Data Harmonization Working Group (DHWG)
Engineering Report
GEOSS Architecture Implementation Pilot, Phase 3

Version 1.0 - FINAL

Revision History

Version	Date	Editor and Content providers	Comments
0.1	23/11/2010	Hervé Caumont	Report Template
			Content and chapters summaries
0.2	30/11/2010	Hervé Caumont	Draft version for review/add-ons by AIP participants
		Erin Robinson	Chapter on the GEOSS Common Record
		Dan Cornford	Several paragraphs on uncertainty management
		Peter Walker	PML scenario (NETMAR Case Study 1)
		Giuseppe Ottavianelli	Recommendations from QA4EO
0.3	07/12/2010	Hervé Caumont	Draft version for review and comments
		Brad Lee	CSIRO scenario and registered components
		Andrew Woolf	GIGAS contribution to AIP
		Ken McDonald	Update on GEO Task DA-09-01b
1.0	05/01/2011	Hervé Caumont	Final version for validation
		Brad Lee	Additional input on CSIRO workflows
		Nigel Fox	Review of QA4EO and uncertainty paragraphs
		Ken McDonald	Additional input on CEOS resources contributions

Document Contact Information

If you have questions or comments regarding this document, you can contact:

Name	Organization	Contact Information
H.Caumont	ERDAS / OGC	herve.caumont@erdas.com
G.Percivall	OGC	gpercivall@opengeospatial.org

Table of Contents

Table of Figures

Data Harmonization Working Group

1. Introduction

1.1 Scope of this document

This Architecture Implementation Pilot, Phase 3 Engineering Report (AIP-3 ER) reflects a set of Data Harmonization requirements, gathered from GEOSS AIP scenarios and use cases, and emphasized in the context of the registration process of contributed resources to the GEOSS Common Infrastructure (GCI).

As a result, the report underlines the prominent role of quality assurance procedures and quality measurements encodings for the development of interoperable, service-based, community applications that aim at supporting the combination of multiple source data products in a coherent way, so the resulting products are known reliable to a usage domain.

Potential impacts and benefits of the provided Data Harmonization requirements and recommendations are presented through both the GEOSS content-oriented discovery facility and the GEOSS web-based assess and access mechanisms, as offered to the end-users known from the several GEO Societal Benefit Areas.

1.2 Role of the GEOSS AIP

The GEOSS Architecture Implementation Pilot (AIP) leads the incorporation of contributed components consistent with the GEOSS Architecture, using a GEO Web Portal and a Clearinghouse search facility to access services through GEOSS Interoperability Arrangements, in support of the GEOSS Societal Benefit Areas. AIP is a GEO task for elaborating the GEOSS Architecture under the purview of the GEO Architecture and Data Committee (ADC).

This Engineering Report (ER) is a key result of the third phase of the Pilot. It is the result of the AIP coordination process of contributions from participant organizations to the Data Harmonization Working Group. A presentation of these participant's responses to the AIP-3 Call for Participation (CFP) is provided by the chapter 7.1 'References'.

AIP-3 was conducted from March to December 2010. A summary of the GEOSS Architecture activities is presented on the GEOSS Best Practices Wiki[1]. A separate AIP-3 ER describes the overall process and results of AIP-3 and thereby provides a context for this report.[2]

---

[1] http://wiki.ieee-earth.org/Best_Practices/GEOSS_Transverse_Areas/Data_and_Architecture/GEOSS_Architecture

[2] Listings of all AIP Engineering Reports: http://www.ogcnetwork.net/AIP2ERs, http://www.ogcnetwork.net/AIP3ERs

2. DHWG approach

2.1 Defining data harmonization activities

"A key challenge in spatial data infrastructures and systems of systems like INSPIRE and GEOSS is interoperability between systems and spatial data from a variety of sources. In this, it is important to note that interoperability has to go beyond any particular community, but take the various cross-community information needs into account." [GIGAS]

During the AIP-3 kickoff meeting, March 2010, the objective set for the Data Harmonization WG was to leverage contributions to AIP related to Information Modeling. Several action items have been identified to address this goal:

- Reconcile the typical **WMO/CEOS 'geophysical parameters'** with the modeling of 'geographical features'
  - o Leverage results from the **GIGAS** contribution, that is identifying ways to reconcile GEOSS 'geophysical observations' with INSPIRE 'features view', highlighting how far the INSPIRE data specifications framework applies equally to satellite and other Earth-observation data.
  - o Elaborate on the **PML** contribution, that is providing pointers to scenarios and use cases, especially the ones defined by the EC NETMAR project, and is designing tools and methods to combine and compare satellite time series with in-situ data and environmental models.
  - o Present the **CSIRO** forecast models, like the DH/DPIPWE/HT hourly flow forecasts model, or the CCAM 48-hour rainfall forecasts model
- Coordinate with the **GEO task DA-09-01a** on Quality Assurance for Earth Observations (QA4EO)
  - o Address the Quality Assurance process for 'Geophysical observations'
  - o Develop interests through the GEO Implementation Team of **QA4EO**, so to address a wider area of "sensor-to-product-to-service" lineage information
  - o Leverage the expertise provided by **ASTON University** / Dan Cornford and the related initiatives (UncertML, UncertWeb). Consider the contribution from **MUNSTER University** through the Air Quality interpolation service
  - o Illustrate the return of experience from **CSIRO** in using points-of-truth for calibration, and potential linkages with uncertainty management
- Coordinate with the **GEO task DA-09-01b** on Data, Metadata & Products Harmonization
  - o Main focus being so far on Earth Observations and **CEOS** activities
  - o Current perspective oriented towards the EC INSPIRE and GMES initiatives
  - o Integrate the return of experience from **CSIRO** in using and harmonizing OGC SWE standards for in-situ measurements

The Geophysical observations/parameters are considered as our "EO domain" starting point, requiring us to conduct harmonization with in-situ measurements, geographical features production workflows, and environmental (predictive) models settings. Quality assurance and uncertainty management are added as key enablers to this end.

The overall objective is to look for various approaches at the level of community adoption, so to first identify actual barriers to interoperability and data usability. And then, shape a convergence of GEO recommendations towards a set of harmonized standards, based on best practices and community orientations.
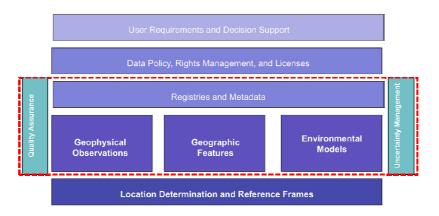

Figure 1.   A GEOSS community driver: reconcile separate models in one information viewpoint (AIP-3 kickoff, updated)

A GEOSS project would typically need to handle most of the building blocks presented in Figure 1 (Information Viewpoint of the GEOSS architecture). This is illustrated e.g. with the 4 scenarios presented later in this report. Orchestrating and putting together the building blocks highlighted with dotted line in the figure is defined in this report as a Data Harmonization activity. During the elaboration of this report, we referred as much as possible to the detailed definition of this overall data harmonization activity as provided by the GIGAS project (see hereafter). It consists in identifying the topics (or "data interoperability components") that are of importance in such an activity, generally requiring some level of agreement. When covered in this report, a data interoperability component as shown below is given a reference to the related chapter (cf. in the left column of the table).

Data Interoperability Component	Definition from the GIGAS project
Architectural support for data interoperability (cf. this chapter and Figure 1)	Description how data interoperability and data harmonization takes place into the general architecture.
Relevance to a community (cf. chapter 3)	Study the relevance of data harmonisation and/or semantic interoperability for a community. It also identifies the relevant reference documents.
Requirements (cf. chapter 4)	Documented requirements with regards to geographic information – including requirements on interoperability and/or harmonisation of data - in the community.
Terminology (cf. chapter 4.4)	A consistent language is key for semantic interoperability, and a process to establish and maintain the terminology is required. This includes terms used in the initiative that are related to the interoperability of data.
Multiple representations (cf. chapter 4 - GIGAS contrib.)	Practices for the aggregation of data across time and space and across different levels of detail.
Data transformation model / guidelines (cf. chapter 4 - GIGAS contrib.)	Rules for the transformation from one data specification to another data specification. Transformations are required for data and – in the case of on-the-fly transformations in a data access service – also for queries. Transformations between source and target application schemas are a key transformation type, but there may be other transformations required, e.g. coordinate transformation, edge-matching, language translation, format transformation, etc.
Spatial and temporal aspects (cf. chapter 4, QA point of view)	While the reference model specifies an overall framework, this aspect deals with the spatial and temporal aspects in more detail, for example, the types of spatial or temporal geometry that may be used to describe the spatial and temporal characteristics of a geographic feature.
Application schemas (cf. chapter 4, QA point of view)	Description of application schemas specified within the initiative as well as application schemas specified by other organisations that are re-used by the initiative (e.g. Observations & Measurements).
Reference model (cf. chapter 4, QA point of view)	The framework for the technical arrangements with regards to data.
Rules for application schemas and feature catalogues (cf. chapter 4, QA point of view)	Application schemas and feature catalogues provide the formal specification of geographic information and promote the dissemination, sharing, and use of geographic data through providing a better understanding of the content and meaning of the data. Across the individual themes, common rules are required to achieve interoperability.
Use of ontologies (cf. chapter 4, QA point of view)	Ontologies are formal representations of semantics and can be used to relate geographic information represented according to different application schemas.
Metadata (cf. chapter 4, QA point of view)	Rules for documenting metadata for datasets and services, on the three levels: discovery, evaluation, and use.
Data & information quality (cf. chapter 4, QA point of view)	Rules for the publication of quality information, e.g. on completeness, consistency, currency and accuracy.
Delivery (cf. chapter 4, QA point of view)	Rules for the delivery of geographic information. This includes the services used to deliver data, where applicable, as well as the encoding formats applied to encode data for the exchange between systems.
Maintenance of information about geographic data (cf. chapter 5 Conclusions)	Rules, processes and technologies applied to maintain information items relevant for the understanding and interpretation of geographic information, e.g. coordinate reference systems, code lists and thesauri, schemas, ontologies, etc.

Publication of information about geographic data (cf. chapter 5 Conclusions)	Rules, technologies and encodings applied to publish information items relevant for the understanding and interpretation of geographic information (see above).
Governance (cf. chapter 5, Conclusions)	Information about the governance of geographic information and their data specifications as far as it is relevant for data interoperability.
Extension points	Rules how data specifications can be extended to provide flexibility to adapt to different contexts and requirements. This is a key point in harmonisation initiatives.
Coordinate referencing and units of measurement model	Specification of the conceptual schema for spatial and temporal reference systems as well as units of measurements – including the parameters of transformations and conversions.
Identifier management	Specification of the role and nature of unique object identifiers.
Object referencing modelling	Rules for the specification of the spatial characteristics of a feature based on already existing geographic features, rather than directly via coordinates.
Multi-lingual text and cultural adaptability	Rules for the support for multi-lingual information in geographic information.
Maintenance	Rules for the maintenance of geographic features.
Portrayal	The schema for portrayal rules and symbology for geographic features and a description how portrayal is to be used within the infrastructure.
Consistency between data	Rules for the consistency between the representation of the same entity in different geographic data sets (for example along or across borders, themes, sectors or at different resolutions).
Data capturing rules	Guidelines which entities are to be represented as geographic features in a data set.
Conformance	Rules for the description of conformance tests in data specifications.

Figure 2.   GIGAS Comparative Analysis – Topics from Chp.5 Data Harmonization and Semantic Interoperability

Besides all these interoperability components are important and shall be treated at some point within GEOSS, the working group focused through AIP-3 mainly on the following: data harmonization relevance to a given community scenario, the associated requirements, the use of application schemas, the development of semantic interoperability (terminology, use of ontologies), the management of metadata and information about geographic data, the merge of multiple representations to allow data aggregation as well as the transformation rules between data specifications, the rules for the publication of quality information, the services and encoding formats used to deliver data, and the governance of data specifications.

2.2 Enhancing data usability (as our main goal)

Data usability is defined in ISO19157 (Candidate Draft status) as the degree of adherence to a specific set of data quality requirements, that is, a dataset or product adherence to an application or requirements set (see 'purpose of data' in Figure 3 below). It is modeled in ISO19157 as one data quality element coming along with the elements related to the completeness, consistency, positional accuracy, thematic accuracy and temporal accuracy (discussed as 'Quality assurance' in a following paragraph).

In the AIP context, we also found it useful to consider Data Usability through some relevant 'facets', as presented in Figure 1 below, that appear quite in line with the Data Harmonization activities. These facets are discussed hereafter and will guide our report conclusions, in order to build on the several applications and related sets of requirements contributed to AIP-3. One focus was thus to discuss some Data Usability levels (scientific usage for environmental modeling, operational usage for small scale map production,…) matching these application exemplars, related to some well known GEOSS datasets or products, and available through some established GCI components and services. It is expected that such an approach would fit well to the GEOSS Best Practices Wiki content, being used as an outreach vehicle to develop Data Harmonization within GEOSS, and to provide a consistent and rational basis for further DH developments in the GEOSS communities.
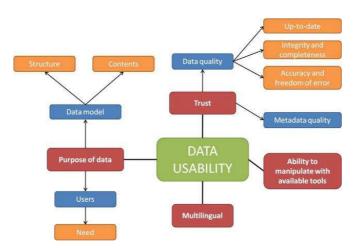
Figure 3.   Data Usability facets, as defined by the GIS4EU project (EC FP7)

Facet 1 - Dataset fitness for purpose: this facet relates to product's data models, generally defined as the response to a data usage requirements class. It is thus expected that a data producer can show for one dataset, with quantitative elements, how far it fits such usage class, i.e. declares the conformance of a product release to a particular product specification. Generalizing, or categorizing the overall product specification classes in the GEO context is one major challenge of the data reusability and data integration within GEOSS communities or Societal Benefit Areas (SBAs).

Of particular focus within this report, defining "Domain Features" is the basic task to perform when deciding to encode some domain objects in a data model like e.g. a GML application schema. Usually, it would be possible to simply pick and aggregate existing GML schema components in a target application schema. The profiling of domain features through a GML encoding is a critical step, with normative rules for constructing GML Application Schemas. It mainly relies on the General Feature Model of ISO19109:2005 « Rules for Application Schema », implying to define a 'feature' as a representation of a 'real world' phenomenon, and its relevant 'properties'. Such properties cover the feature attributes like the position and geometry, the feature associations like within a hierarchy, and the feature operations describing the dynamic changes of the feature over some range of reference, typically the time. Nevertheless, new tools and procedures still have to be developed so to use operationally full Application Schemas (typically through extensions), providing elaborated associations and operations.

User types then, and their set of operational needs, are validating the fitness for purpose of such Data Models. The GEOSS User types are considered here as an entry point to this purpose. They were discussed in the AIP-2 engineering reports[3].

Facet 2 - Quality assurance: this facet relates for us directly to contributions gained from the QA4EO, and is defined here as the need to assign to all data or information products a Quality Indicator (QI). This would allow stakeholders to evaluate a data product suitability for a particular application, given the stakeholders quality requirements. The QA4EO framework requires that such QI definitions must be transparent, internationally consistent and independent of sensor or application domain, and must be unequivocal and universal both in terms of its definition and derivation. The QI and evidence to support its 'value' must also be fully traceable back through the processing chain to the original source data and measurement.

In practice, this rule set is likely to be applied through a wide range of descriptors and terms (e.g. text or numeric) depending on the specific application or users needs. However, the DHWG consider all QIs should be based on a well defined statistically derived value, and this value should be the result of an assessment of its traceability to an agreed reference standard as propagated through the data processing chain, or obtained by rigorous in field validation.

---

[3] http://www.ogcnetwork.net/system/files/AIP2_Summary_Version_1_1_final.pdf

QA4EO is a set of guidelines and principles relevant to quality assessment in Earth observation, and is based on three main principles:

1. **Data Quality**: All data and derived products must have associated with them a Quality Indicator (QI) based on documented quantitative assessment of its traceability to community agreed reference standards.
2. **Data Policy**: Cal/Val data must be freely and readily available / accessible / usable. This implies that all Cal/Val data and associated support information (metadata, processing methodologies, QA, etc.) is associated with the means to effectively implement a quality indicator. In return, the provider must be consistently acknowledged.
3. **Communication and Education**: All stakeholders must have a clear understanding of the adequacy of the information, which should be accessible through a single portal and should be fully traceable to its origins.

Calibration is the process that enables an instrument's readings to be converted to physical units. For one satellite example this might be units of radiant energy e.g. Wsr^{-1}. Satellite instrument calibration activities take place throughout the lifetime of the instrument and beyond through retrospective re-calibration where coefficients of a stable reference may have been updated or a model improved. Prior to launch, instruments are typically calibrated in laboratories against known SI traceable references e.g. in this case sources of radiant energy (e.g. a blackbody of known thermodynamic temperature. Calibration continues while the instrument is in space either through some form of onboard Calibration device or by a vicarious method e.g. against a "surface based" target which might be a high level geophysical variable or a reflectance/radiance for example. Such post-launch calibration relies on the value assigned to the external target being well-calibrated either by some in-situ method or by another independent "reference satellite".. Validation refers to validation of the geophysical products generated from the instrument's observed 'measurands' e.g. radiances, and is a critical part of characterizing the instrument quality in the field, for both satellite and in-situ sensors.[4]

The distinction between calibration and validation in many ways is subtle and can largely be considered dependent on the level of uncertainty estimated for each measurement (a calibration requires that the instrument under test is viewing a target with smaller uncertainty).

QA4EO in its current state mostly uses satellites as case studies although its principles are in fact generic and can be adapted to all EO environments. The principles are supported by a set of guidelines which are written to provide advice and templates which can be used to help those seeking to implement:

- The document **QA4EO-QAEO-GEN-DQK-001** outlines the overall strategy that needs to be adopted for a QA system and starts with the need to define overall requirements for a community in terms of key quality/performance information needed to enable both interoperability and also to assess suitability for a particular application andis based on the concept of identifying and assigning quality indicators. A quality indicator is typically synonymous with an uncertainty judgment. The document takes the example of a measurement from a satellite sensor and briefly discusses the high-level issues that need to be considered for pre-launch and post-launch activities.
- The document QA4EO-QAEO-GEN-DQK-002 which is the core of the QA4EO guidelines states that "*the requirement driving QA4EO is the need to assign to all data / information products a Quality Indicator (QI), which will allow all stakeholders to unequivocally evaluate the products' suitability for a particular application. This requires that the basis for such a QI must be transparent, internationally consistent and independent of sensor or application domain. The QI and evidence to support its 'value' must also be fully traceable back through the processing chain to the original source data / measurement. This processing chain can be considered as a set of linked activities or processes (e.g. data collection, correction / conversion algorithm, dissemination, etc.), some operating in a direct linear path, others providing ancillary information to aid the next processing step*". Other parts of the document indicate that the QI must also be a "quantifiable" assessment of the degree of traceability. This leads to the conclusion that some form of uncertainty assessment is the most natural form of quality indicator. This is confirmed in subsequent specific guidance documents e.g. DQK-003 as stated below.

---

[4] See also the WMO Implementation Plan for a Global SpaceBased InterCalibration System (GSICS)

- The document **QA4EO-QAEO-GEN-DQK-003** defines a key concept in QA4EO, namely the definition of a reference standard is a measurement of reality or at least a community agreed proxy with known (and ideally minimal) uncertainty. The reference standard can then be used to assess the uncertainty in the observations or derived information to create the quality indicator, which is further defined: "*This requires that all data has associated with it a Quality Indicator (QI), which must be unequivocal and universal in terms of its definition and derivation. In practise there is likely to be a wide range of actual descriptors and terms used (e.g. text or numeric) depending on the specific application or users needs. However, all should be based on a statistically derived value and this value should be the result of an assessment of its traceability to an agreed reference standard (ideally SI) as propagated through the data processing chain*". Such forward assessment of uncertainties (propagating uncertainties through the processing chain) requires validation in a statistical sense, using reference standards. The QA4EO process, at this stage, emphasizes approximate methods (reasonable first order approximations) for propagating uncertainty, which might not always be robust in practice, and thus certainly require *validation in the field* of the final uncertainty assessment. There are more statistically motivated approaches to uncertainty and sensitivity analysis to consider for strongly non-linear processing chains, and deserving further development within the QA4EO guidance.
- The document **QA4EO-QAEO-GEN-DQK-004** discusses issues of organizing comparisons which are seen to be the basis of establishing the evidence to support the assignment of quality indicators. The document describes that in the simplest form, a calibration is a comparison and in its most sophisticated, it is an evaluation of consistency between peers of an uncertainty calculation. There is significant discussion on the importance of uncertainty budgets in this document, and the importance of breaking these down is emphasized. It should be noted that for the data end-user, it is really only the final uncertainty value that is of direct interest; however this is unlikely to be accepted without access to the provenance of this information, requiring documentation of all stages in its definition as specified in the QA4EO process. It is of course true that for some users and some applications, the full-breakdown of uncertainties may also allow greater analysis.
- The document **QA4EO-QAEO-GEN-DQK-005** provides the link to existing generic "best practice" guidelines for testing and validating models or algorithms and derived software used for measurement systems. These guidelines are recommended for use within the processing chain of Earth Observation (EO) data products. It emphasizes the need for providing detailed guidance on appropriate techniques for validation and assessment of the "fitness for purpose", and the means to assign a "quality indicator".
- The document **QA4EO-QAEO-GEN-DQK-006** reviews the application of the ISO Guide to the expression of Uncertainty in Measurement, often called "GUM". This describes the well established and internationally accepted metrological approach to managing uncertainty, and is the recommended practice for measurement systems in all manufacturing and academic sectors by ISO, standards laboratories, and accreditation bodies. Recent extensions to the GUM have been written to consider more complex problems involving for example monte-carlo methods.
- The document **QA4EO-QAEO-GEN-DQK-007** describes how to establish the evidence needed to underpin the whole process, and serves as a summary and guidance to a data provider or end user about what information is needed to support the documentation of a quality indicator. This essentially lists the desirable provenance / lineage information provided along with the quality indicator, and will vary according to the criticality and complexity of the application or the relative maturity (in terms of QA) of the community involved.

Overall QA4EO is a very laudable and ambitious attempt to harmonize the treatment of quality information in Earth observations, including both satellite and in-situ measurements and derived products. It can suffer from a lack of clarity in places, and from its expressed desire to remain a set of guidelines and recommendations, rather than defining processes, tools and standards that must be adopted to support and ensure interoperability of the results. It considers that it is for the user to specify its applicability and importance depending on their applications. And also given the relative immaturity and breadth of the GEO community (e.g. climate change to biodiversity) it is seen better to allow communities to have flexibility to define standards and tools where they feel it is appropriate and possible.

Furthermore, looking at the "accuracy and freedom of error" facet from Figure 1, there is an obvious stake in addressing approaches to manage, propagate and visualize uncertainty in a processing chain, up to the visualization and analysis tools, and of course this is the intended consequence of implementing the QA4EO guidelines.

Under this general framework, uncertainty is defined as "an absence of certainty", or sometimes "a state of having limited knowledge", although this is normally described specifically as epistemic uncertainty. Several types of uncertainty are recognized:

1. **Aleatory uncertainty** - arising from the fact that something it truly random or variable, for example radioactive decay. This is also sometimes called "statistical uncertainty". For example if we choose to summarize the climate (temperature) at a given location, over a given time period (assuming we can make perfect measurements) then the actual temperatures over the time period are known but there is variation so there will be irreducible statistical uncertainty assuming we summarize the climate, for example with a mean and standard deviation. This sort of uncertainty is often associated with objective, or frequentist notions of probability.

2. **Epistemic uncertainty** - arising from a lack of knowledge. If we talk about the temperature at a specific location at some time a few days in the future, this is not variable, there will be a single value, but unless we have a perfect forecasting model there will be epistemic uncertainty. We could reduce this by using a better forecasting model (hence this is often called reducible uncertainty), and the uncertainty is to do with our state of knowledge, and is not a property of the system. Thus this sort of uncertainty is typically associated with subjective interpretations of probability.

3. **Ontological uncertainty** - in Tannert et al (2007) the following definition is given: "ontological uncertainty is caused by the stochastic features of a situation, which will usually involve complex technical, biological and/or social systems. Such complex systems are often characterized by non-linear behavior, which makes it impossible to resolve uncertainties by deterministic reasoning and/or research". This is not so much about lack of knowledge or variability, but rather that for complex systems prediction might be impossible even with perfect knowledge - an example might be the oft (mis)used "Lorenz butterfly effect", essentially representing chaotic response to initial conditions. An example of ontological uncertainty (also subject to epistemic uncertainty) is making a temperature forecast for a given location for 200 years in the future - this will depend on factors, such as societal change, which are so complex that one has no real ability to even define a probability. Such situations are often tackled using scenario based modeling.

4. **Semantic uncertainty** - the lack of knowledge about the concepts we are discussing, for example "what is a forest?". This is often associated with fuzzy approaches, and is sometimes confused with ontological uncertainty. It can also be addressed with a (Bayesian) probability framework.

This description of uncertainty is not complete, nor unique. In any given circumstance it is likely that all aspects of uncertainty will be present, and surely there are others not listed, however the listed uncertainties are those which dominate most application scenarios relevant to GEOSS and more broadly to the physical sciences.

Facet 3 - Interoperability: this third selected facet of Data Usability refers to the ability to discover and use geospatial data resources, such as Earth observation data. Interoperability is often achieved by harmonization, based on common standards and practices, and can be achieved at a variety of levels, as proposed by Dan Cornford through the contribution of ASTON University to AIP-3[5]:

1. **Machine encoding interoperability** -- common underlying representation of basic data values, e.g. big-endian, byte order assumptions, often IEEE standards based, etc.
2. **Format encoding interoperability** -- common data format used, specifying e.g. order of elements, delimiters, tags etc, e.g. NetCDF, GML and O&M XML application schema, Shapefile, etc.
3. **Semantic dictionary interoperability** -- understanding of the meaning of the data values, based on semantics and ontology description languages, for example RDF / OWL. This is still an open research issue, and only recently have controlled vocabularies started to develop to enable this, e.g. SKOS. This is 'hardwired' semantics via a dictionary.
4. **Semantic machine interoperability** -- the real goal of semantic integration that machines can 'understand'

---

[5] https://wiki.aston.ac.uk/foswiki/bin/view/UncertWeb/UncertaintyAndStandards, 'Information interoperability' topic

concepts and reason with these, typically without having a central controlled vocabulary. This remains an active research area.

5. **Information interoperability** -- the relation of the data to reality or to other sources of the same information is quantified so that the data can be used appropriately in the given application. At present this is little addressed in the EO community (some might be tempted to call this quality or accuracy, but it is really uncertainty) other than through the new QA4EO.

The above list is not intended to be exhaustive, but provides an ordered set of levels of interoperability. Information interoperability is not often discussed, but is essential to data harmonization. Machine, format and semantic (either level) interoperability allow a data set to be accessed and made sense of, but it is not possible to use the data in a rational manner without some sort of information interoperability. In practice information interoperability requires the definition of the relation between the data set and the reality to which it pertains, or at least to a well defined and agreed reference. Note this requires a rigorous definition of reality or the reference. The relation between data and reality is naturally encoded using probabilistic uncertainty representations and these lie at the heart of achieving information interoperability. These are also very central to QA4EO.

2.3 Developing coordination (as our main process)

The AIP-3 Data Harmonization Working Group coordinated with several GEO tasks and GEOSS-related initiatives.

2.3.1 Quality Assurance Strategy (coordination with GEO Task DA-09-01a)

This GEO task aims at developing a GEO quality assurance strategy, beginning with data related to space-based observations, and evaluating expansion to in-situ observations. It encompasses a wide variety of disciplines, and a multitude of monitoring tools and procedures that require a quality metric to be associated to their outputs if to be reliably integrated into an expanded set of systems and services supporting the GEO SBAs needs.
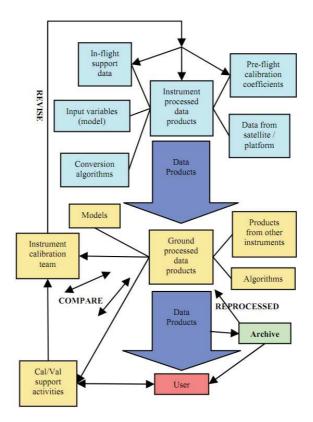

Figure 4.   Interdependencies of the EO data processing chain (QA4EO)

One key element in the QA4EO Approach is the processing chain, which can be considered as a set of linked activities / processes (e.g. data collection, correction / conversion algorithm, dissemination, etc.) some operating in a direct linear path and others providing ancillary information to aid the next processing step.

The objective of QA4EO is to assign a QI to the result of every critical independent step in an EO information product processing chain or an appropriate aggregate of such steps. Figure 4 provides a schematic data processing chain for a satellite sensor, where the complexity and interdependency of the various activities in the process chain can be visualized. Each activity can be considered as:
- a measurement – the use of an instrument to obtain information about an entity or
- a process – the transformation of information from one form to another. This may involve the combination of other information and/or the use of theoretical models

The 2009 QA4EO Workshop on Facilitating Implementation was held from September 29th to October 1st, 2009 in Antalya, Turkey, agreed to a series of steps to facilitate implementation of QA4EO into the GEOSS community. Amongst them, an action plan shall target applicability of QA4EO to the wider EO community and engagement of data providers and users, and the QA4EO framework document shall contain a set of guidelines, based on best practices, to provide guidance through templates and where appropriate exemplars.

Coordination of QA4EO and AIP should facilitate the translation of the QA4EO principle within the various application domains, to aid its implementation across all GEO communities.

2.3.2 Data, Metadata and Products Harmonization (coordination with GEO Task DA-09-01b)

This GEO Task aims at facilitating the development, availability and harmonization of data, metadata, and products commonly required across diverse societal benefit areas, including base maps, landcover data sets, and common socio-economic data. The task is a continuation of the GEO Task DA-06-04, with the same description.

- Earlier activities described the ongoing efforts of members and participating organizations, but did not address harmonization at GEO level.
- Current activities are to review and characterize contents of GEO registries as description of member/participating organization capabilities, and assess other well-known capabilities that may not yet be in the GEOSS registries.
- The task team has also promoted the interaction and exchange of information by members and other GEO tasks on community harmonization initiatives in an effort to establish best practices and lessons-learned. These include:
    1. CEOS/WGISS Integrated Catalog (CWIC) of satellite data collections.
    2. EC harmonization activities of EuroGEOSS and INSPIRE.
    3. WMO WIS
    4. CEOS/WGCV initiatives to characterize and document data quality.
    5. The task also hosted a joint workshop with GEO Task DA-09-02a, Data Integration and Analysis Systems.

The current GEO Task DA-09-01b participants and point of contacts are as follows:
- Co-Leaders
    1. **CEOS/WGISS/NOAA** - Ken McDonald (POC)
    2. **USA/FGDC/GSDI** Secretariat - Doug Nebert
- Contributors
    1. **Australia/Geosciences Australia** - Chris Body
    2. **CEOS/WGCV/USGS** – Greg Stensaas
    3. **CEOS/WGISS/NASA** – Yonsook Enloe
    4. **EC/EuroGEOSS** - Massimo Craglia
    5. **IEEE** - Steven Browdy
    6. **OGC** – George Percivall, Herve Caumont
    7. **WMO/OBS/WIS-DM** - Omar Baddour, Eliot Christian

2.3.3 GEOSS, INSPIRE and GMES an Action in Support (overall coordination)

The GIGAS EC FP7 project is fostering the coherent and interoperable development of the GMES, INSPIRE and GEOSS initiatives, through their concerted adoption of standards, protocols, and open architectures.

The GIGAS project contribution to the Data Harmonization topic of the GEOSS AIP-3 investigates the possibility of a common foundation between GEOSS and INSPIRE, using a marine forecasting scenario from GMES. The report highlights the important **role of the Observations and Measurements standard**, as a common foundation between INSPIRE and GEOSS in harmonizing the geophysical parameters view of Earth observations with the feature-based view of INSPIRE and spatial data infrastructures. Organized through the AIP-3 Data Harmonization Working Group, a presentation[6] of this work was made to the GEO Task DA-09-01b on June 24th, 2010.

Figure 5.   GIGAS project for the harmonization of standardization processes

AIP also made an overall presentation[7] to the CEN/TC 287 workshop on Interoperability between INSPIRE, GMES, and GEOSS, 8-9 November 2010, JRC, discussing the AIP coordination with CEN and JRC with regard to the upcoming GMES priorities for Data products harmonization.
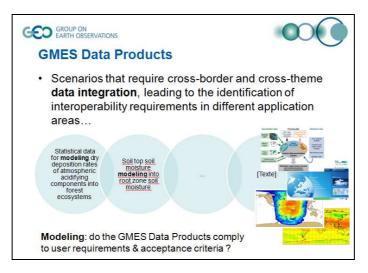

Figure 6.   GMES Data Products discussed from the AIP process perspective

---

[6] https://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-telecons/dhteleconjune24th2010/100624AIP-3DataHarmonizationTelecon.ppt

[7] https://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-discussions/computing-trust/20101108_AIP_to_CEN_workshop.ppt?attredirects=0&d=1

As a perspective, PML is also contributing to the EC InterRisk project, and developed as such an experience in the deployment and adaptation of OGC services, in particular WMS, WFS, WPS, and the development of visualization web portals. "*The network of InterRisk services will be embedded in the ESA Service Support Environment (SSE), which will provide the underlying infrastructure for the InterRisk system. Incremental development of services and SSE components will be used to facilitate rapid feedback from users and service providers, leading to improved products and services that can be sustained in the future. InterRisk will be user driven, INSPIRE compliant and it will contribute to the implementation of GMES and GEOSS*".

Within the actual AIP-3 DHWG, together with the GIGAS project, the CEOS activities related to CAL/VAL and the QA4EO contribution to the GEOSS quality assurance strategy beginning with space-based observations and evaluating expansion to in-situ observations, we developed a comprehensive approach on Data Harmonization that integrates the end-to-end spectrum, from ground segments specific DH issues to end user access to harmonized data and products.
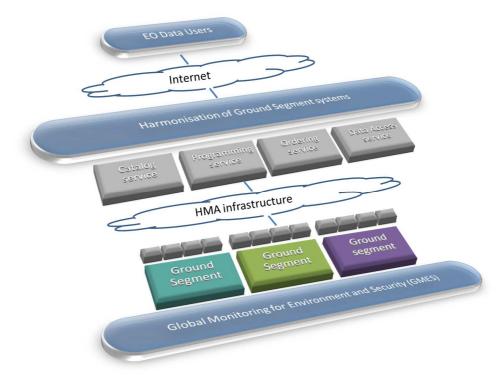

Figure 7.   Harmonization efforts between Ground Segments and Services to users (ESA & EC GMES)

Altogether, this general understanding addresses several key assets for inclusion in the next AIP-4 activities, thus providing a dedicated focus to the next generation on Ground Segment systems, as illustrated through different initiatives like the GMES HMA interoperability, the various EC projects related to Space and In-situ components integration, or the GMES and INSPIRE approach to process near real-time and deliver cross-border datasets.

3. AIP process for requirements gathering

The DHWG participants have addressed through their contributions a set of scenarios related to GEOSS communities. These scenarios describe how the application of Earth Observation data will be of societal benefit, and in summary present the scientific basis and the end goals.

This section contains descriptions of such scenarios and their references to some engineering Use Cases, providing solutions or illustrations to the Data Harmonization issues within GEOSS. It includes sub-sections to describe the Engineering use cases of the Scenarios. The general presentation of the AIP-3 Use Cases is contained in a separate AIP-3 ER.

3.1 Building a Service Oriented Architecture for GEOSS

The AIP-2 defined a reusable Service Oriented Architecture (SOA) process[8], based on operational scenarios and related technical use cases, to leverage the GEOSS Common infrastructure (GCI) and components in support of many SBA communities.

Scenarios show how GEOSS is applied in various Communities of Practice. Use Cases are reusable transverse technology approaches for implementing the scenario steps.

The AIP defined and piloted such a process for using and augmenting the GEOSS Common Infrastructure functionalities, supporting a global service oriented architecture progressively implemented through recognized International Standards and ad-hoc Interoperability Arrangements.
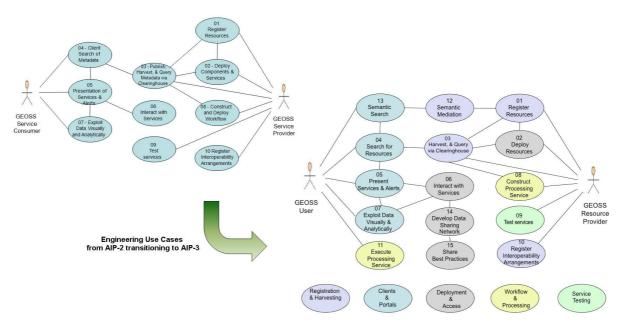

Figure 8.   Generic Engineering Use Cases (GEOSS AIP)

In the AIP-3 scenarios and use cases contributed to through the DHWG and presented hereafter, we emphasized on the elements that are leading to a set of requirements directly related to Data Harmonization. Recommendations to the GEOSS governance bodies are then formulated based on these requirements. DH requirements and recommendations are presented in the following chapters as the results of the AIP-3 Data Harmonization WG.

---

[8] http://www.ogcnetwork.net/pub/ogcnetwork/GEOSS/AIP3/pages/AIP-2_ER.html#summary

3.2 Contributed DH Scenarios & Use Cases

3.2.1 MyOcean's GMES Marine [GIGAS / MyOcean]

This GMES Marine scenario, based on the *oceanographic data assimilation problem*, is encapsulated in the GMES Marine Core Service (MyOcean, http://www.myocean.eu). The overall approach is to *analyze and predict geophysical properties of the ocean*, thus the scenario incorporates geophysical observations, geographic features, and environmental models. It explores Data Harmonization across the full breadth of the Information Viewpoint as presented in Figure 1 : throughout the processing pipeline in oceanographic data assimilation and analysis, different stages occur where either a feature view or a coverage view is predominant. The individual measurements may be considered features in their own right, or else as contributing to a discrete point coverage.

- **Model integration steps**
    1. Scientist selects input observations: in-situ observations are made of various oceanic parameters, and satellite observations are made of the sea surface (*coverage views, with sampling features*). The "observation pattern" applies to all MyOcean observation types, in-situ and remote – they are each measured with a specific instrument on a specific geophysical parameter, and each has a corresponding result. For example, observations can be a "Sampling down the water column" (with result a temperature coverage over a vertical profile), a "Fixed in-situ sampling" (with result a time series of water speed and direction), a "Remote sensing altimeter" (with result an along-track, or a gridded, sea surface height fields)
    2. Scientist exploits input observations *(related DH use case developed hereafter)*, taking individual measurements of a geophysical parameter (e.g. temperature) for integration within a numerical model, to perform an objective analysis of *oceanographic fields*
    3. Scientist performs run of numerical model to produce gridded analyses and forecasts (*coverage view*). The ocean is considered as a feature with properties corresponding to geophysical parameter fields (temperature, salinity, current speed, etc.) that vary throughout and in time. These may all be represented as *four-dimensional continuous coverage types* (i.e. values varying over three dimensional space, and in time). Additional coverage properties may be restricted to a smaller dimension – for instance the ocean 'surface height' varies horizontally and in time (*a three-dimensional coverage*).
- **Feature detection steps**
    4. Scientist performs feature detection based on analyzing a *geophysical parameter field*. For example, gradients in the oceanic temperature and/or salinity fields are often indicative of fronts and major current systems. The mesoscale variability of these features is often investigated by analyzing property fields as they change over time. Indeed, eddies and rings in the ocean are often classified according to their properties as seen in the temperature field (e.g. 'warm-core eddy').
    5. Scientist identifies the detected individual dynamic features, for example like fronts, upwelling zones, etc. (*feature view*).
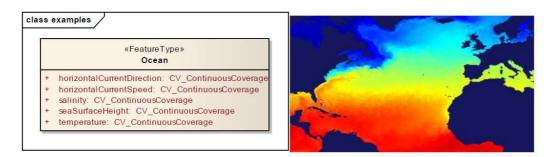

Figure 9.   Modeling the ocean as a feature with coverage-valued properties - temperature illustrated (GIGAS)

Engineering use case - UC 02 "Deploy resources with application schemas supporting transformation between meta-models"

Scenario step related to Data Harmonization	AIP Use Case description	Engineering Use Case requirements	GEOSS Registry reference
Step 2 - Scientist exploits input observations	N°02 - Deploy Resources This use case shall provide the conditions and steps to configure and deploy a component with associated service interfaces	**Requirement** (ontologies): Publish/standardize ontologies to create bridges between different vocabularies **Requirement** (rules): Develop ad-hoc rules for the representation of coverage functions in application schemas (not explicitly covered in the standards yet) Use of Observation and Measurement to convey differences between meta-models in appropriate application schemas. **Requirement** (delivery): Deliver with proper encoding depending on the application state / workflow step	See Annex of this document

Use Case's basic flow of events:

Title	Deploy resources with application schemas supporting transformation between meta-models, so that observations from different studies may be combined
Overview	Describes the service operation request and response encoding, and the parameters included. Provides best practice for service Registration in the GEOSS Common Infrastructure, ensuring proper service discovery, retrieval, and testing by the GEOSS communities of practice.
Actors and Interfaces	Service Provider GEOSS Common Infrastructure Registry
Initial Status and preconditions	Service Provider has a resource of interest for GEOSS. For example: a model, an observation or another process (grid transformation, workflow...)
Evolution (basic flow steps) [GIGAS]	Service Provider implements in his component the best available Web Service standard (OGC, W3C...) to make its resource available on GEOSS, based on best practices for that type of data, and the availability and familiarity of software tools. Service Provider configures or validates all the information about its Service interface as provided in the service Capabilities document: • Service Type, Version, Title and Abstract, Supported languages • Contact information (service provider PoC) • Supported service operations request and response encoding • Contents : Layers Names (normalized terms for M2M processing) and Titles (human readable) • Domains of validity : dimensions, units, range, scales, reference systems **Service Provider publishes ontologies** (a) Identify each entry within a vocabulary or an ontology using a URI. (b) Associate multiple 'labels' with each entry, allowing synonyms to be linked directly to an entry, and supporting multi-lingual requirements. (c) Include multiple relationships between entries, to encode richer semantics and also support inferences and equivalences. (d) Publish the ontology as a web resource, if possible through a formal query interface e.g. SPARQL. (e) Provide alternative representations of each concept, e.g. in HTML for human consumption, in RDF for machine reasoning. Service Provider publishes its Component and associated Service interfaces to the GEOSS registry. Metadata about the Service (ISO 19139) is generated automatically from the service, or additional information (e.g. metadata not found in the getCapabilities) needs to be added when registering the service
Post Condition	The following pieces of information about a Service must be available from the GEOSS registry: Service getCapabilities URL Service type (WMS, WCS, SPS, etc..) Type of procedure (Model, Sensor or Platform…) Input (e.g. Phenomena, configuration variables...) Output (e.g. Phenomena, file...) Ideally all of these pieces should have an associated URI (Uniform Resource Identifier)

3.2.2 The South Esk Hydrological Sensor Web [CSIRO]

This scenario is about surveying the South Esk river sub-catchments in North-Eastern Tasmania, a mostly unregulated river experiencing a highly variable climate and flow regime. The « Integrated Water Information Systems Theme » is providing near real-time situation awareness of river flow in regional catchments. Two operational outcomes are envisioned through this decision-aid scenario: apply water restrictions so that water supply surety is maintained, and announce flood take during peak flow events.

- **Integrate sensor observations and harmonize data**
  1. Acquire input observation data : rainfall, water level, climate
  2. The Kepler proxy system processes gridded rainfall surfaces based on the rainfall gauge observations provided through various CSIRO partner's SOS instances.
- **Model gridded rainfall surfaces and compute forecasts**
  3. The gridded rainfall surfaces produced in Kepler feed into a semi-distributed flow forecast model (set up by Hydro Tasmania (HT) Consulting).
  4. A Flow forecast model generates hourly Riverflow forecasts at the DPIPWE and HT stream-gauge locations.
  5. A Weather prediction model generates 48-hour Rainfall forecasts at a one-kilometre grid-cell resolution (CSIRO Cubic Conformal Atmosphere Model, CCAM).
- **Publish results**
  6. Short term riverflow forecasts are published via a dedicated 'model prediction' SOS service.
  7. Weather rainfall forecasts are exposed via a dedicated 'rainfall forecast' OpenDAP service.
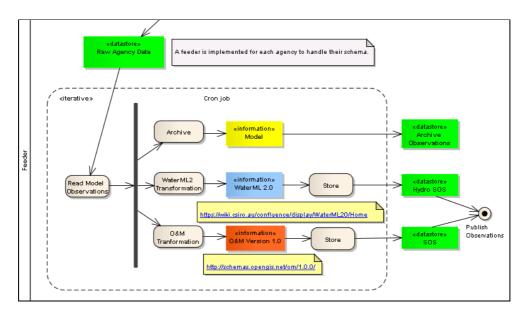

Figure 10.   Flow forecast data collection activity diagram - Feeder component (CSIRO)

Note: CSIRO is also developing a real-time quality assurance approach, as part of the provenance management framework for Hydrological Sensor Webs. This would lead to uncertainty management requirements too – all the process introducing additional uncertainty, and several bits of information that need to be combined.

Engineering use case - UC 02 "Deploy resources profiling SWE information models and web service interfaces"

Scenario step related to Data Harmonization	AIP Use Case description	Engineering Use Case requirements	GEOSS Registry reference
Step 2 - Kepler scientific workflow system, supporting integration, harmonization and encoding tasks	N°02 - Deploy Resources This use case shall provide the conditions and steps to configure and deploy a component with associated service interfaces	**Requirement** (information quality) : Flow forecast model needs stream gauges in each sub-catchment to serve as points-of-truth for calibration **Requirement** (data transformation): Harmonize time-series in observations provided through various partner's services. Fix basic errors such as gaps in the time-series and data spikes **Requirement** (application schemas): HydroSOS output of **Rainfall surface** shall support time-series in **WaterML v2.0** **Requirement** (application schemas): SOS output of **Riverflow** forecast model shall support **O&M v1.0** output encoding	See Annex of this document

Use Case's basic flow of events:

Title	Deploy resources profiling SWE information models and web service interfaces.
Overview	Describes the service operation request and response encoding, and the included parameters. Describes application schemas supporting data models to allow correct encoding and transmission of water information. Provides best practice for service Registration in the GEOSS Common Infrastructure, ensuring proper service discovery, retrieval, and testing by the GEOSS communities of practice.
Actors and Interfaces	A scientist user shall retrieve availability of spatio-temporal subsets, discover product and service metadata, or retrieve spatio-temporal subsets (including time series) to be visualized. The Hydrological Sensor Web provides an interoperability layer, built over existing sensing infrastructure operated by partner agencies in the South Esk river catchment. South Esk SOS for data access implies profiling SWE information models and web service interfaces to allow correct encoding and transmission depending on the SOS function requested. Observations from different partners may be exploited for distinct decision making goals: apply water restrictions, or announce flood take.
Initial Status and preconditions	Partner agencies have agreed to allow public access to their sensor observations via SOS interfaces.
Evolution (basic flow steps) [CSIRO]	Service Provider implements in his component the best available Web Service standard (OGC, W3C...) to make its resource available on GEOSS, based on best practices for that type of data, and the availability and familiarity of software tools. Service Provider configures or validates all the information about its Service interface as provided in the service Capabilities document: • Service Type, Version, Title and Abstract, Supported languages • Contact information (service provider PoC) • Supported service operations request and response encoding • Contents : Layers Names (normalized terms for M2M processing) and Titles (human readable) • Domains of validity : dimensions, units, range, scales, reference systems **Service Provider publishes service's application schema :** WaterML v2.0, O&M v1.0, and OpenDAP netCDF for numerical weather predictions Service Provider publishes its Component and associated Service interfaces to the GEOSS registry. Metadata about the Service (ISO 19139) is generated automatically from the service, or additional information (e.g. metadata not found in the getCapabilities) needs to be added when registering the service.
Post Condition	The following pieces of information about a Service must be available from the GEOSS registry: Service getCapabilities URL Service type (SOS, etc..) Type of procedure (Model, Sensor or Platform…) Input (e.g. Phenomena, configuration variables...) Output (e.g. Phenomena, file...) Ideally all of these pieces should have an associated URI (Uniform Resource Identifier)

3.2.3 Relationships between physical and biological variables [PML / NETMAR]

This scenario supports scientific researchers wishing to identify and use long term time series in order to quantify ecosystem responses to natural variability, climate change or the impact of anthropogenic activities (cf. NETMAR Case study). Examples may include *comparing* long term change in zooplankton concentration to water temperature or *relating* optical properties to chlorophyll concentration. It also supports operational users, *comparing*, in near real time, contemporary satellite and in-situ data in order to provide input to *water quality monitoring systems*, for example, on phytoplankton chlorophyll concentration.

- **Discovery steps**: the Scientist will use discovery tools to find datasets for the desired parameter within the specified geographical and temporal bounds. Tools will use semantic metadata and mappings to identify (and match) parameters named in different vocabularies.
    1. The Scientist selects the precise geographic area of interest
    2. The Scientist selects the time period of interest
    3. EUMIS displays a list of the in-situ data available through an existing database linked to an existing web feature server (WFS): e.g. in-situ measurements of phytoplankton or physical properties from time series either in single locations (such as the L4 station off Plymouth) or along track series (such as, for example, ships of opportunity or ferrybox systems).
    4. The Scientist selects the relevant in-situ data parameters
    5. EUMIS displays a list of the EO data layers available that are compatible with selected parameter: e.g. satellite time series of optical properties, chlorophyll-a concentrations, sea surface temperature and sea-surface elevation.
    6. The Scientist selects the relevant EO data layer
- **Build workflow**: the Scientist will use service chaining editor to build a workflow to merge the 2 datasets (in-situ & EO) and generate a statistical analysis of the result.
    7. EUMIS displays a list of available processes (based on *semantic metadata* attached to dataset)
    8. The Scientist selects the initial process (e.g. merge data)
    9. EUMIS displays a list of available processes (based on *semantic metadata* attached to process output)
    10. The Scientist selects a further process as above (e.g. run statistical comparison)
- **Run workflow steps**: the Scientist will use the service chaining manager to run the workflow. *Uncertainty information*, held in the metadata, will be passed through the chain and included in the output.
    11. The Scientist runs the workflow, providing the datasets identified in steps 4 & 6 as inputs
    12. The Scientist downloads the output from the whole workflow and, optionally, intermediate results produced during the processing
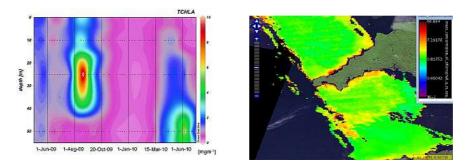

Figure 11.   Chlorophyll-a data from in-situ time series & MODIS instrument (PML)

Engineering use case - UC 08 "Construct Processing Services for combining in situ satellite and modeling data"

Scenario step related to Data Harmonization	AIP Use Case description	Engineering Use Case requirements	GEOSS Registry reference
Steps 7 to 10 – Build workflow	N°08 – Construct & Deploy Workflow This use case shall provide the conditions and steps to configure and deploy workflows, consisting of one or more services	**Requirement** (application schemas): User shall extract values for a given location or area and time from either in situ or satellite data, and compare with model hindcasts allowing for differences in resolution **Requirement** (application schemas): Results need to be returned in a format specified by the user allowing for in situ, satellite or model error characteristics	See Annex of this document

Use Case's basic flow of events:

Title	Construct Processing Services for combining in situ satellite and modeling data
Overview	Describes the walkthrough for designing, deploying, and executing a workflow, the related services operation request and response encodings, and the included parameters. Provides best practice for service Registration in the GEOSS Common Infrastructure, ensuring proper service discovery, retrieval, and testing by the GEOSS communities of practice.
Actors and Interfaces	Scientist, Developer of WPS and workflow chain Web Processing Service – This is a geospatial Web service with standard OGC interfaces.
Initial Status and preconditions	**Data** – Data or observations from sensors are rudimentary format and projection. Further processing is required to generate useful information for decision-makers. **Algorithms** – The algorithms may be in abstract description. The algorithms needs to be implemented as geospatial Web Processing Processes in order to re-use them in different workflows. **Standard Web Services** - Individual working Web services. For the specific use case of severe weather detection, we have a Web Feature Service that supports transaction, a Web Coverage Service for serving GOES data, a Web Processing Service for the severe weather event detection algorithms.
Evolution (basic flow steps) [PML/NETMAR]	Service Provider implements in his component the best available Web Service standard (OGC, W3C...) to make its resource available on GEOSS, based on best practices for that type of data, and the availability and familiarity of software tools. Service Provider configures or validates all the information about its Service interface as provided in the service Capabilities document: • Service Type, Version, Title and Abstract, Supported languages • Contact information (service provider PoC) • Supported service operations request and response encoding • Contents : Layers Names (normalized terms for M2M processing) and Titles (human readable) • Domains of validity : dimensions, units, range, scales, reference systems Service Provider publishes its Component and associated Service interfaces to the GEOSS registry. Metadata about the Service (ISO 19139) is generated automatically from the service, or additional information (e.g. metadata not found in the getCapabilities) needs to be added when registering the service.
Post Condition	The following pieces of information about a Service must be available from the GEOSS registry: Service getCapabilities URL Service type (SOS, etc..) Type of procedure (Model, Sensor or Platform…) Input (e.g. Phenomena, configuration variables...) Output (e.g. Phenomena, file...) Ideally all of these pieces should have an associated URI (Uniform Resource Identifier)

3.2.4 Uncertainty enabled pressure correction chain [ASTON / UncertWeb]

This scenario is a plausible scenario developed within the UncertWeb project to assess different mechanisms for propagating uncertainty in service chains. The essence of the concept is that a user wishes to correct a pressure measurement to sea-level pressure, taking into account the uncertainties in the process. This is not meant as a real world case, but rather a demonstrator for how uncertainty can be propagated and managed within a workflow in a service oriented architecture.

- **Harvest input data into the Model chain**
    1. The model chain harvests into an observations service (SOS) the user-contributed air pressure measurements (together with their uncertainties) done at station locations, taken from the Weather Underground sensor network for the UK.
- **Orchestrate the Model chain processing steps**
    2. For each of these measurement locations, surrounding elevation values are retrieved from a service over the Shuttle Radar Topography Mission data, noting the SRTM supplier's estimate of uncertainties on the elevation data.
    3. These values are passed to an automatic interpolation service (INTAMAP) which computes the predicted elevation of the station location in question, taking into account the noted elevation error.
    4. A correction service is used to correct the air pressure measurement to pressure at sea level, using the predicted elevation, and their uncertainty estimates.
- **Present results with uncertainty estimates**
    5. The model chain makes use of a variety of representations of uncertain values (e.g., simple statistics such as variance, and samples from a probability density function) to produce a summary of the distribution of possible output values that encapsulates intervening sources of uncertainty.
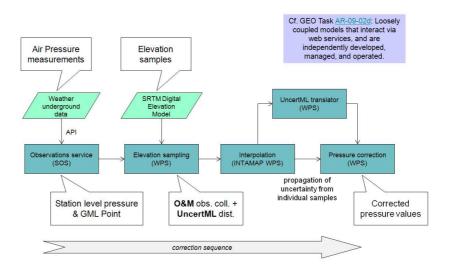

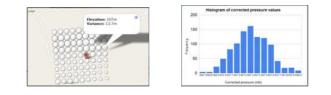*Figure 12.   Uncertainty enabled pressure correction chain (ASTON University)*

Figure 13.   Elevation sampling & Corrected pressure values (ASTON University)

Engineering use case - UC 08 "Construct Model Chain offering propagation of uncertainty"

Scenario step related to Data Harmonization	AIP Use Case description	Engineering Use Case requirements	GEOSS Registry reference
Steps 3 to 5 - (all WPS components)	N°08 – Construct & Deploy Workflow This use case shall provide the conditions and steps to configure and deploy workflows, consisting of one or more services	**Requirement** (quality): Model the uncertainties from in the inputs to the chain, as issues of data quality, and represented with UncertML **Requirement** (quality): Process an input taking into consideration the main sources of Uncertainty **Requirement** (data transformation): Convert from a representations of uncertainty to another where this is required within the chain	See Annex of this document

Use Case's basic flow of events:

Title	Construct Model Chain offering propagation of uncertainty
Overview	Describes the walkthrough for designing, deploying, and executing a workflow, the related services operation request and response encodings, and the included parameters. Provides best practice for service Registration in the GEOSS Common Infrastructure, ensuring proper service discovery, retrieval, and testing by the GEOSS communities of practice.
Actors and Interfaces	Scientist, WPS and workflow developer Web Processing Service – This is a geospatial Web service with standard OGC-specific interfaces.
Initial Status and preconditions	**Data** – Data or observations from sensors are rudimentary format and projection. Further processing is required to generate useful information for decision-makers. **Algorithms** – The algorithms may be in abstract description. The algorithms needs to be implemented as geospatial Web Processing Processes in order to re-use them in different workflows. **Standard Web Services** - Individual working Web services. For the specific use case of severe weather detection, we have a Web Feature Service that supports transaction, a Web Coverage Service for serving GOES data, a Web Processing Service for the severe weather event detection algorithms.
Evolution (basic flow steps) [ASTON/UncertWeb]	Service Provider implements in his component the best available Web Service standard (OGC, W3C...) to make its resource available on GEOSS, based on best practices for that type of data, and the availability and familiarity of software tools. Service Provider configures or validates all the information about its Service interface as provided in the service Capabilities document: • Service Type, Version, Title and Abstract, Supported languages • Contact information (service provider PoC) • Supported service operations request and response encoding • Contents : Layers Names (normalized terms for M2M processing) and Titles (human readable) • Domains of validity : dimensions, units, range, scales, reference systems Service Provider publishes its Component and associated Service interfaces to the GEOSS registry. Metadata about the Service (ISO 19139) is generated automatically from the service, or additional information (e.g. metadata not found in the getCapabilities) needs to be added when registering the service. Additional information can be found in the UncertWeb requirements for service chains in the biodiversity, agricultural crop modeling and air quality forecasting application domains[9].
Post Condition	The following pieces of information about a Service must be available from the GEOSS registry: Service getCapabilities URL Service type (SOS, etc..) Type of procedure (Model, Sensor or Platform…) Input (e.g. Phenomena, configuration variables...) Output (e.g. Phenomena, file...) Ideally all of these pieces should have an associated URI (Uniform Resource Identifier)

---

[9] http://www.uncertweb.org/uploads/deliverables/b8aea972456fbcd80aeebb181098bc18114615e0.pdf

As a conclusion, the working group noticed from the examples above that one key element in the descriptors that is often missing, despite all meant to be implied within the processes described, is when any data is ingested, it must be associated with an uncertainty statement or QI. Without this, none of the following steps have all their meaning and any delivered result similarly won't have all confidence on its viability.

➔ The DHWG recommends AIP to strengthen its scenario and use case templates, so to include guidance and support with regard to the expression of uncertainty on measurements and processing.

3.2.5 Coordination with other AIP-3 working groups

During the AIP-3 process, some Data Harmonization technical common group have been discussed with other groups, having developed some approaches or prototypes in this regard, notably related to the uncertainty management topic.

AIP **Water Management** coordinated with Will Pozzi (Center for Research in Environment and Water)
- Coordination occurred while looking at the issue of cross-boundary databases containing groundwater aquifer water level and water storage information, along with semantic integration to overcome semantic heterogeneity (i.e., data harmonization)
- Discussions on a repository of ontologies, where Water ontologies should be stored, and considering the option of the GEOSS Best Practices Wiki

AIP **Air Quality** coordinated with Erin Robinson / Edzer Pebersma (University of Munster)
- Short videos that demonstrate how the Aguila software, enabling the user to interact with spatial data that comes as probability distributions (coming from the intamap wps) : http://intamap.org/videos.php
- The workflow features an interpolation function, needed to estimate the concentrations of polluants at some points of interest. Interpolated concentrations are thus estimates. Estimation errors need also to be communicated. An interpolation Web Processing Service is demonstrated with several uncertainty types, like quantiles, which shall be returned, and can be specified in the request.
- The AIP-3 Demo Video on Air Quality also refers to this tool for visualization of uncertainty. See the online video at 3:15 : http://www.ogcnetwork.net/GeossArchProg

AIP **Renewable Energy** coordinated with Lionel Ménard (Mines Paris Tech)
- Informative exchanges as the EnerGEO portal (ESRI GeoPortal) aims to provide with extension for Data Quality Management

4. Data Harmonization - Requirements and recommendations

This section provides a review of the Data Harmonization requirements that were gathered out of the demonstrated Use Cases. It provides a set of recommendations for addressing the GEOSS Data Harmonization requirements that were gathered through the demonstrated use cases. It is envisioned as a synthesis of the AIP discussions held on 'which encoding standards, for which purposes in the GEOSS DH context'. These recommendations would address primarily the established GCI components for the standards registration tools, the content-oriented discovery facility and the web-based assess and access mechanisms.

4.1 Sensor Calibration and Validation, sensors inter-calibrations

The CEOS Working Group on Calibration and Validation (WGCV) has proposed an approach for GEOSS CAL/VAL requirements. Such process supports the quantitative definition of a system's responses to known, controlled signal inputs, and the assessment, by independent means, of the quality of the data products derived from the system outputs.

Such data products have been discussed and referred to in the AIP-3 DHWG meeting minutes as "level 1 products", or "Products for primary dissemination formats". As a corollary, level 2 or 3 products (or products for secondary dissemination formats) do require specific attention within the GEO tasks, due to the expanding thematic usages and processing of EO products towards GEO Societal Benefit Areas.

These notions are relevant to distinguish and leverage the Data Harmonization strategies, depending on the physical meaning of observations and measurements, and the processing lineage applied to such values and the subsequent quality statements. They are discussed hereafter as recommendations.

➔ The DHWG acknowledges the need for more outreach of Inter-calibration domain towards data users that need to integrate multiple data sources in their applications, delineating accurate information products from operational environmental satellites. Coordination with the Global SpaceBased InterCalibration System (GSICS) appears relevant here, as *'calibration' ties a satellite instrument's readings to physical quantities such as units of radiant energy, [so] the 'Inter-calibration' of instruments achieves comparability of measurements from different instruments*.

The CSIRO Tasmanian ICT Centre has developed the Hydrological Sensor Web (HSW) based on OGC-SWE standards. Near real-time hydrologic observations and flow forecasting are published and accessed through the OGC Sensor Observation Service (SOS). The contributed "South Esk Catchment" scenario, described in section 3.2.2, presented a requirement for usage of points-of-truth for calibration, as a number of partner agencies operate sensor assets in the river sub-catchments. This requirement has linkages with uncertainty management as well, that was emphasized during the Working Group telecons, and would provide material for additional AIP work. One typical aim supported by such an inter-calibration is the foreseen ability to implement another workflow that fuses rainfall forecasts with in-situ rain gauge data, to produce an enhanced rainfall surface, that can be ingested into the a semi-distributed flow forecast model.

	Rainfall	Climate	Water Level
BOM	6	1	5
CSIRO	7	7	
DPIPWE	4		6
FT		2	
HT	1		1
Total	18	10	12

Figure 14.   Observed phenomena in South Esk Catchment (CSIRO)

Figure 15a resumes the generation process of flow forecasting results: firstly, rainfall observations are collected from different sensor sites, owned by different agencies, and stored in databases. The observations are published on the Hydrological Sensor Web via SOS. A Kepler workflow obtains rainfall observations from SOS and generates the gridded rainfall surface. Then a forecast model consumes the gridded rainfall data, and produces flow forecasts. Finally, the forecasting results are published onto the HSW through SOS.
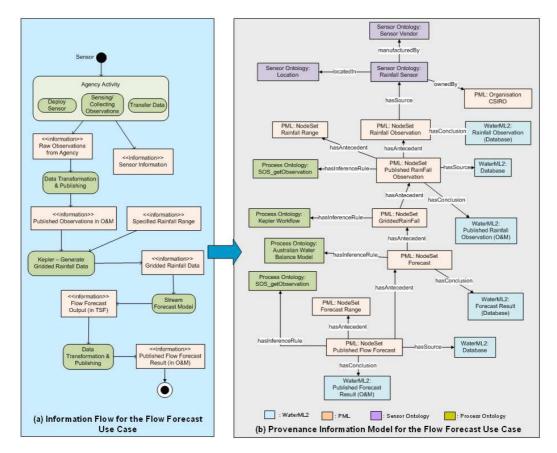
Figure 15.   Provenance information model (CSIRO)

It can be seen that different agencies are involved in producing flow forecasting. So, for this generation process, a provenance information model has been developed, which is demonstrated in Figure 15b. Three sets of ontologies have been adopted, which are the Sensor Ontology, the WaterML2 Ontology and the Process Ontology, to describe information/ knowledge in the sensor domain, the water domain and the data processing domain, respectively. Then, the Proof Markup Language (pML) is used to describe the generation processes of information products, and link multi-domain ontologies together. This allows tracking the lifecycle of hydrologic data products, as well as record-related factors that may impact on data qualities, e.g., sensor setting, model calibration.

4.2 EO Data Products Specifications

By 'EO data products' we intend to discuss here the Earth Observations data resources that are encompassing a quite large extent of acquisitions types: near-real time, forecast, historical, composite, interpolated…

Harmonization of such EO data products thus calls for a detailed knowledge of the functional and technical specifications of these products, within each of these categories, which was not possible to conduct during the process.

Nevertheless, through the GIGAS contribution, a way forward was proposed to advance harmonized data models, encompassing 'features', 'coverages' and 'observations'. It is based on the following assumption:

- Key to scientific interest in 'features' (ocean's tides, atmosphere poluants, geologic layers…) is generally the fact that the value of one or more feature's properties varies spatially within the scope of the studied feature.

- For historical reasons related to the development of GIS tools and applications, simple types of features, in which each property has a single or constant value, can be generally thought not to apply to coverages.

- However, spatially varying properties are allowed by the General Feature Model, and would be represented as a 'coverage'.

For coverages, "*The chief requirement for making their feature-type explicit comes from cross-domain applications, where a scope that is implicit within the original community needs to be made explicit across community boundaries*". In response, the GIGAS contribution to AIP-3 emphasized a proposed unifying model – *Sampling Coverage Observations* – for spatially varying feature properties. We resume hereafter the four-stages approach to address this topic.

Stage 1) Spatially varying properties are allowed by the General Feature Model, and would be represented as a 'coverage'. A coverage function could e.g. be represented as a GF_Operation that maps values from the spatiotemporal domain to the range values.
Feature of interest:
- Ocean

Feature (ocean) properties ➔ (modeling) n-dimensional continuous coverage types, namely:
- Temperature, salinity, current speed…

Stage 2) Individual measurements may be considered features in their own right, or else as contributing to a discrete point coverage. Various pieces of information extracted or interpreted from the coverages will characterize a detected feature.
Measurement:
- Coverage type

Sensed phenomena (coverage type) detected features ➔ (interpretation) feature properties, namely:
- Location, spectral characteristics, gradients, orientation…

Stage 3) We have different views on *measurements* but 'linked' by the *observations*. Where the observed property varies within the scope of the feature-of-interest, then the result should be expressed as a coverage (Discrete Coverage Observation). In all these cases, it is not the full ocean which is being observed, but rather the portion of the ocean defined by the sampling manifold; thus the 'feature of interest' is a spatial 'sampling feature'.
Results represent an *estimate* of the value of a feature property, and therefore should also include a measure of the quality of the estimate (as indicated in the O&M model).

Feature of interest ⬅➔ Data capture procedure (instrument or algorithm) **⬅➔ Observed Coverage property**

➔ Lists of observed properties (e.g. the CEOS Geophysical Parameters [CEOS/WMO (2009)]) shouldn't be populated by members that mix the actual observed-property with a specific feature-of-interest and/or the observation-procedure. The DHWG recommends promoting the observed property as a key to observation semantics: observations from different studies may be combined, providing the observed property is the same, so this may be a key data discovery parameter.
   o **sea-surface-dynamic-height** is height (a property) of the sea-surface (the feature-of-interest) measured dynamically (the procedure)
   o **temperature of tropopause** mixes observed property (temperature) and feature-of-interest (tropopause)
   o similarly for **sea-ice thickness** (CEOS/WMO Parameter 065)

Stage 4) A unifying model is then proposed based on the following equivalence. Some consistency requirements in this process are identified:

- the coverage result must have a range type consistent with the observed property
- the spatial domain of the coverage result must conform to the sampling geometry
- the sampling time must define the temporal elements of the coverage result, and should be reflected in the observation's 'phenomenon time'

Spatial Sampling Feature ←→ **Sampling Coverage Observation** ←→ Discrete Coverage Observation

➔ The DHWG recommends the SIF to register and promote ISO/DIS 19156 Observations & Measurements. This standard provides a basic model for sampling regimes, to support cross-domain communication, through some common names defining the spatial sampling features, linked from various disciplines. (Note that currently, a query like '19156', when submitted to the Standards and Interoperability arrangements Registry (SIR), does not provide a satisfying result)
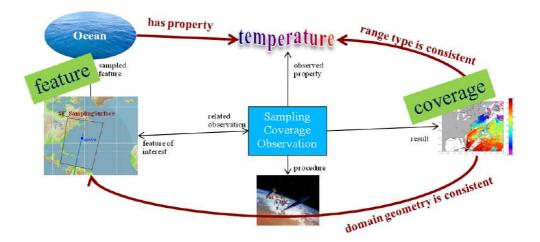
Figure 16.   Role of O&M "Sampling Coverage Observation" in reconciling separate information views (GIGAS)

The information viewpoints are entirely complementary, and each can be described in terms of the other, so transformation between meta-models can be achieved. For example, contributing features from a 'vector' representation will provide values at irregular locations used to constrain interpolation or modeling a regular gridded coverage. Some recent standards developments are also of interest in this context, like the ISO 19131, a template for data product specifications, and the ISO 19123 and the OGC application schema for coverages (OGC 09-146r15), adding a rangeType element to the GML coverage types, including a description of the range semantics.

4.3 Environmental Data and Models

Key to encoding our **understanding about reality and observations** are models. Models take several forms, and are viewed in different ways in different communities. For example physicists build models of reality, which are often called physical laws. In other disciplines, models are representations of the modeler's beliefs about reality, but are almost always recognized to be approximations, and thus subject to uncertainty. A good example, currently very much under scientific, political and public debate is climate modeling.

The issue of **model uncertainty** (also called model error or model discrepancy) and the relations between models, observations and reality is a very complex subject, but in the context of this document the focus is on **data uncertainty**, which is conceptually somewhat simpler. However, as many, if not most, data products are processed, we need to consider the processing model uncertainty when dealing with level 2 or higher products (that is, processed observations or derived products). Furthermore, when describing reality we need to select and define an

appropriate **range of spatial and temporal scales**. Without a clear definition of what is meant by reality it is impossible to define a notion of uncertainty and thus make rational links between different models or observations of the same thing.

Consider an example: a user is interested in obtaining an estimate of nitrogen dioxide concentration over a large city. Data is available from in-situ sensors (with different spatial and temporal supports), and from satellite based retrievals from the GOME instrument. These different sets of observations represent very different samplings of reality, with given spatial and temporal support, and associated uncertainty due to measurement , processing , etc. The only consistent way to combine these is through having a common model for the uncertainties of each, with respect to reality, or some pre-defined and agreed approximation to it, defined over a certain spatial and temporal scale. This is **information interoperability**, and allows the data to be used rationally across many applications.

➔ The DHWG recommends the SIF to promote additional guidance on the GEOSS Best Practices Wiki, for consistent means to define information interoperability across in-situ and satellite-based observations (GMES use cases). The AIP in coordination with QA4EO shall help push things further and qualify the mechanisms (standards and tools) to allow this to happen.

4.4 Vocabularies, ontologies and Registers supporting semantic interoperability

4.4.1 GEOSS registries

A demonstration video from the AIP-3 Data Harmonization WG contribution (CSIRO Use Case, registration of Components and Services) was released to present the current CSR registration process capabilities and limitations[10]. The supporting slides are also available from the AIP-3 collaboration site[11]. Hereafter, we discuss the issue of vocabularies, ontologies and schemas that can only be registered as "other" at this point in the CSR.

During the AIP-3 process, CSIRO expressed interest in contributing to the development of service registries, particularly in aspects relating to data provenance, including encoding standards for provenance (e.g. Open Provenance Model, Proof Mark-Up Language) and technology for harvesting provenance information at service interfaces.

A registry is an information system on which a register is maintained; whereas, a register is a set of files containing identifiers assigned to items with descriptions of the associated items (definitions from ISO 19135).
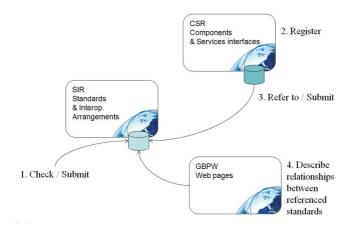
Figure 17.   Describing standards in the GEOSS Registries – SIR, CSR and BPW

---

[10] http://www.ogcnetwork.net/pub/ogcnetwork/GEOSS/temp/AIP-3/pages/Demo.html, then select the entry 'Data Harmonization'

[11] https://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-resources/Data_harmonzation_Demo_capture.ppt?attredirects=0&d=1

A registry provides access to the registers that it maintains. The GEOSS registries and their current owners are:

- The **GEOSS Components and Services Registry** (CSR) is similar to a library catalogue. All of the governments and organizations that contribute components and services to GEOSS provide essential details about the name, contents, and management of their contribution. This assists the Clearinghouse, and ultimately the user, to identify the GEOSS resources that may be of interest.
- The **GEOSS Standards and Interoperability Registry** (SIR) enables contributors to GEOSS to configure their systems so that they can share information with other systems. This Registry is vital to the ability of GEOSS to function as a true system of systems and to provide integrated and crosscutting information and services. Contributors can also share ideas and proposals informally via the associated Standards and Interoperability Forum.
- The **GEOSS Best Practices Wiki** (BPW) provides the GEOSS community with a means to propose, discuss and converge upon best practices in all fields of earth observation.
- The **GEOSS User Requirements Registry** (URR) will publish User Types, Activities and Requirements to support identification of linkages between those items as well as down a value chain/network that benefit from an Activity or Requirement. This registry is under development and is not discussed in this report.

Discussions occurred within DHWG for the addition of the QA4EO Framework and guideline documents to the SIR, under the Quality Assurance / Quality Control category (indicating these entries as a best practices). It was noticeable that in two distinct but similar cases (REST / QA4EO), there is an obvious need to cross-link SIR entries that pertain to 'best practices' to some related technical standards (e.g. REST with HTTP/1.1, or QA4EO with UncertML…). How would these linkages be achieved within the SIR is still undergoing. Although REST is in the SIR, the SIF is trying to stay away from using the SIR for very broad horizontal standards (HTTP, HTML, XML, etc.). Whilst the QA4EO documents themselves as an overarching generic framework; principles and guidance are really standards and should sit in SIR, their translation and adaptation into sector or SBA specific best practices (including those specific to the space sector) would logically sit in the BPW. The cross-linking should then take place between the BPW and the SIR, as well as between the BPW and the CSR. Unfortunately, this functionality does not currently exist. It has been talked about and categorized as future plans.

➔ The DHWG recommends the SIF to provide additional guidance directly on the CSR website and the GEOSS Best Practices Wiki, for consistent means to cross link best practices documents and some related implementation standards, which may arise as a chosen interoperability arrangement.

4.4.2 Sensor Ontologies

One illustration was given through CSIRO contribution. CSIRO plans to modify SOS instances to replace the use of SensorML, with the Sensor Ontology being developed by the **W3C Semantic Sensor Network Incubator Group (SSNIG)**[12]. At present, there is no capability within the CSR process to register specific ontologies. The interim solution is to register specific ontologies using the GEOSS Best Practices Wiki (cf. Figure 17). It is noted that the registration process does permit the specification of SPARQL query results as a supporting *data format*.

Figure 18.   GEOSS Component resources – Possible category for Ontologies registration

---

[12] http://www.w3.org/2005/Incubator/ssn/

From this particular return of experience, some general conclusions can be made in order to leverage the existing tools and enhance the GEOSS Common Infrastructure 'user experience'.

➔ The DHWG recommends the GCITF to provide additional guidance directly on the CSR website and the GEOSS Best Practices Wiki, for consistent means to declare an online ontology, as a simple example of a 'Registry', thus acknowledged under the CSR category "Catalog, Registry or Metadata Collection".

More related to governance issues, we also emphasize that interoperability is not a static achievement, thus requiring support for some degree of agility : as organizations evolve and build new partnerships, interoperability arrangements must be found or consolidated, and the tooling for their implementation readily available to the software development communities.

➔ DHWG recommends wider access and use within GEOSS of some common data-related resources, to be provided online, and in a reusable form: conceptual schemas, encoding schemas (e.g. XML schemas and Schematron), controlled vocabularies, codelists, and glossaries. The dependencies between these components shall be described and managed under stated governance rules.

For example, the OGC Meteorology & Oceanography Domain Working Group underlined that ISO 19156 Observations and Measurements time values can be applied to create a standard vocabulary for time.

In the context of GEOSS observations, it would be expected to find references to items in remote registers, following this pattern if the encoding is GML, in the following locations (for instance): Coverage/rangeType Observation/observedProperty Observation/procedure Specimen/processingDetails.

4.4.3 General ontology content encoding

The current state-of-the-art for general ontology content is to use RDF, typically through the OWL or SKOS applications. OWL is a full ontology representation, with a profile that supports Description Logic. SKOS is an OWL application oriented around 'concepts' and 'concept schemes' with a built-in set of relationships that map SKOS concepts to a rigorous version of the conventional thesaurus model. For specialized vocabularies, such as units of measure and coordinate reference systems which include specific mathematical structures as part of the concept definition, there may be specialized data formats in common use. For example, GML includes specific modules for coordinate reference systems, and SensorML is designed for procedures used for observations. A vocabulary shall be also external from the local application schema, so that several terms or labels can be associated with each item, which is accessed primarily using a URI. In GML applications, the Xlink W3C Standard for linking is the primary mechanism for carrying references.

The need is to publish 'standard vocabularies' as ontologies within e.g. remote registers of reference. Example of the "ISO 19156 Observations and Measurements" time values, as applied to create a standard vocabulary for time (from OGC MetOc DWG). At present, there is no capability within the GCO registration process to register specific ontologies. The interim solution is to register specific ontologies using the Best Practices Wiki. In order to reconcile, or synchronize, Data Products *structure* and *content*, they need to be managed within a same control body. Examples of some well-governed vocabularies discusses within the AIP-3 DH WG:

- CEOS Geophysical Parameters
- CUAHSI Parameters
- WaterML
- OGC SensorML
- UncertML
- W3C SSNIG Sensor Ontology…

➔ The DHWG recommends the SIF initiating a process for the identification of 'control bodies' in charge of both standards and terminology within the GEOSS community.

4.5 Mark-up languages for Data and Metadata

4.5.1 Geography Markup Language (GML)

4.5.1.1 Coverages

Some applications have focus on "property-fields", or coverages, where the property values are associated with spatial position. The coverage model from ISO 19123 defines a classical spatial function (especially through a domainExtent and a rangeType), which maps from positions in the spatio-temporal domain to values in a range. The definition includes also a number of operations. This is because a spatial function provides values everywhere (in the domain), so the key to implementation is provision of methods to access the value at each valid location.

Although ISO 19123 has a "rangeType" attribute to describe the range of the coverage, it is intended for describing the range record structure rather than its semantics, and supports nothing more than a name to describe a range component. The OGC application schema for coverages (OGC 09-146r1) adds a rangeType element to the GML coverage types, including a description of the range semantics in addition to its record structure. This allows the semantics of a stand-alone coverage instance to be understood.

The GIGAS report contributed to AIP-3 also describes coverage functions in the General Feature Model.

4.5.1.2 Features

Geographers/topographers often prefer to regard the world as composed of identifiable structures, with objective properties. Technically, this approach adopts the 'General Feature Model' (ISO 19109, also published as OGC Abstract Specification Topic 5) as an object-like conceptual schema language. This is the approach adopted by the European INSPIRE spatial data infrastructure (EC Directive 2007/2/EC). This is a classical approach to modeling where the 'domain of discourse' is defined primarily on the basis of the 'feature-types' of interest. Each 'feature-type' is defined by a characteristic set of properties required for the application. Instances are discrete objects with a defined type. Feature properties have arbitrary precision, and spatial properties are described using coordinates, hence 'vector'. Because the current version of ISO 19109 predates ISO 19123, it does not explicitly address coverage functions in the General Feature Model.

4.5.2 Observations and Measurements Markup Language (O&M)

A CSIRO software application, accessible online[13], has also been contributed to the GEOSS CSR. It is based on a Google Maps thin client, for visualizing observations provided by SOS (O&M v1.0 output) and HydroSOS (WaterML v2.0 output, an O&M profile), as described in the Scenario and Use Case in section 3.2.2.
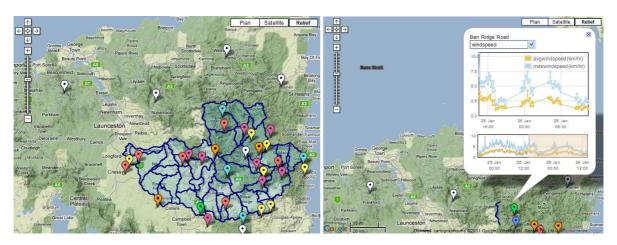
Figure 19.   South Esk Hydrological Sensor Web – GMaps thin client (CSIRO)

---

[13] http://www.csiro.au/sensorweb/au.csiro.OgcThinClient/OgcThinClient.html

This application illustrates the state of the art when accessing to O&M output encodings from a SOS Service. Markers are color-coding the different agencies operating sensor assets, where recent sensor time series data can be rendered for viewing. Observations are covering diverse phenomena like humidity, atmospheric pressure, wind speed and direction, barometric trend, air temperature, rainfall, direct and reflected solar radiation, evaporation, water level and water discharge. In this approach, "*the Sensor Web is an advanced type of spatial data infrastructure in which different sensor assets can be combined, to create a macro-instrument with massive sensing capability. This macro-instrument can be instantiated in many ways, to achieve multi-modal observations across different spatial and temporal scales*".

It is foreseen that applications of this kind will extend their approach in such a way as presented in the GIGAS MyOcean scenario. By explicitly modeling an observed property which varies (spatially or temporally) giving a coverage result, together with observation on a 'sampling feature', this may develop a unifying model that reconciles coverage-based and feature-based views. Then, observations from different studies may be combined, providing the observed property is the same, so this may be a key **data discovery** parameter, to relate with the discussion on discovery and assessment in the next chapter on the "GEOSS Common Record".

4.5.3 Water Markup Language (WaterML)

The first version of WaterML was aimed to encode the semantics of hydrologic observation discovery and retrieval, and implement water data services in a way that is both generic and unambiguous across different data providers. WaterML 1.0 follows and shares terminology with the information model of ODM (Observation Data Model), described at http://www.cuahsi.org/his/odm.html. Conformance with OGC specifications was not the goal of this initial version.

The goal of WaterML v2 is harmonization of WaterML with the Observations and Measurements (O&M) data model and encoding and Sensor Observation Service (SOS) interface. O&M and SOS are from Open Geospatial Consortium (OGC), and O&M is on track to become an International Standard ISO 19156. Harmonization would enable (a) delivery and consumption of water observations data using systems developed to conform to the SOS standard (b) integration of water observations data with information from closely related domains in environmental sciences, such as geology and meteorology, where OGC-conformant systems are being deployed (e.g. OneGeology).

4.5.4 Sensor Markup Language (sensorML)

SensorML provides a standardized date models and an associated XML encoding for the description of sensors and/or of measurement processes (encompassing dynamic and stationary platforms, in-situ and remote sensors).
While the CSIRO SOS will return observations using the observations and measurements 1.0.0 encoding standard, its DescribeSensor operation will return sensor information encoded using SensorML.
Despite the relevance of this standard for the Data Harmonization topic, there was no significant discussion thread of the OGC SensorML within the WG participants.

4.5.5 Uncertainty Markup Language (UncertML)

Out of the ASTON contributed scenario and its WPS engineering use case, the Working Group derived many discussion items and could link the UncertML topic with other considered items as the QA4EO framework, the EO Data Products specifications and encodings (through the ability to include the UncertML schema in existing XML representations like a GML Application schema or the O&M XML encoding), as well as the ISO standards for Data Quality encoding.

The Uncertainty Markup Language (UncertML) is a simple conceptual model and associated XML encoding (and JSON and netCDF best practice encodings) for the transport and storage of information about uncertain quantities, with emphasis on quantitative representations based on probability theory. UncertML is not intended to be a solely geospatial standard, since the fundamental principle of UncertML (interoperable representation of probabilistic uncertainty) appeals to a wider set of application domains.

The concepts discussed in ISO19115 would suggest that UncertML is a realization of the "DQ_QuantitativeResult" element, although UncertML does not include units of measure, since these are encoded in different ways in different application domains, and are not considered in scope for UncertML, which is designed to be used with other standards such as O&M, NetCDF and GML. Note that UncertML focuses very clearly only the issue of uncertainty and is associated with a controlled vocabulary.

UncertML as a conceptual model

UncertML is a conceptual model and associated set of XML schema for representing probabilistic uncertainty. The conceptual model is based on the concept that uncertainty and the probabilistic representation of uncertainty is the natural mechanism for dealing with incomplete or erroneous data. The natural framework for integrating uncertain information from multiple sources is probabilistic modeling. Providing well defined probabilistic representations allows far greater integration of uncertain information in a principled and traceable manner, as outlined in QA4EO.

This lies at the heart of data integration and data fusion. The emphasis for quality information should be the utility of this information in subsequent processing of the data, not on whether we can create a very customized solution which happens to fit the very specific training and bias of the domain in which a 'novel' errorStatistic was defined. We should always aim to simplify and make generic cross cutting issues, and uncertainty is a perfect example of this. Like mathematics, the treatment of uncertainty is something universal (given one accepts the basic axioms, but this is part of being a scientist), so we should seek common solutions. UncertML is designed to be used within other standards, for example ISO19115 (and thus ISO19139, O&M 2.0, ...). The key principle behind the design of UncertML was that probabilistic approaches to addressing uncertainty are universal, so we did not want to bring in any domain specific information into UncertML. Thus UncertML 2.0 imports no external schema, maintains a simplicity that makes usage in a range of application simple, and has several encodings including XML, JSON and NetCDF.

UncertML takes a rigorous statistical approach, but does allow flexibility in representation of uncertain values (random quantities). Random quantities can be described using probability distributions (including mixture models and hierarchical distributions), realizations (from both random and deterministic sampling) or summary statistics (a quite complete and sufficient set of commonly used types is currently given, but this could be extended in the future under a defined governance arrangement).

UncertML encoding and interoperability

UncertML's ability encode uncertainty in many representations is important and allows a degree of interoperability between uncertainty representations. At the most complete specification a probability distribution completely describes the random quantity, and realizations and statistics can be derived from them automatically (using our developing Uncertainty Transformation Service, or the UncertML API, for some commonly used distribution types). Given sufficient realizations, it is also possible to automatically derive statistics. However going up the hierarchy is more complex, since statistics typically do not uniquely define a full probability distribution and thus automated conversion is not possible and additional information is required (or assumptions must be made).

We believe that the use of a common model and encoding for uncertainty will allow far greater interoperability in managing uncertainty than the use of strings (for errorStatistic) as advocated in ISO19115. If users provide reliable uncertainties which also have good resolution, described using UncertML this takes us closer to the aim of "information interoperability". UncertML can equally well represent an uncertain result, or the uncertainty estimates usually associated with metadata quality statements, e.g. in ISO19115.

One conclusion is that linking such uncertainty encoding within an ISO19115 metadata record would leverage some extended search and discovery use cases addressing 'compatible datasets' in the realm of their fit for purpose. This would then impact some on the fields defined in the GEOSS Common Record, and the GEOSS Clearinghouse data model and search interface, which are briefly presented hereafter.

UncertML applications and tools

Within the UncertWeb project, the UncertML 2.0 specification will be further developed to provide a Java (and possibly R) API, and an Uncertainty Transformation Service. Further tools are being developed to support the management of uncertainty in workflows, including:
- elicitation tools to support the expert elicitation of uncertainty
- sensitivity analysis tools to support the exploration and partitioning of sources of uncertainty in a workflow; this will also include emulator technology
- visualization tools to support the visual exploration of uncertain results

All these tools will be based on and supported by UncertML. At present UncertML is not integrated into the mainstream geospatial standards, for example Geography Markup Language, Observations & Measurements, and associated metadata (ISO19115/19139). UncertWeb will develop profiles of these information models / encodings, to show how this can be done, and will illustrate these in 4 use cases.

➔ Recommendation to AIP: further work will usefully demonstrate more closely notions of uncertainty into the developing information models, which currently have a very deterministic feel. This will enable information interoperability and create a system capable of integrating the widest types of observations and models into a coherent decision and policy support mechanism.

4.5.6 Metadata for Geographic Information (ISO 19139 XML)

This standard addresses the scope of applicability and the interoperability levels for the encoding and the communication of quality assurance procedures, observations/measurements 'inner' quality estimates, and data products 'specified' quality attributes.
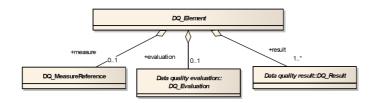
Figure 20.   Data quality element descriptors (ISO19157)

Figure 21.   Required tiers to encode spatial domain, units of measure, and uncertainty (from UncertML)

The main issue with the ISO 19115 metadata implementation, which is mirrored in the XML implementation (ISO19139), is the actual standards flexibility. It essentially allows the user to specify pretty much any measure in the 'DQ_Element', and then since the 'errorStatistic' in the 'DQ_QuantitativeResult' is also a string, this could be anything. Such flexibility can be seen as assisting the deployment of this standard - basically we can all keep using whatever language (semantics) we like to describe the 'errorStatistic' we are using to characterize the quality of the data. There is also no support here for defining what "truth" the error is computed with respect to, something that is essential to define for geospatial data due to the complexity of the scaling of spatio-temporal phenomena. To achieve better interoperability one needs a controlled vocabulary for 'errorStatistic' and an implementation of this in several easy to use encodings, such as envisaged in UncertML and being deployed in UncertWeb.

4.6 Discovery and assessment: the GEOSS Common Record

4.6.1 Roles for metadata in discovery and access for the GEOSS Clearinghouse

The End-to-End process designed during AIP-2 discussed the roles for metadata in discovery and access[14]. GEOSS is built to support data providers contributing standardized data access services, allowing users to bind to a dataset without needing specialized instruction or contact from the providers. In order for data users to find these services, service metadata from providers needs to answer two questions: (1) What metadata is needed for discovery? (2) Once one has discovered a dataset, how can they access the service ?

The GEOSS components and services are described by various metadata standards and profiles, typically driven by the context in which they were created or organization that created them. In order for all datasets to be found using the GEOSS Common Infrastructure (GCI), specifically the GEOSS Clearinghouse, there must be a minimum set of metadata fields that allow for discovery and access across distributed organization specific metadata.

4.6.2 GEOSS Common Record

Through the collaboration between a variety of stakeholders in AIP-2 and AIP-3, a key set of fields for all Earth observations has emerged as a potential GEOSS Record:
- Dataset title
- Data originator or distributor
- Abstract
- Geolocation (Bounding Box, Region, Point)
- Temporal Extent
- Keywords
- Service Type (Type of resource)
- Metadata File ID (Unique ID)
- Associated Component (catalog information)
- Last date of modification of the resource
- Reference to resource

Upon which the following additional / more specific parameters have also been proposed:
- **Observed footprint** – in the case of satellite observations, a bounding box may be too coarse to be sufficiently discriminating, and there are many cases where the location an observation was made is not in fact identical to the footprint of that part of the earth to which the observation applies ("feature of interest").
- **Phenomenon** – while this may roughly correspond to dc:subject, a precise indication of what the observation served to estimate (not necessarily the same as what was measured) is the critical property for most earth observation applications. This may or may not include the other indicator of what was examined, medium.
- **Medium** – the type of earth material that exhibits the observed phenomenon (e.g. water vs air temperature). This often is the critical property that distinguishes observation data of interest to a particular community.
- **Model** – this term is used broadly to describe either the process / sensor used to derive an observation from measurements, or the simulation model used to predict or interpret an observation from older or indirect measurements. This is another parameter that is critical for both discovery and selection of EO resources, and is only vaguely accounted for in more general metadata elements (e.g. dc:source). The combination of csw:Record elements and the above additional EO parameters has been suggested as a mandatory "geoss:Record" for contributed GEOSS resource descriptions.

➔ Though GEOSS should anticipate heterogeneous metadata provenances, we should promote minimum documentation for specific purposes, e.g. GEOSS Record for discovery. Also, GEOSS should aim to use existing standards/specifications and work through the proper channels where modifications to these standards/specifications are deemed appropriate. A crosswalk towards the main established practices was defined during AIP [15].

---

[14] http://www.ogcnetwork.net/system/files/AIP2_E2EDA_ER_v1_0_final.pdf

[15] https://spreadsheets.google.com/ccc?key=0AqFCPt1U1kS4dEJGSUlLcXB6QmxjMVZuMTNRNEIxOGc&hl=en#gid=1

4.6.3 Implications of User Expectations for Provider Documentation

Based on these general Web experiences, Issue-oriented Users expect from the GEO Portal to find relevant resources based on a simple search interface that is similar to those they encounter on the Web. This search interface may support keyword as well as spatial searches. The users will also expect fast responses. The combination allows them to iterate through searches, modify searches by adding more specific words, limiting results to a specific area and such like.

The above set of fields in <u>chapter 4.6.2</u> allows for the GEOSS Clearinghouse to perform general text searches of all fields. However, because of the additional fields it also allows for several more specific searches. It allows data providers and distributors to find their own records. The inclusion of spatial and temporal extent allows for spatial and temporal searches. Information about the service type allows filtering by specific service standard. Searching along the associated component links the metadata record of a given dataset to a larger community catalog. Additionally, if a common vocabulary is used in the keywords by a community of organizations, discovery by measurement platform or observed phenomena could also be found.

These fields are flexible enough for describing and finding a wide variety of Earth Observations. These fields are meant to be a coarse filter that allows for building more community-specific search interfaces on additional metadata included in the records. The GEOSS Clearinghouse performs the aggregation service for distributed communities, but isn't burden with needing to provide all community-specific search interfaces. The common record also realizes the GEOSS mantra of one dataset being applicable for many uses, not just original intended use.

4.6.4 Metadata standards and crosswalks

With the advent of ISO 19115/19119/19139 there is a framework for describing these resources more fully for the purpose of understanding. However, these international standards allow for variations through the creation of profiles. The large SDI initiatives (North America and INSPIRE particularly) already have created profiles of the ISO standard that are different in various aspects. At the core (metadata to support discovery) there is generally sufficient overlap but considerable work in AIP-II and AIP-III was needed to define this overlap and document resulting gaps

To identify the fields necessary for a GEOSS Common Record that allowed discovery and access we started with the ISO 19115 Core Metadata fields (Nogueras-Iso, 2005) and also the fields that were needed for a Catalog Service for the Web (CS-W) Record. The common record is independent of metadata standard, but ISO 19115 and CSW are two widely used standards for description and query. Starting with this as a basis ensured a wide number of organizations already were compatible with the common record.

Therefore, the GEOSS Common Record fields can be in any metadata format, it is just important that the community agrees that 1) this set is needed to find any Earth Observation service, 2) the crosswalk that has been developed does capture the correct linkages between various metadata standards.

The AIP-2 and AIP-3 communities have developed collaboratively this crosswalk[16]. A joint publication[17] "Enhancing data discovery, understanding and usage through an Air Quality metadata system" describes the process of identifying a GEOSS record that could be common for all Earth observations. A perspective on Data Spaces[18] is also provided in this work. "*Data Spaces are an abstraction in data management that aim to overcome some of problem encountered in data integration system. The aim is to reduce the effort required to set up a data integration system by relying on existing matching and mapping generation techniques (…). Data Spaces are not a data integration approach. Rather, they are more of a data co-existence approach*." [Wikipedia]

> ➔ The DHWG recommends further investigating similarities between O&M "sampling feature" and the GEOSS Common Record "Observed footprint", and between ISO19115 "processing lineage" and the GEOSS Common Record "Model".

---

[16] https://spreadsheets.google.com/pub?key=0AqFCPt1U1kS4cEcwY0QzNVNCX0EtNExSa1FUZXNlUFE&hl=en&output=html

[17] http://capita.wustl.edu/capita/capitareports/090504Stresa_IT/Metadata_for_GEOSS/090406_CatalogMetadata_Submitted.doc

[18] http://en.wikipedia.org/wiki/Dataspaces

4.6.5 Perspectives from the CEOS Community Catalog

NOAA, working with the CEOS agencies, is providing key support toward the development of a prototype for an Earth Observation community catalog. The goal of this effort is to provide a CEOS WGISS Integrated Catalog (CWIC), satisfying the need in the GEOSS operations concept for a catalog of Earth Observation satellite products. This catalog develops a two-tiered discovery and access model. Once users identify satellite data *collections* that are of interest, more detailed queries are sent to inventory systems where individually relevant *products* are identified for processing and data access. The CWIC design team is composed of staff from NOAA, NASA, USGS, JAXA, ESA, and EC, and started its activities at the beginning of 2010.

"It is important to acknowledge that there are very different demands and expectations related to the discovery and access of data from different levels. For satellite data, the data processing level can have a strong influence on the ability to present the data as coverages or layers:

- Level 0 data are full-resolution raw instrument/payload data. They are intended primarily for further processing to higher levels, not for direct use.
- Level 1A data are reorganized full-resolution versions of Level 0 data, with calibration information appended but not applied.
- Level 1B data are Level 1A data that have been geolocated and processed to sensor units.
- Level 2 data are derived geophysical variables at the same resolution (usually) and location as the Level 1 data.
- Level 3 data are temporally and/or spatially averaged data on regular grids.

For example, there are many more instances of observation oriented data (particularly Level 1a and 1b). A collection of Level 1b data may have many millions of granules, each representing a different observation. Additionally, these granules may be very significant in size, perhaps measured in gigabytes per granule. In contrast, Level 3 data is summary in nature, perhaps a monthly average. Looking for the right set of granules out of millions is much more of a challenge than selecting the right month for analysis. Furthermore, there are many other issues pertaining to the delivery and usage of data that is much larger. (For example, it might be reasonable to ask for a series of summary products representing a year to be delivered to your desktop. However, asking for a year's worth of Level 1b data is quite another story.) The remote sensing community has well established conventions of establishing directories to support the discovery of types of data and inventories to manage the representations of the hundreds of millions of individual granules". [19]

CWIC will reuse components, services and standards that are currently operational at the CEOS WGISS agency data systems, to integrate common inventory search with specific data access mechanisms. Clients will access CWIC through the OGC Catalog Service for the Web (CS-W) interface, using a subset of ISO 19115 metadata for queries.

Only such an integrator work for community catalogs and portals, designed along with the general purpose GCI Clearinghouse and GWP for a discovery in support inter-disciplinary communities developments, would contribute to a scalable system of systems, for effective services provided to the breadth of GEO users. The data and service providers in the GEO community being highly varied, the GCI by itself can only support providers on one end of that space. Community catalogs and portals, as integrated information services, can themselves be considered as GEO services that provide enhanced capabilities for particular disciplines or data types.

The CEOS WGISS agencies are working together to identify a minimum, common metadata set that must be supported across all satellite data providers. Once CWIC implementation has progressed to the point where it can provide useful services, it will be registered in the GCI registry.

➔ The DHWG recommends the GCI-TF to promote a GEOSS development strategy in the realm of the integration work of community-based information systems, including strong support to the development of newly required inter-disciplinary communities, that would benefit from the GCI to search and understand the needed resources and interfaces, so to address their ad-hoc requirements.

---

[19] "Challenges to Seamless Data Discovery and Access to Remote Sensing Data", CEOS WADC Project

4.7 Processing chains

4.7.1 Chaining and discovery strategies

The CSIRO scenario and DH use case implement a number of workflows (time-series harmonization, time-series gap fixing, data spikes generation) in the "Kepler Scientific Workflow System", to process observations provided by various input SOS instances. Kepler can then be viewed either as a Sensor Web client, or as a proxy for a Web Processing Service.
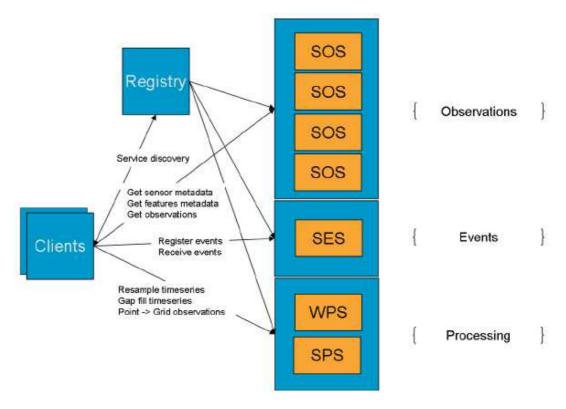
Figure 22.   High-level architecture of the Hydrological Sensor Web (CSIRO)

4.7.2 Integration of uncertainty measurements

The UncertWeb project is developing a series of prototypes and real applications in which uncertainty is managed in a service oriented framework, using widely used standards such as the W3C standards, OGC Web Service standards and Observations and Measurements and SWE standards.

It is incorporating UncertML within these standards to define best practices for including uncertainty in the information models and service interfaces. The project is moving rapidly to define the best approaches for managing uncertainty in processing workflows. Further developments are expected in the coming months that will define what is seen as the best approach to including uncertainty and thus also quality information in the GEOSS context.

A further project, GeoViQua starting in February 2011, will consider the links between data quality, uncertainty and user trust. In particular this will focus on how to best provide quality information in the GEOSS context. This will further develop the usage and tools to support the usage of UncertML.

4.8 Standards and Governance

4.8.1 From quality assurance rules to data quality encoding standards

The data products registration in the CSR shall be adapted to encourage the association of a Quality Indicator to each dataset, which will be linked to the QA4EO questionnaire (so, used as a published evidence supporting a declared quality information).

When users access these datasets, the GEOSS infrastructure will allow access to this quality information and should preserve its association to each dataset. This will be further explored within the GeoViQua project.

The GIGAS contribution on presenting the O&M standard, and the DHWG mindshare on how to downlink Quality Assurance to Data Quality encodings, taking into consideration uncertainty management, indicate possible tradeoffs from the current GEOSS and GCI tools.

Several discussions occurred during the DHWG telecons, on how to reflect and interlink in the GCI the assets that are supporting such an approach for data harmonization, sharing, and reuse.

➔ The DHWG recommends strengthening QA4EO by providing stronger guidance on the appropriateness of a Bayesian approach to the evaluation of uncertainty.

➔ The DHWG recommends targeting quality indicator for all datasets as a 'probabilistic assessment of uncertainty' with respect to a well defined reality, and supported by statistical validation. This could also include a more comprehensive statistical context.

4.8.2 Recommendations emphasized from the QA4EO contribution

The Quality Assurance framework and guidelines for Earth Observations (QA4EO), in the continuity of the GEO task DA-06-02, and of the new GEO task DA-09-01a, is aimed at "Develop a GEO data quality assurance strategy, beginning with space-based observations and evaluating expansion to in situ observations, taking account of existing work in this arena".

➔ In this regard, the DHWG recommends:
   o To develop coordination with the GEO SIF to serve the QA4EO outreach to GEOSS SBAs
   o To further coordinate with the AIP, to leverage quality assurance guidelines and tools provided through exemplar systems, as deployed and referenced in the frame of the GCI
   o To provide initial guidance documents on the choice of encoding standards all along the processing chains, that is expanding the QA for the 'metrology-related' Space component domain (e.g. calibration and validation), to a selected set of GEO thematic applications domains, where a rational end-to-end use of EO products have been pursued, through a statistical approach to quantify and manage uncertainty
   o To evolve in its forthcoming development stages to more explicitly encompass and in particular provide examples not only guidance for "level 0" and "level 1" data, and the related topic of "Products for primary dissemination formats", but also "level 2" or "level 3" products, and "products for secondary dissemination formats". For example, the steps 2 and 3 of the QA4EO document "Reference Standards for QA", chapter "Characterizing a reference standard", shall encompass and consider the GEOSS Standards and Interoperability Registry (SIR)
   o To support the future integration of the QA4EO questionnaire into the GEO registration procedure.
   o To encourage the participation to the upcoming QA4EO Workshop on Providing Harmonized Quality Information in Earth Observation Data by 2015 …

4.8.3 Liaisons to the SIF

Standardized application schemas increase the ability to share information among applications. These schemas are especially useful for information systems, software developers and data users of geographic information having spatial and temporal characteristics, to provide consistently understandable data structures. Examples of GML Application Schemas are provided on the OGC Website, the INSPIRE initiative website...

➔ The DHWG recommends GEOSS to consider adopting a methodology for the development and registration of application schemas. For consideration, an INSPIRE Drafting Team has developed an initial "Data Specifications" methodology and toolset (UML models, XSD schemas, technical guidelines), and is now entering in a next stage of this work, that will fully encompass the GMES data products. It provides a coordination opportunity with the GEO community that suits well with AIP mandate.

4.8.4 Liaisons to the GEOSS Task Forces

4.8.4.1 GCI-TF

Two main recommendations were identified more suited to the GEOSS Common Infrastructure Task Force.

➔ The DHWG recommends supporting the development of controlled vocabulary for 'errorStatistic', and an implementation of this in '*easy to use*' encodings.

➔ The DHWG recommends that the notions of "series" or "collection" (e.g. a "product specification", a "datasets series",…), defined as a unique combination of a location, scale, observed variables and time intervals that specify a sequence of observations, shall be formally coordinated and defined within GEOSS SBAs, and serve the GEOSS discovery strategies.

4.8.4.2 DS-TF

One main recommendation was identified more suited to the Data Sharing Task Force.

➔ The DHWG recommends for future work to investigate the context of Open Linked Data and Semantic Web tools for the licensing choices made already by government bodies sharing online data, especially when related to the INSPIRE initiative.

5. Conclusion

During AIP-3, the Data Harmonization Working Group conducted eleven dedicated WG telecons, and participated to eight coordination telecons (with DA-09-01a or DA-09-01b). Topics advanced by the working group were as follows:

- Reconcile 'geophysical parameters' with 'geographical features' information viewpoints: tools and methods to combine and compare satellite time series with in-situ data and environmental models, forecast models…
- Promote Quality Assurance for Earth Observations (QA4EO): process for 'Geophysical observations', "sensor-to-product-to-service" lineage information, uncertainty management and visualization, points-of-truth for calibration, and potential linkages with uncertainty management…
- Contribute to transition from Earth Observations (e.g. CEOS activities) towards the EC INSPIRE and GMES initiatives, standards for in-situ measurements, geographical features production workflows, and environmental (predictive) models settings…
- Lower barriers to interoperability and data usability, shape a convergence of GEO recommendations towards a set of harmonized standards, based on best practices and community orientations…

The contributed resources from participant organizations covered applied scenarios, engineering use cases, and components and services registered to the GEOSS Components and Services Registry (CSR). From the provided scenarios and use cases, several requirements for the deployment, configuration and registration of the contributed resources were emphasized. Out of these requirements, recommendations to the GEOSS governance bodies were then formulated. A set of about 10 recommendations mainly address the following topics:

- The development of the geosphysical parameters semantics and use for GEOSS data discovery, along with online management of supporting conceptual schemas, encoding schemas (e.g. XML schemas and Schematron), controlled vocabularies, code lists, and glossaries.
- The registration of additional supporting standards into the SIR, and the need for additional guidance on their use from the CSR and the Best Practices Wiki (through cross-linking strategies)
- Some enhanced governance and liaisons within GEOSS for the quality assurance processes and data products quality indicators.

Perspectives for the AIP-4

As the GEOSS infrastructure develops on its own roadmap, the Semantic Web technology is growing in several application domains related to the GEOSS SBAs (Health, Life sciences…). This Engineering Report has highlighted some aspects of this technology addressing requirements to some Data Harmonization components (use of URIs for terminology component and services endpoints). Semantic Web technology is offering an interesting potential to facilitate cross-disciplinary data integration. Access to Governmental information is another example currently expanding and relying on the Open Linked Data concept. There is a considerable amount of work on this new topic to be investigated in a collaborative, cross-disciplinary approach, as typically provided by the GEOSS AIP process. It represents a powerful approach for the development within GEOSS, and towards the identified SBA end-users, of Data Harmonization and Data Sharing tools and practices. There is a also a requirement to progress the treatment of uncertainty and data quality to permit the harmonized representation of data quality within the GEOSS, infrastructure but more particularly to support the creation, discovery and use of the data and associated uncertainty and other quality statements.

6. Glossary

AIP	Architecture Implementation Pilot (GEO task)
EC	European Commission (http://ec.europa.eu)
CAL/VAL	Calibration/Validation (CEOS)
CCAM	Cubic Conformal Atmospheric Model (CSIRO / WMO)
CEN	European Committee for Standardization (http://www.cen.eu)
CEOS	Committee on Earth Observation Satellites (http://www.ceos.org)
CGMS	Coordination Group of Meteorological Satellites (WMO)
COVE	CEOS Visualization Environment (http://www.ceos-cove.org/)
CSIRO	Commonwealth Scientific and Industrial Research Organization (http://www.csiro.au)
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science, Inc. (http://www.cuahsi.org/)
DHWG	Data Harmonization Working Group (AIP)
DPIPWE	Department of Primary Industry, Parks, Water and Environment (Tasmania)
DSTF	Data Sharing Task Force (GEO)
BPW	Best Practices Wiki (GEO / IEEE)
EUMIS	European Marine Information System (EC NETMAR project)
FP7	7th Framework Programme (EC)
GCI	GEOSS Common Infrastructure (GEO)
GEO	Group on Earth Observations (http://earthobservations.org/)
GEOSS	Global Earth Observations System of Systems (GEO)
GFM	General Feature Model, defined in ISO19109 (ISO)
GIGAS	GEOSS , INSPIRE and GMES an Action in Support (EC FP7)
GMES	Global Monitoring for Environment and Security (http://gmes.info)
GML	Geography Markup Language (OGC)
GOS	Global Observing System (WMO)
GSICS	Global Space-based Inter-Calibration System (WMO)
GWP	GEOSS Web Portal (GEO)
IEEE	Institute of Electrical and Electronics Engineers (http://www.ieee.org)
ISO	International Organization for Standardization (http://www.iso.org)
MIM	Mission, Instrument, and Measurements database (CEOS)
NETMAR	Open Service Network for Marine Environmental Data (EC)
OGC	Open Geospatial Consortium (http://www.opengeospatial.org/)
OMXML	Observations & Measurements Markup Language (OGC)
PML	Plymouth Marine Laboratory (http://www.pml.ac.uk)
QA	Quality Assurance
QA4EO	Quality Assurance for Earth Observations (GEO)
RDF	Resource Description Framework (W3C)
SIF	Standards and Interoperability Forum (GEO)
SKOS	Simple Knowledge Organization System (W3C)
SOS	Sensor Observation Service (OGC)
SPARQL	SPARQL Protocol and RDF Query Language (W3C)
SSNIG	Semantic Sensor Network Incubator Group (W3C)
SWDWG	Semantic Web Deployment Working Group (W3C)
SWE	Sensor Web Enablement (OGC)
W3C	World Wide Web Consortium (http://w3c.org)
WCS	Web Coverage Service (OGC)
WFS	Web Feature Service (OGC)
WGCV	Working Group on Calibration and Validation (CEOS)
WGISS	Working Group on Information Systems and Services (CEOS)
WMO	World Meteorological Organization (http://www.wmo.int/)
WMS	Web Map Service (OGC)
XML	Extensible Markup Language (W3C)

7. References

7.1 Summary of participants' responses to the AIP-3 CFP

Aston University proposed submission represents **a chain of web processing services**, and some other web service components which take a sea level pressure at a given location, and correct this to mean sea level pressure taking into consideration some of the main **sources of uncertainty** in this operation. The uncertainties in the inputs to the chain, **taken to be related to issues of data quality are represented as UncertML** (www.uncertml.org), and this is propagated through the chain, including some conversion between different representations of uncertainty where this is required. The services invoked include sampling a digital elevation model, interpolation of elevation, conversion of representations of uncertainty and correction of pressure to mean sea level given the uncertain elevation and other (uncertain) factors. The **chain uses BPEL for orchestration of the services**, and **most services are web processing services**, augmented with a WSDL description to facilitate BPEL chaining

CSIRO intends to contribute to the AIP-3 on three fronts: 1) **Scenarios - water resource management use cases**. 2) Component and service registration - **profiles of SWE information models and service interfaces** that allow correct encoding and transmission of water information. 3) Architecture - possible **contributions to the Data Harmonization and Vocabularies technology themes**. CSIRO believes it can contribute in these areas because of its **expertise in the water domain**, contributions to the **international hydrology efforts and experience** designing and building hydrology systems. CSIRO intends to do this by providing **use cases that provide detailed requirements relevant to the scenarios**.

CEOS Systems Engineering Office (SEO) is offering systems engineering services to benefit the development of the GEO portals. As the space arm of GEOSS, CEOS data structures, tools, and engineering processes should be common to GEOSS. The CEOS SEO has developed, with participating organizations like the WMO, **data formats that are utilized within CEOS in the online Measurements, Instruments, Missions (MIM) database** and other tools. These other tools are SEO developed and include the **Systems Database for online gap assessments** and the **CEOS Visualization Environment (COVE) tool** which is a web-based tool that uses Google Earth to display satellite sensor coverage areas and coincident scene images for multiple sensors.

Utilizing these existing data structures within the **GEO Portal will allow for MIM data to be easily transitioned from CEOS to GEOSS**, allow for user ease within the CEOS and GEO Portal web sites, and allow for other information sharing in an efficient manner. The CEOS SEO proposes to provide **engineering services** to support data format adaptation, user viewpoints, and MIM data transition.

GIGAS proposal addresses "Architecture and Interoperability Arrangement Development", with a focus on **data interoperability**. The main contribution proposed here is to investigate the **possibility of a Common Foundation between GEOSS and INSPIRE**, in order to enhance the opportunity for interoperability - both between these initiatives, and within GEOSS. The main deliverable from our contribution will be a **report to be placed on the GEOSS Best Practice wiki**, describing our proposal for a **common foundation that reconciles the 'geophysical parameter' and 'geographic feature' views of Earth observations**. Our contribution to such activities would **encourage the use of the General Feature Model and INSPIRE Generic Conceptual Model, and use the Observations and Measurements model** if appropriate.

PML seeks to increase the **availability of environmental monitoring data** and participates in projects with this aim. Some current examples include **DevCoCast** (EU FP7, providing ocean monitoring data to African and South American countries via EUMETCAST), **NCOF** (a UK initiative encouraging collaboration between researchers, data providers and operational forecasters) and **NETMAR** (EU FP7, extending the capabilities of OGC/web-based **visualisation** and **analysis**). The societal benefit areas we most frequently seek to address are **ocean-related aspects of ecosystems, health and disasters**. Use Case 1: **Relationships between physical and biological variables**. Use Case 2: **Ecosystem model validation**.

QA4EO Task Team is proposing to provide the needed guidance to any GEOSS element or SBA that needs to improve its **data quality assurance strategy in the framework of AIP**. Specifically, the QA4EO Task Team will provide: a questionnaire that can be used by any interested entity for the **tailoring of QA4EO**; general guidance and know-how on specific matters to support the SBAs; and **recommendations for GEOSS Interoperability Arrangements**.

AIP-3 CFP responses can be consulted here: http://www.ogcnetwork.net/node/635

7.2 Contributions to the GEOSS Components and Services Registry (CSR)

7.2.1 GIGAS / GMES Marine Services

To date, the GMES initiative as a whole is registered as a GEOSS Component, and the GMES Core (business) services are registered as GEOSS Services. This highlights a discrepancy also revealed within the CSR by other entries, on the understanding and usefulness of GEOSS Services either as informative resources or as computational services (with standardized software interfaces).

The MyOCean project is registered as a GEOSS Service.

https://geossregistries.info/geosspub/service_details_ns.jsp?serviceId=urn:uuid:48142827-2c18-4602-8482-d629d5a87a28

7.2.2 CSIRO

The Hydrological Sensor Web is first of all defined as a GEOSS component.

https://geossregistries.info/geosspub/component_details_ns.jsp?compId=urn:uuid:b7225be5-0f6f-4e56-9c76-e1f64fd12809

There are two associated services for that component, a Catalog (RESTful) and a Sensor Observation Service (SOS).

To date, the CSIRO South Esk catalog is presenting a search web page and a listing page of catalogued SOS.

This catalog service describes a 'REST' interoperability arrangement.

Such a 'REST' entry is provided in the SIR as a best practice / architectural style.

HTTP/1.1 was submitted to the SIR as a new entry,

https://geossregistries.info/geosspub/service_details_ns.jsp?serviceId=urn:uuid:1bac831f-b7d4-4c84-b7b4-cf86b9d499ff

To date, the CSIRO South Esk Sensor Observation Service (SOS) is responding with SensorML (describeSensor), O&M (getObservation) and WaterML v2.0 output encodings, depending on the SOS function requested.
A thin client for visualization is also referenced (accessible through the Information URL).

https://geossregistries.info/geosspub/service_details_ns.jsp?serviceId=urn:uuid:ebe78116-79f1-4896-b22d-a489ecf40e85

7.2.3 PML

The Plymouth Marine Laboratory response to AIP-3 CFP identified the following Data services offering for registration in GEOSS:

- Satellite time series of optical properties, chlorophyll-a concentrations, sea surface temperature and sea-surface elevation processed from NASA & ESA source data.
- In situ measurements of phytoplankton or physical properties from time series either in single locations (such as the L4 station off Plymouth) or along track series (such as, for example, ships of opportunity or ferrybox systems).
- Ecosystem model output for the UK waters

Also, Web Processing Services combining and comparing satellite time series with in situ and model data were envisioned for regitration.

7.2.4 ASTON University

ASTON University contributed a web based system that allows users to examine and understand the components of a processing chain, how they interact and how the uncertainty propagates (see http://uncertws.aston.ac.uk/client/). Access to the chain and components is provided at a low level of activity, run on prototype / development servers (not full deployment servers).

This contribution is aimed at providing expertise of chaining web services and uncertainty management in these chains with any interested parties within GEOSS, and is envisioned for registration.

7.2.5 CEOS

The CEOS online Measurements, Instruments, Missions (MIM) database is registered as a Component, with no Service instance associated.

Other services like the COVE (CEOS Visualization Environment) were also envisioned for registration.

https://geossregistries.info/geosspub/component_details_ns.jsp?compId=urn:uuid:a979415d-5ac8-4693-a622-dab396211d0e

7.3 Contributions to the SIR and the GEOSS Best Practices Wiki (BPW)

7.3.1 CSIRO

WaterML version 2.0 registered in the SIR on Sep 28th, 2010.
http://seabass.ieee.org/groups/geoss/index.php?option=com_sir_200&Itemid=157&ID=301

7.3.2 QA4EO

A dedicated entry was provided in the GEOSS Best Practices Wiki on the Quality Assurance Framework for Earth Observation (QA4EO).
http://wiki.ieee-earth.org/Best_Practices/GEOSS_Transverse_Areas/Data_and_Architecture/Data_Quality_Assurance
Cross links to the CSR (conformant components and services) and the SIR (selected implementation standards) are to be defined. These links would be provided on a 'per project' basis.

7.4 Bibliography

ISO 19101:2002 Geographic information – Reference model
ISO 19109:2005 Geographic information – Rules for application schema
ISO 19115:2003 Geographic information – Metadata
ISO 19123:2005 Geographic information – Schema for coverage geometry and functions

ISO/CD 19157 Geographic information – Data quality, July 15th, 2010
The creation of new project team ISO 19157 is merging and harmonizing the following standards:
> ISO 19113:2002 Geographic Information – Quality principles
> ISO 19114:2003 Geographic Information – Quality evaluation procedures
> ISO/DTS 19138:2005 Geographic Information – Data quality measure

Open Geospatial Consortium Inc. Discussion Paper – Uncertainty Markup Language (UnCertML)
OGC 08-122r2, v0.6, April 08th, 2009
Matthew Williams, Dan Cornford, Lucy Bastin & Edzer Pebesma

Open Geospatial Consortium Inc. Implementation Standard
OGC 10-157, v0.2, November 11th, 2010
Profile of Observations and Measurements (O&M) for describing Earth Observation products (EO products)

Open Geospatial Consortium Inc. Discussion Paper – Harmonizing Standards for Water Observation Data
OGC 09-124r1, v0.2, December 4th, 2009
Peter Taylor (CSIRO), Gavin Walker (CSIRO), David Valentine (CUAHSI)

Open Geospatial Consortium Inc. Interface Standard
OGC GML Application Schema – Coverages (a GML coverage structure extending the GML 3.2.1)
OGC 09-146r1, v1.0, October 27th , 2010
Peter Baumann, Andrei Aiordachioaie, Jinsongdi Yu

INSPIRE D2.5 Generic Conceptual Model, Version 3.2
INSPIRE D2.6 Methodology for Specification Development, Version 3.0

GIGAS – GEOSS, INSPIRE and GMES an Action in Support
GEOSS AIP-3 Contribution - Data Harmonization, 23-Nov-2010
Andrew Woolf (STFC), Simon Cox (JRC), Clemens Portele (interactive instruments)

GEO task DA-09-03 on Global Datasets
   * DA-09-03a: Global Land Cover
   * DA-09-03b: Global Meteorological and Environmental Data
   * DA-09-03c: Digital Geological Map Data
   * DA-09-03d: Global DEM

QA4EO Framework
GEN-DQK-001 – QA4EO guide to establish a Quality Indicator on a satellite sensor derived data product
GEN-DQK-006 – QA4EO guide to expression of uncertainty of measurements

CEOS WGISS Architecture and Data Contributions (WADC) Project
Challenges to Seamless Data Discovery and Access to Remote Sensing Data
Michael Burnett, Chris Lynnes, Lyndon Oleson, January 2011

IGARSS Conference, Hawaii, August 2010
The CEOS WGISS Integrated Catalog for GEO

7.5 Web pages

The AIP-3 DHWG references are listed on the Google Sites project pages:
http://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization
The AIP-3 DHWG Tasks and Deliverables:
http://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-tasks
The AIP-3 DHWG Resources:
http://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-resources
The AIP-3 DHWG Scenarios and Use Cases:
http://sites.google.com/a/aip3.ogcnetwork.net/home/home/data-harmonization/harmony-scenarios

CEOS Virtual Constellations for GEO
http://www.ceos.org/index.php?option=com_content&view=article&id=275
CEOS Land Surface Imaging Constellation Portal
http://wgiss.ceos.org/lsip/
CEOS Working Group on Calibration and Validation
http://calvalportal.ceos.org/cvp/web/guest/home
http://www.ceos.org/index.php?option=com_content&view=category&layout=blog&id=75&Itemid=113
CEOS MIM database
http://database.eohandbook.com/
CEOS WGISS Interoperability Hanbook, February 2008, Issue 1.1
http://www.ceos.org/images/WGISS/Documents/Handbook.pdf

QA4EO website
http://qa4eo.org/

Aston University UncertWeb demo client
http://uncertws.aston.ac.uk/client/index.php
Dan Cornford's wiki page developing a paper on relations between uncertainty, data quality, metadata and data
https://wiki.aston.ac.uk/foswiki/bin/view/UncertWeb/UncertaintyAndStandards

Intamap Project page
http://www.intamap.org/
Videos demonstrating how "Aguila" software allows to interact with spatial data that comes as probability distributions (coming from the Intamap WPS):
http://intamap.org/videos.php

InterRisk (FP6-IST GMES)
http://interrisk.nersc.no/
http://www.npm.ac.uk/rsg/projects/interrisk/

NETMAR Project's Case Studies and Deliverables
http://netmar.nersc.no/?q=node/36
http://netmar.nersc.no/?q=node/39

GML Application Schemas for geospatial information:
http://www.ogcnetwork.net/node/210
http://inspire.jrc.ec.europa.eu/index.cfm/pageid/2 (cf. 'Other documents' section)

CSIRO Water Resources Observation Model, running a numerical weather prediction model (Cubic Conformal Atmospheric Model) in Tasmania's South Esk region
http://www.csiro.au/sensorweb/ccam/index.html
http://www.csiro.au/thredds/catalog.html

CSIRO Integrated Water Information Systems theme (IWIS).
One of the research topics in the theme is near real-time situation awareness of river flow in regional catchments.
CSIRO has established a prototype Hydrological Sensor Web in the South Esk river catchment, NE Tasmania, to deliver short-term river flow forecasts.

CUAHSI Hydrologic Information System (HIS)
http://his.cuahsi.org/
CUAHSI Ontology files - Hydrologic Ontology for Discovery
http://his.cuahsi.org/ontologyfiles.html
CUAHSI Water Ontology and Controlled vocabulary
http://water.sdsc.edu/hiscentral/startree.aspx
http://water.sdsc.edu/hiscentral/startree.html
CUAHSI WaterML 1.0 and 1.1
http://river.sdsc.edu/wiki/Default.aspx?Page=WaterML&NS=&AspxAutoDetectCookieSupport=1