symbol)

Is there a quality statement from the producer for this dataset?

If not, is there a quality statement from the producer for the parent dataset (e.g., series or global coverage)?

Is there any user feedback available in the feedback catalogue? (including citations of publications) If not, is there any user feedback for the parent dataset?

What's the average user rating for the dataset?

Level 2 (further detail into which the user may drill down)

What is in the quality statement? How complete is the general metadata of the dataset? Show a clickable summary (e.g., a positional accuracy record and 2 lineage records). Here,
it would make sense to include a few things that are found outside the 'data quality' section of the metadata document, e.g., spatial resolution.

What is in the user feedback? Show a clickable summary (e.g., 1 cited publication, 2 reports of usage and 78 identified problems)

What are the individual user ratings and associated comments?

7.7 Providing general feedback on a metadata record

User G is a climate forecaster who purchased climate raster data X from a data provider that he has never used previously. The provider’s website did not offer much detail on the datasets they provide so user G had to purchase this data blindly. After analysing the data, user G observed that the data itself is of good quality and the metadata record provided complies with the ISO 19139 standard, but elements such as lineage, in which he is particularly interested, were not present. He assesses the completeness of the metadata using the online NOAA / GeoViQua rubric\(^{13}\) (shown in Figure 33), and wishes to share the results with other potential users, and comment on his minor dissatisfaction with the data producer.

- **User G creates a GVQ.FeedbackItem with a primaryTarget pointing at the unique identifier of that metadata record, and adds a subject which is a simple comment or keyword 'Metadata record completeness problems'**.
- **User G adds a GVQ.UsageReport to the item with a reportAspect of 'Problem'**.
- **He supplies some user information about himself (at minimum, this will contain a CI_Role of 'User')**.
- **In the usage report he puts some text (possibly minimal) in the usageDescription to say how he assessed the data and metadata record**.
- **User G also adds to the GVQ.UsageReport a GVQ.DiscoveredIssue with some text describing the knownProblem (incomplete metadata record, no conformance to international standards)**.
- **He adds a publication referenceDoc which contains an onlineResource pointing at the NOAA / GeoViQua rubric analysis of this metadata document (see Figure 33, next page)**

An XML encoding of this feedback element can be seen at: http://uncertgeo.aston.ac.uk/GVQ_v2.2/example_documents/FeedbackUseCase_7_7_metadata.xml

\(^{13}\) This can be used by supplying the URL of the resource to be analysed as a parameter to http://www.ogc.uab.cat/cgi-bin/geossback/GEOSSBack.cgi
ISO 19115 Completeness Report with Quality extension

This report is the composition of two efforts first started by NOAA and complemented by GeoViQua with the quality extensions.

This report is produced using this stylesheet. Please contact Ted.Habermann if you have questions or suggestions.

Title: Set of 1km grid fields covering Germany for RSMS values for the reference period April where each element belongs to a year between 1971 and 1980

Total Spiral Score: 17/48

Summary of the results. Each spiral is represented by a row in the rubric. The columns show the % of the elements in that spiral that exist in the record. Click the spiral name for more details:

<table>
<thead>
<tr>
<th>Spiral</th>
<th>None</th>
<th>1-33%</th>
<th>34-66%</th>
<th>67-99%</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Spiral</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Connection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distribution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Context</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisition Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 33: A screen shot of the NOAA / GeoViQua completeness evaluation rubric for metadata.
7.8 Providing general feedback on a specific metadata record elements

User H is a land use data archivist who found an aerial photography dataset X that contains data on agricultural land use in South Yorkshire in 2009. After analysing both data and metadata record, user H discovered that the spatial coverage recorded in the metadata record does not truly correspond to the actual photography. User H wants to report this discovered issue with the spatial coverage metadata record element and also wants to provide comments on an intercomparison of dataset X with other dataset that she previously archived for South Yorkshire area.

- **User H creates a GVQ_FeedbackItem with a primaryTarget pointing at the unique identifier of the metadata element, and adds a subject which is a simple comment or keyword 'Metadata spatial coverage attribute problem'.**
- **User H also adds a secondaryTarget pointing at the unique identifier of other dataset that was used as a comparison.**
- **She adds a GVQ_UsageReport to the item with two reportAspects: 'Problem' and 'Alternative'.**
- **She supplies some user information about herself (at minimum, this will contain a CI_Role of 'User'). In this context, an obligatory rolecode and expertise level would be useful.**
- **In the usage report she puts some text (possibly minimal) in the usageDescription to say how data and metadata record were analysed.**

*User H also adds to the usage report a GVQ_DiscoveredIssue with some text describing the knownProblem (spatial coverage not being correct). The user can specify workarounds (provide actual spatial coverage information), and alternativeDatasets.*

7.9 Searching for a domain-specific rating on a resource

User I wants the average dataset rating resulting from usage by hydrologists FOR A DATASET, AND FOR THE SERIES TO WHICH THAT DATASET BELONGS.

*How to do it: Find the identifier of the parent dataset, and for both identifiers, find feedback records whose GVQ_ReportAspectCode is 'Usage' and whose GVQ_UserInformation>applicationDomain or GVQ_FeedbackItem>domainURN is a urn mapping to 'hydrology'. Then aggregate any ratings in those records.*
7.10 Searching for information on problems with a dataset for a specific space/time window.

User J wants to find out about any problems reported by research end-users using the data for analysis in Africa between 1990 and 1997

How to do it: Find records where this data is the primary target. In the case of the feedback server, query the GVQ_UserRoleCode to find research end-users and find coincident extents with the spatial and temporal extent of any datacentric focus elements. Look for discoveredIssue elements. If required, the search can be further refined by looking for usageReports with a reportAspect of 'Problem'.

7.11 Searching for individual products which are cited in more than 15 unique journal articles

How to do it: Searching GVQ_Publication records associated with GVQ_FeedbackItems. A simple search would simply count the number of unique citations (relying on ISSN) per resource. A more sophisticated search might take the hierarchy of data resources into account, and include publications relating to subsets of the specified resource. This would require access to an external taxonomy of GEOSS resources and their relationships.

7.12 Finding datasets which meet user-specified quality thresholds

These searches involve interrogation of producer data – namely, the metadata documents that are registered in GEOSS when a resource is published. Examples 1 and 2 are possible with existing standards. Example 3 shows how the GeoViQua model can allow a more sophisticated search.

1. Find all datasets with a RMSE Positional Accuracy better than 50m

How to do it: this raises tricky issues with units and measures here, including the problem of verifying that 'RMSE' was what was measured in every case. Possibly, a prototype search will only consider valueType urns referring to metres and kilometres, and convert between the two. (Other datasets could be presented to the user as 'units unknown' so they can investigate themselves).

2. Find me datasets with a pixel resolution smaller than 750m, and a thematic classification correctness greater than 85%

How to do it: This search is similar to the one above, but requires looking in MD_SpatialRepresentation as well as the dataQuality section of the metadata document.
3. Find me datasets whose quality was assessed with reference to UK Ordnance Survey’s Mastermap.

   *How to do it:* Get the dataset identifier(s) for the Mastermap product and look for it/them in the referenceDataset of any reported quality evaluation elements.

4. Find me datasets with an attribute accuracy < 1degC^2 for all pixels / locations in my region of interest.

   *How to do it:* This is a very tricky example, but if we had pixel / object level information on accuracy, it can potentially be carried out. Whether this can be done within the scope of GeoViQua remains to be seen!
Appendix 1: example metadata document based on the GLC 2000 dataset.

This is the complete example of a metadata record that follows the GeoViQua Producer Quality Model. Fragments of this example are explained throughout the document.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<gvq:GVQ_Metadata
 xmlns:updated19115="http://www.geoviqua.org/schemas/19115_updates"
 xmlns:xlink="http://www.w3.org/1999/xlink"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:gmd="http://www.isotc211.org/2005/gmd"
 xmlns:gco="http://www.isotc211.org/2005/gco"
 xmlns:gvq="http://www.geoviqua.org/schemas/ProducerModel/3.0"
 xmlns:gml="http://www.opengis.net/gml/3.2"
 xmlns:gmd19157="http://www.geoviqua.org/schemas/gmd19157"
 xsi:schemaLocation="http://www.geoviqua.org/schemas/ProducerModel/3.0 ../schemas/GeoViQua_Metadata.xsd http://www.uncertml.org/2.0 ../schemas/uncertml/uncertml.xsd">
  <gmd:fileIdentifier>
    <gco:CharacterString>c0dc2fd0-88fd-11da-a88f-000d939bc5d8</gco:CharacterString>
  </gmd:fileIdentifier>
  <gmd:language>
    <gco:CharacterString>eng</gco:CharacterString>
  </gmd:language>
  <gmd:characterSet>
    <gmd:MD_CharacterSetCode
      codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_CharacterSetCod
e" codeListValue="utf8" />
  </gmd:characterSet>
  <gmd:contact>
    <gmd:CI_ResponsibleParty>
      <gmd:individualName>
        Etienne Bartholomé
      </gmd:individualName>
      <gmd:organisationName>
        JRC
      </gmd:organisationName>
      <gmd:positionName>
        Project Coordinator
      </gmd:positionName>
      <gmd:contactInfo>
        <gmd:CI_contact>
          <gmd:phone>
            <gmd:voice>
              (+39)0332-78-9908
            </gmd:voice>
            <gmd:facsimile gco:nilReason="missing" />
          </gmd:phone>
        </gmd:CI_contact>
      </gmd:contactInfo>
    </gmd:CI_ResponsibleParty>
  </gmd:contact>
</gvq:GVQ_Metadata>
```
Acronym: GeoViQua
Project title: QUAlity aware Visualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development
The Land Cover of the World 2000 is a product of the GVM Unit that is coordinating and implementing the GLOBAL LAND COVER 2000 Project (GLC 2000) in collaboration with a network of partners around the world, under the coordination of the European Commission's Joint Research Centre. To achieve its objective GLC 2000 makes use of the VEGA 2000 dataset: a dataset of 14 months of pre-processed daily global data acquired by the VEGETATION instrument on board the SPOT 4 satellite, made available through a sponsorship from members of the VEGETATION programme, including JRC. The legend of the map is based on FAO Land Cover Classification System (LCCS). Later version will also include regional legends.

The general objective is to provide for the year 2000 a harmonized land cover database over the whole globe. The year Two Thousand is considered as a reference year for environmental assessment in relation to various activities, in particular the United Nation's Ecosystem-related International Conventions.
Acronym: GeoViQua
Project title: QUALity aware Visualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development

<gmd:fileName><gco:CharacterString>glc2000_s.gif</gco:CharacterString></gmd:fileName>
<gmd:fileDescription><gco:CharacterString>thumbnail</gco:CharacterString></gmd:fileDescription>
<gmd:fileType><gco:CharacterString>GIF</gco:CharacterString></gmd:fileType>
<gmd:MD_BrowseGraphic>
<gmd:graphicOverview>
<gmd:fileName><gco:CharacterString>glc2000.gif</gco:CharacterString></gmd:fileName>
<gmd:fileDescription><gco:CharacterString>large_thumbnail</gco:CharacterString></gmd:fileDescription>
<gmd:fileType><gco:CharacterString>GIF</gco:CharacterString></gmd:fileType>
</gmd:MD_BrowseGraphic>
</gmd:graphicOverview>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:keyword><gco:CharacterString>Global Landcover 2000</gco:CharacterString></gmd:keyword>
<gmd:type><gmd:MD_KeywordTypeCode codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_KeywordTypeCode" codeListValue="theme" /></gmd:type>
</gmd:MD_Keywords>
<gmd:keyword><gco:CharacterString>SPOT Vegetation</gco:CharacterString></gmd:keyword>
<gmd:type><gmd:MD_KeywordTypeCode codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_KeywordTypeCode" codeListValue="place" /></gmd:type>
</gmd:descriptiveKeywords>
<gmd:descriptiveKeywords>
<gmd:MD_Keywords>
<gmd:keyword><gco:CharacterString>World</gco:CharacterString></gmd:keyword>
<gmd:type><gmd:MD_KeywordTypeCode codeList="http://www.isotc211.org/2005/resources/codeList.xml#MD_KeywordTypeCode" codeListValue="place" /></gmd:type>
</gmd:MD_Keywords>
</gmd:descriptiveKeywords>
<gmd:resourceConstraints>
<gmd:MD_LegalConstraints>
<gmd:accessConstraints>
The GVM unit delivers products and services derived from the analysis of satellite remote sensing data to the various DGs of the European Commission, Space Agencies, the scientific community at large and other users. For more
information about GVM and other collaborating partners see the GVM website http://www.gvm.jrc.it/default.htm
</gco:CharacterString>
</gmd:supplementalInformation>
</gvq:referenceDoc>
<gvq:referenceDoc>
<gvq:GVQ_Publication>
<gmd:title>
<gco:CharacterString>
A new land-cover map of Africa for the year 2000
</gco:CharacterString>
</gmd:title>
<gmd:date>
<gmd:CI_Date>
<gmd:date>
<gco:Date>2004-06-01</gco:Date>
</gmd:date>
</gmd:date>
</gmd:date>
</gmd:date>
</gmd:CI_Date>
</gmd:date>
</gmd:ISSN>
<gvq:target xlink:href="GLC2000">
<!-- the ID is used for xlinking within the document (see 'discoveredIssue'), the UUID for potential outside links -->
</gvq:target>
<gvq:doi>
</gco:CharacterString>10.1111/j.1365-2699.2004.01073.x</gco:CharacterString>
</gmv:doi>
<gvq:purpose>
<gvq:GVQ_PublicationPurposeCode
codeListValue="publication" />
</gvq:purpose>
<gvq:scope xlink:href="#datasetScope" />
<gvq:category>
<gvq:GVQ_PublicationCategoryCode
codeListValue="journalArticle" />
</gvq:category>
<gvq:onlineResource>
<gmd:CI_OnlineResource>
<gmd:linkage>
</gmd:linkage>
</gmd:CI_OnlineResource>
</gvq:onlineResource>
</gvq:referenceDoc>
</gvq:GVQ_DataIdentification>
</gmd:identificationInfo>
Acronym: GeoViQua
Project title: QUAlity aware VISualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development

<gmd:distributionInfo>
  <gmd:MD_Distribution>
    <gmd:transferOptions>
      <gmd:MD_DigitalTransferOptions>
        <gmd:onLine>
          <gmd:CI_OnlineResource>
            <gmd:linkage>
              <gmd:URL>
              </gmd:URL>
            </gmd:linkage>
            <gmd:protocol>
              <gco:CharacterString>
                WWW:LINK-1.0-http—link
              </gco:CharacterString>
            </gmd:protocol>
            <gmd:name gco:nilReason="missing" />
            <gmd:description>
              <gco:CharacterString>
                Global 2000 landcover Website
              </gco:CharacterString>
            </gmd:description>
            <gmd:function>
              <gmd:CI_OnLineFunctionCode
                codeListValue="information" />
            </gmd:function>
          </gmd:CI_OnlineResource>
        </gmd:onLine>
        <gmd:onLine>
          <gmd:CI_OnlineResource>
            <gmd:linkage>
              <gmd:URL>
              </gmd:URL>
            </gmd:linkage>
            <gmd:protocol>
              <gco:CharacterString>
                WWW:LINK-1.0-http—link
              </gco:CharacterString>
            </gmd:protocol>
            <gmd:name gco:nilReason="missing" />
            <gmd:description>
              <gco:CharacterString>
                Global legend overview
              </gco:CharacterString>
            </gmd:description>
            <gmd:function>
              <gmd:CI_OnLineFunctionCode
                codeListValue="information" />
            </gmd:function>
          </gmd:CI_OnlineResource>
        </gmd:onLine>
      </gmd:transferOptions>
    </gmd:MD_DigitalTransferOptions>
  </gmd:MD_Distribution>
</gmd:distributionInfo>

61
http://www-gvm.jrc.it/default.htm

<gmd:URL>
  <gmd:linkage>
    <gmd:protocol>
      <gco:CharacterString>
        WWW:LINK-1.0-http-link
      </gco:CharacterString>
    </gmd:protocol>
    <gmd:description>
      Global Vegetation Monitoring Unit website
    </gmd:description>
    <gmd:function>
    </gmd:function>
  </gmd:linkage>
</gmd:CI_OnlineResource>

<gmd:CI_OnlineResource>
  <gmd:linkage>
  </gmd:linkage>
  <gmd:protocol>
    <gco:CharacterString>
      OGC:WMS-1.1.1-http-get-map
    </gco:CharacterString>
  </gmd:protocol>
  <gmd:name>
    <gco:CharacterString>
      global_land_cover_2000
    </gco:CharacterString>
  </gmd:name>
  <gmd:description>
    The Land Cover of the World in the Year 2000
  </gmd:description>
  <gmd:function>
  </gmd:function>
</gmd:CI_OnlineResource>

<gmd:CI_OnlineResource>
  <gmd:linkage>
  </gmd:linkage>
  <gmd:protocol>
    <gco:CharacterString>
      WWW:DOWNLOAD-1.0-http-download
    </gco:CharacterString>
  </gmd:protocol>
</gmd:CI_OnlineResource>
Calculation of omission and commission statistics from sample polygons.

Validation of the global land cover 2000 map
FP7 Project Nr: 265178
Acronym: GeoViQua
Project title: QUALity aware Visualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development

<updated19115:MD_Identifier
   uuid="jrc.europa.eu:GLC2000_groundtruth">
   <gmd:code>
      <gco:CharacterString>GLC2000_groundtruth</gco:CharacterString>
   </gmd:code>
   <updated19115:codeSpace>
      <gco:CharacterString>jrc.europa.eu</gco:CharacterString>
   </updated19115:codeSpace>
</updated19115:MD_Identifier>
<updated19115:MD_AssociatedResource>
   <gmd:CITitle>Integrating new data visualisation approaches of earth Systems into GEOSS development</gmd:CITitle>
   <gmd:CIType xlink:href="http://www.isotc211.org/2005/resources/codeList.xml#DS_InitiativeTypeCode" codeListValue="study" />
   <gmd:DS_AssociationTypeCode codeList="http://www.isotc211.org/2005/resources/codeList.xml#DS_AssociationTypeCode" codeListValue="crossReference" />
FP7 Project Nr: 265178
Acronym: GeoViQua
Project title: QUAlity aware VISualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development

<!-- Report which contains 'traceability' information about the thematic classification assessment above -->
<gvq:GVQ_Traceability>
  <gmd19157:evaluation>
    <gmd19157:evaluationMethodDescription>
      <gco:CharacterString>
        1. Analysis published in a peer-reviewed journal article with full description of methodology. 2. Original data and reference dataset published via FAO and JRC.
      </gco:CharacterString>
    </gmd19157:evaluationMethodDescription>
    <gmd19157:evaluationMethodType>
    </gmd19157:evaluationMethodType>
  </gmd19157:evaluation>

  <gmd19157:result>
    <gmd19157:DQ_DescriptiveResult>
      <gmd19157:resultScope xlink:href="#datasetScope" />
      <gmd19157:statement>
        <gco:CharacterString>

        </gco:CharacterString>
      </gmd19157:statement>
    </gmd19157:DQ_DescriptiveResult>
  </gmd19157:result>
</gvq:GVQ_Traceability>

<!-- Report which contains 'traceability' information about the thematic classification assessment above -->
The thematic validation procedure was carried out using design-based inference which conforms to the CEOS Cal-Val recommendations. The resulting research was published in a peer-reviewed journal, and the component datasets generated for the validation may be obtained from JRC.

The thematic validation procedure was carried out using design-based inference which conforms to the CEOS Cal-Val recommendations.

Stratified sampling of ground truth blocks.

Systematic sampling to obtain an irregular stratification with different sampling rates for each stratum.

Joint Research Centre of the European Commission

GIS & Map Librarian; Research Assistant

codeList="http://www.isotc211.org/2005/resources/codeList.xml#CI_RoleCode" codeListValue="processor" /
Schematic sub-sampling of all GLC2000 landcover categories
Interpretation by regional experts of 3x3km sample boxes at specified sample sites. Regional interpreters used ancillary data like aerial photographs, thematic maps and NDVI profiles at coarse resolution in support to the Landsat interpretation.

Local knowledge required to verify exact landcovers at each ground truth point.

Collected table of landcover categories for all sample sites.
Cross-validation of independently interpreted sites against GLC2000 data to generate confusion matrix and Kappa statistic.

Identification of classes prone to confusion, and overall assessment of the accuracy of the classification.

Confusion matrix and overall accuracy value.

Confusion matrix and overall accuracy value.
Acronym: GeoViQua
Project title: Quality aware Visualisation for the Global Earth Observation system of systems
Theme: ENV.2010.4.1.2-2
Theme title: Integrating new data visualisation approaches of earth Systems into GEOSS development

<gvq:discoveredIssue>
<gvq:target xlink:href="#GLC2000" />
<!-- an internal document link for efficiency -->
<gvq:knownProblem>
<gvq:characterString>
Legend issues with South East Asia
</gvq:characterString>
</gvq:knownProblem>
<gvq:workAround>
<gvq:characterString>
Legend mapping available from FAO
</gvq:characterString>
</gvq:workAround>
<gvq:alternativeDataset>
<gvq:MD_DataIdentification>
<gvq:citation>
<gvq:CI_Citation>
<gvq:title>
South East Asia legend mapping: regional classes to GLC2000
</gvq:title>
</gvq:CI_Citation>
</gvq:citation>
</gvq:MD_DataIdentification>
</gvq:alternativeDataset>
This table gives a full hierarchical mapping from all regional SEA landcover categories to those used in GLC2000.