Geophysical Research Abstracts
Vol. 13, EGU2011-3230-2, 2011
EGU General Assembly 2011

Meta-propagation of Uncertainties for Scientific Workflow Quality Management in Interoperable Spatial Data Infrastructures

Didier G. Leibovici, Amir Pourabdollah, and Mike Jackson
University of Nottingham, Centre for Geospatial Science, United Kingdom (didier.leibovici@nottingham.ac.uk)

Abstract:
Addressing multidisciplinary interoperability, one of the goals of the European FP7 project EuroGEOSS [1] is to facilitate the derivation of new product datasets from existing data and scientific models. A scientific workflow instantiates a scientific model by combining the needed chosen resources within an interoperable environment using OGC standards [2] for (geo)computational processes and data embedded in web services. When using/testing different sources, changing the scale or adpating the scientific model with various scales, the user or the modeller needs some means to evaluate the «fit to purpose» of the workflow instance. Quality assessment provides quantification of the reliability of the workflow in term of the expected uncertainties, and, accumulation of evidence for its usability [3]. This is crucial for decision making and any proper use of the seamless ability to reuse existing workflows along with discovered or retrieved datasets and processes. The knowledge of how the data uncertainties are defined and propagated through the processes within the workflow is of concern. This paper proposes a framework, within existing interoperability/standards settings, that is able to encode quality information, and to assess the quality of an instance of a scientific model at workflow and sub-workflow levels using them directly [4, 5]: meta-propagation. Specific quality metadata for processes that allows simple data uncertainties to be propagated are derived and encoded along with the scientific model within a XPDL [4] file representing the workflow. A WPS (Web Processing Service) profile realising and performing the workflow will be described to allow querying for metadata propagation. This type of WPS could be thought as a WWS (Web Workflow Service) which could reify the use of WPS in a single hard-coded task and leaving the WWS for higher level combination of tasks.

Keywords: scientific workflow, web services, quality, uncertainty, metadata, error propagation
References:
1. EuroGEOSS, A European approach to GEOSS. FP7-ENV.2008.4.1.1.1: European Environment Earth Observation system supporting INSPIRE and compatible with GEOSS (Global Earth Observation System of Systems), 2009-2012, http://www.eurogeoss.eu
2. OGC, standards, 2010, http://www.opengeospatial.org
3. Leibovici, D.G. Hobona, G. Stock, K. and Jackson, M., Qualifying geospatial workfow models for adaptive controlled validity and accuracy. In: IEEE proceedings 17th International conference on GeoInformatics, August 2009, USA, pp. 1-5. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5293485
4. XPDL WfMC, Workflow Process Definition Interface – XML Process Definition Language (XPDL).Workflow Management Coalition, Document WfMC-TC-1025, 2008. http://www.wfmc.org/index.php?option=com_docman&task=doc_download&Itemid=72&gid=132
5. Leibovici, D.G. Pourabdollah A., Workflow Uncertainty using a Metamodel Framework and Metadata for Data and Processes " OGC TC/PC Meetings, 20-24 September 2010, Toulouse, France. http://portal.opengeospatial.org/index.php?m=projects&a=view&project_id=82&tab=2&artifact_id=40240

Meta-propagation of Uncertainties for Scientific Workflow Quality Management in Interoperable Spatial Data Infrastructures

session ESSI8: Uncertainty in Environmental Data and Models

Didier G Leibovici, Amir Pourabdollah and Mike Jackson

Centre for Geospatial Science
University of Nottingham

FP7 European project

- motivation: integrated modelling /scientific workflow

  model discovering / model building / reusing / rescaling / model refining

- aim: quality assessment

  quality description / error propagation / uncertainty analysis / user's perspective

- means: meta-model for workflows

  interoperability / standards / metadata (data & processes)

  quality principles & measures for processes / workflow notation /encoding / enrichment

integrated modelling/ scientific workflow

Figure 1: Example of geo-processing workflow model for ground condition forecast as a GEOSS-type model (Alonso & Hagen (1997)

model building

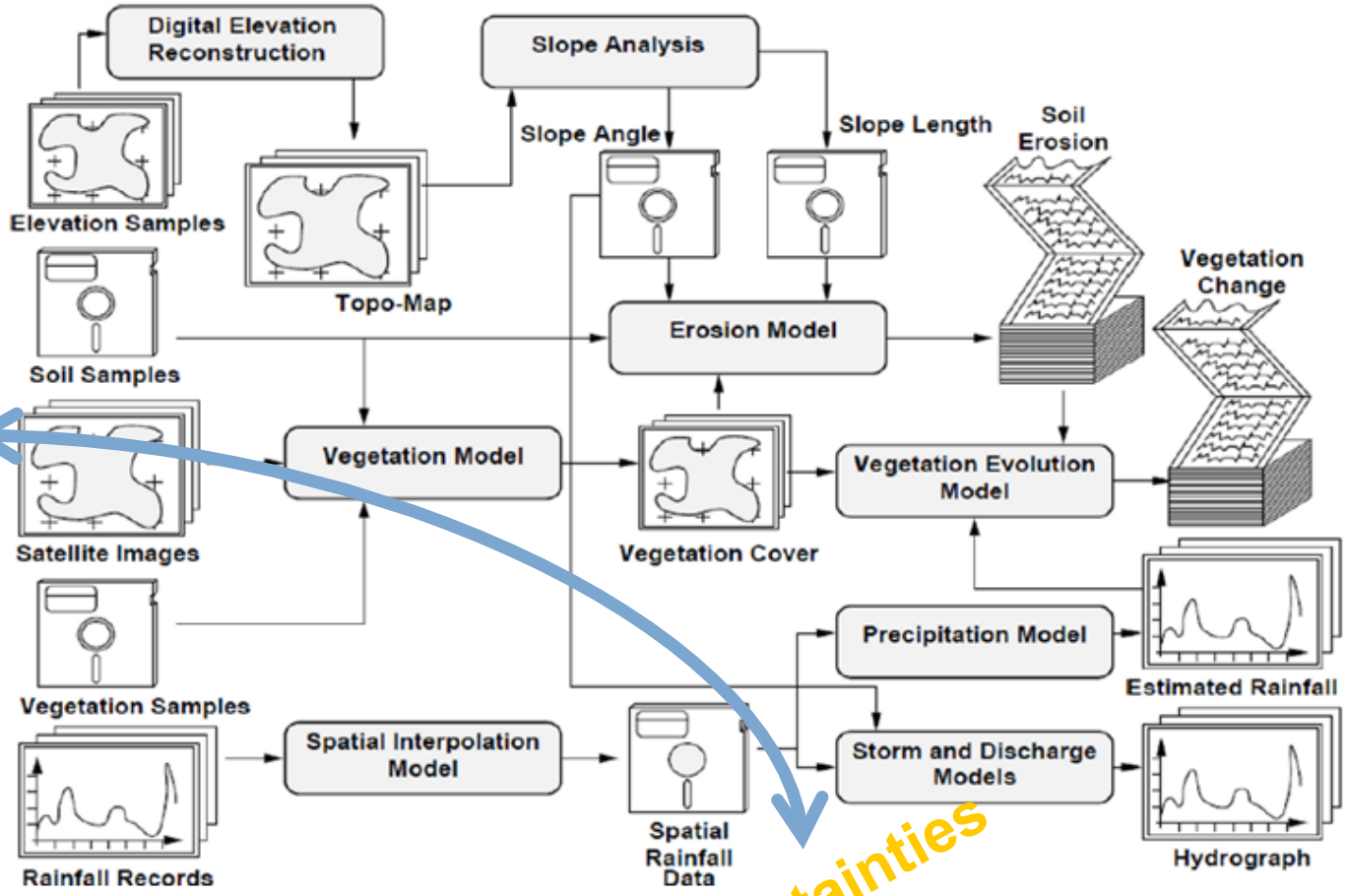
reusing

user's perspective

multidiscipline

rescaling

quality assessment



list of issues

- sharing

  representation / exchanging format

- publishing

  web access / discovering / semantic

- running

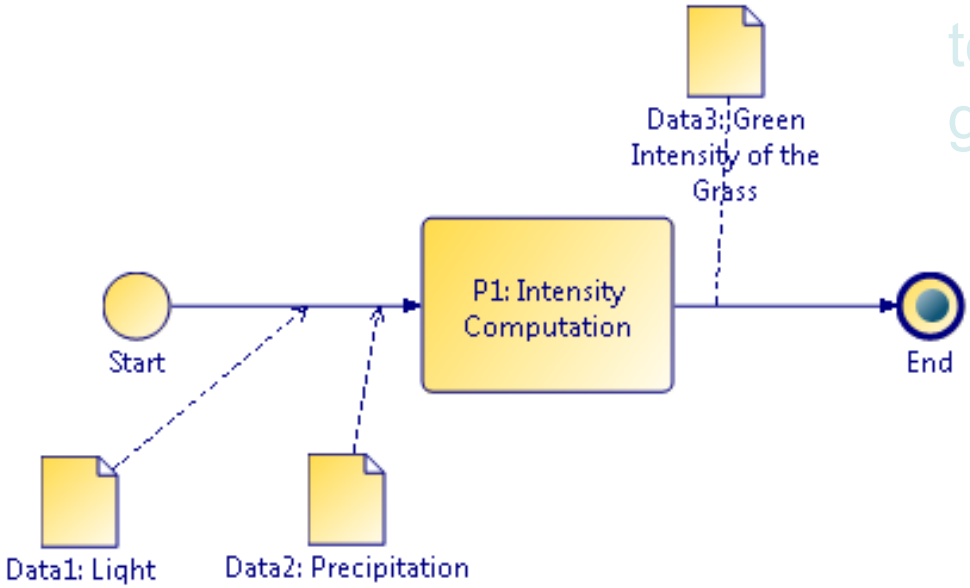
  engine / interoperable services

- assessing

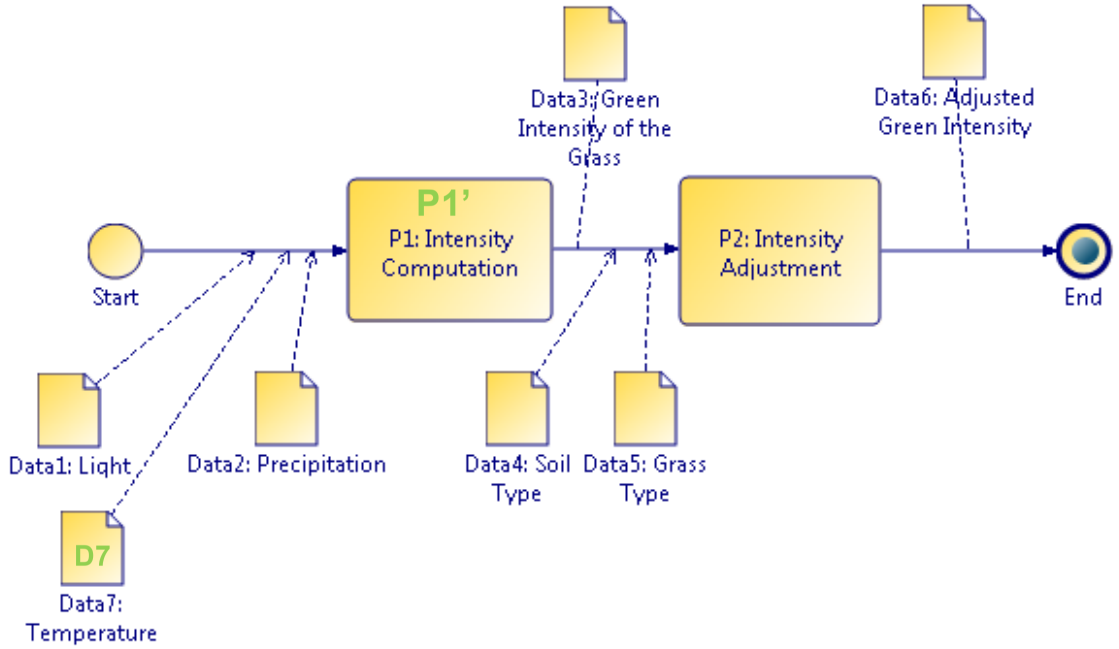
  quality / popularity /fit for purpose

BPMN representation

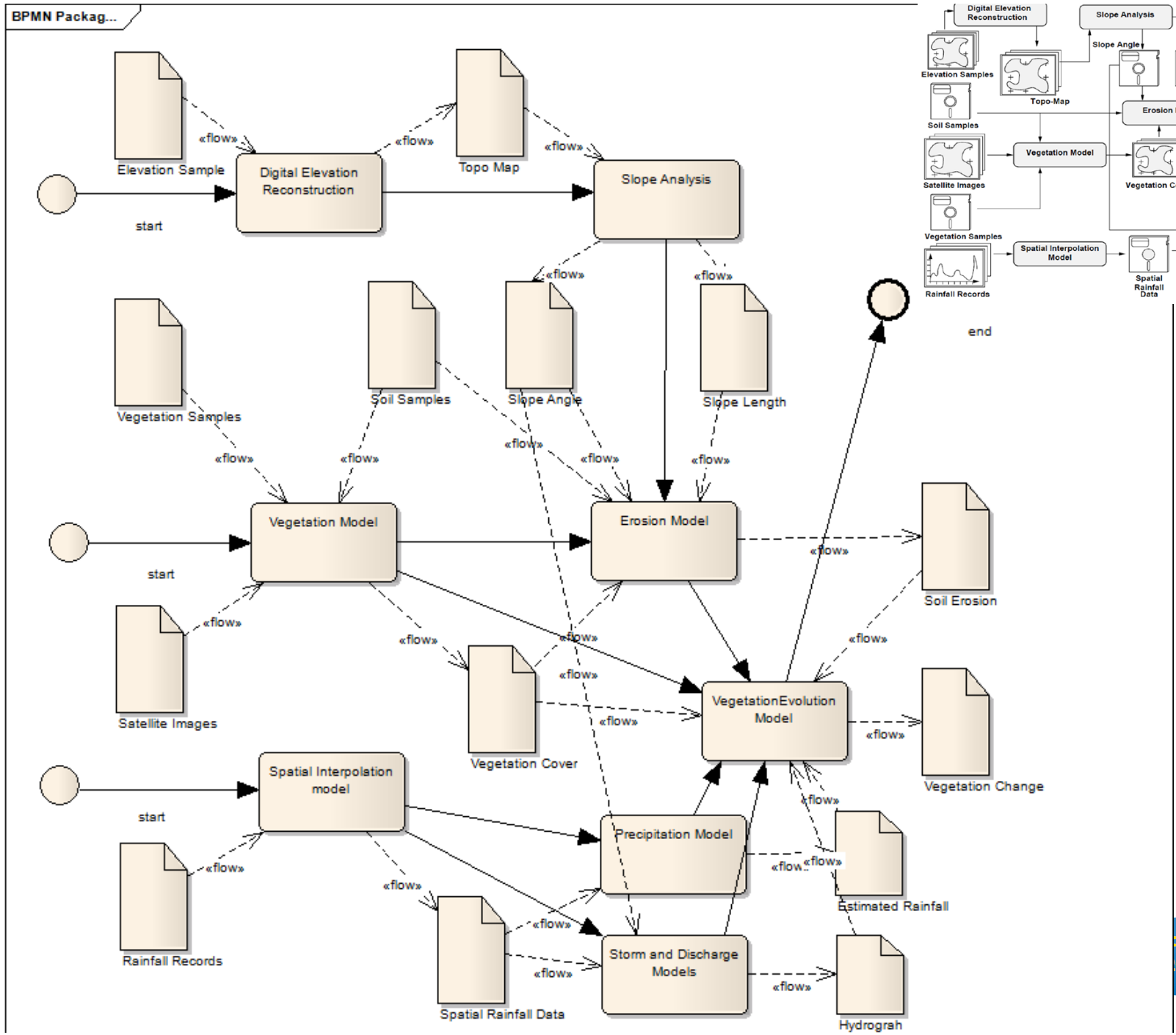
toy example:
greenness model



Data3= P1(Data1, Data2)



Data3= P1' (Data1, Data2, Data7)
Data6= P2(Data3, Data4, Data5)


BPMN Packag...
Elevation Sample
Topo Map
Digital Elevation Reconstruction
Slope Analysis
start
«flow»
Vegetation Samples
Soil Samples
Slope Angle
Slope Length
Vegetation Model
Erosion Model
Soil Erosion
Satellite Images
Vegetation Cover
VegetationEvolution Model
Vegetation Change
Spatial Interpolation model
Precipitation Model
Estimated Rainfall
Rainfall Records
Spatial Rainfall Data
Storm and Discharge Models
Hydrograh
end
Digital Elevation Reconstruction
Slope Analysis
Slope Angle
Slope Length
Soil Erosion
Elevation Samples
Topo-Map
Vegetation Change
Erosion Model
Soil Samples
Vegetation Model
Satellite Images
Vegetation Cover
Vegetation Evolution Model
Precipitation Model
Estimated Rainfal
Vegetation Samples
Spatial Interpolation Model
Storm and Discharge Models
Rainfall Records
Spatial Rainfall Data
Hydrograph

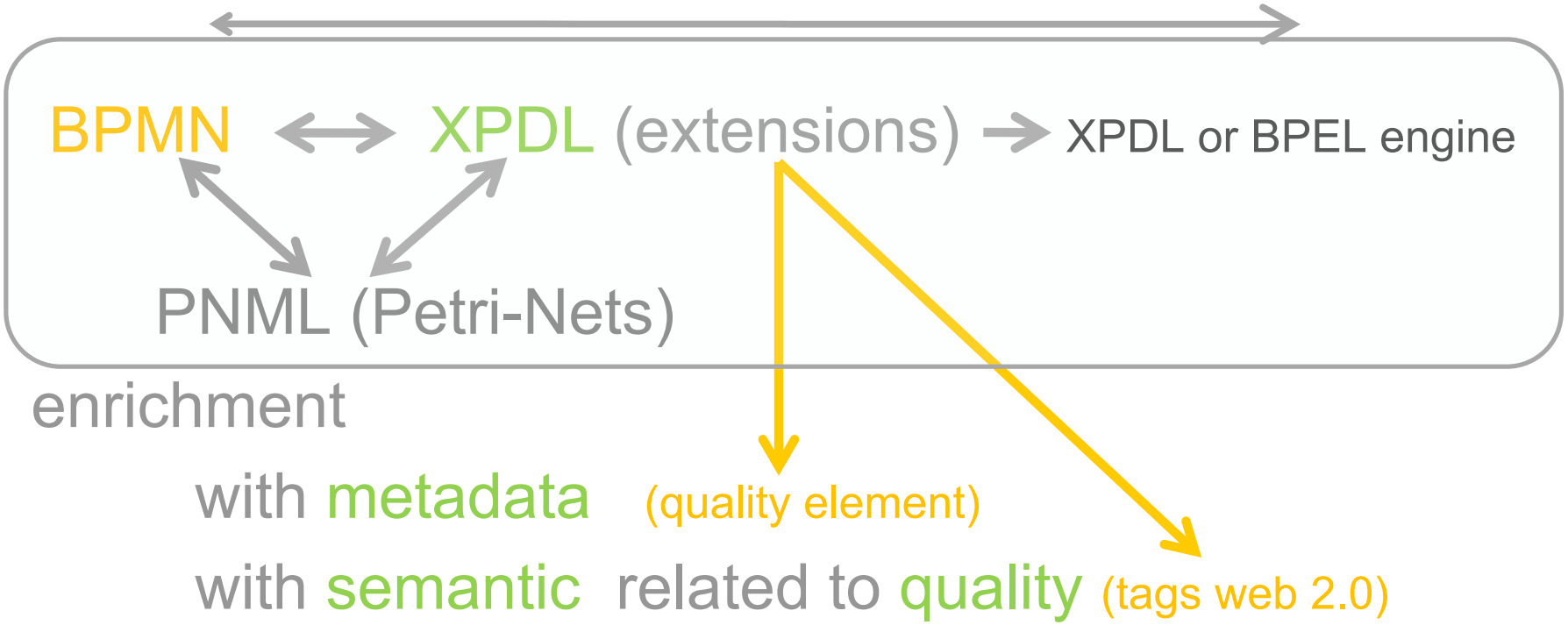
meta-model for workflows

WE SET THE STANDARD

WORKFLOW MANAGEMENT COALITION

- representing / storing & navigate / execute

notation encoding enrichment engine



Centre for Geospatial Science

The University of Nottingham

European Geosciences Union
General Assembly 2011
Vienna | Austria | 03 – 08 April 2011

XPDL profile / extended attributes

- Without namespace

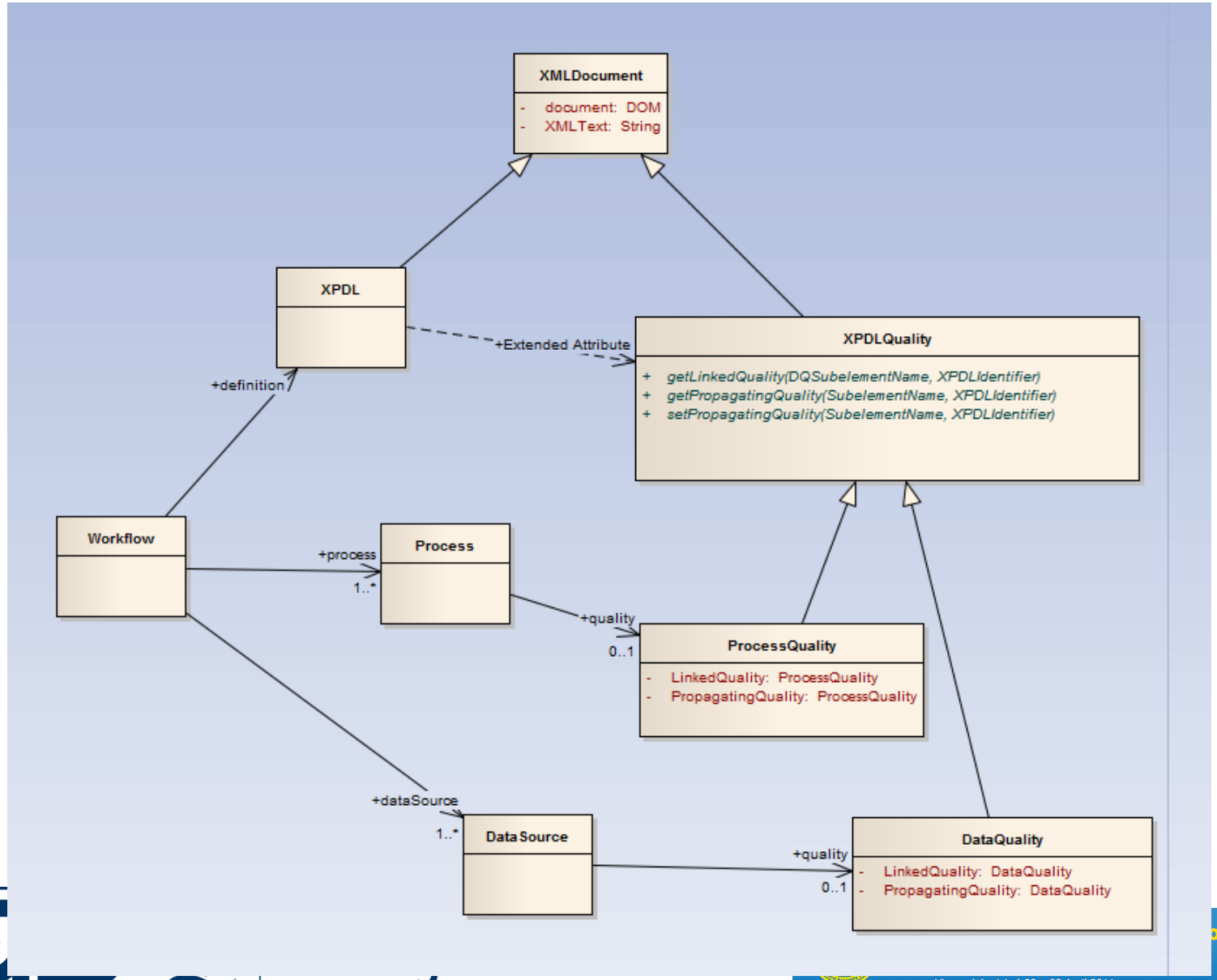
```
<xsd:element name="ExtendedAttribute">
    <xsd:complexType mixed="true">
        <xsd:choice minOccurs="0" maxOccurs="unbounded">
            <xsd:any namespace="##other" processContents="lax" minOccurs="0" maxOccurs="unbounded"/>
        </xsd:choice>
        <xsd:attribute name="Name" type="xsd:NMTOKEN" use="required"/>
        <xsd:attribute name="Value" type="xsd:string"/>
    </xsd:complexType>
</xsd:element>

<xsd:element name="ExtendedAttributes">
    <xsd:complexType>
        <xsd:sequence>
            <xsd:element ref="xpdl:ExtendedAttribute" minOccurs="0" maxOccurs="unbounded"/>
        </xsd:sequence>
    </xsd:complexType>
</xsd:element>
```

- With namespace

```
<xsd:element name="VendorExtension">
  <xsd:complexType>
    <xsd:sequence>
      <xsd:any namespace="##other" processContents="lax"
        minOccurs="0" maxOccurs="unbounded"/>
    </xsd:sequence>
    <xsd:attribute name="ToolId" type="xsd:string" use="required"/>
    <xsd:attribute name="schemaLocation" type="xsd:anyURI" use="optional"/>
    <xsd:attribute name="extensionDescription" type="xsd:anyURI"
      use="optional"/>
    <xsd:anyAttribute namespace="##other" processContents="lax"/>
  </xsd:complexType>
</xsd:element>
```

```
<xsd:element name="VendorExtensions">
    <xsd:complexType>
        <xsd:sequence>
            <xsd:element ref="xpdl:VendorExtension" minOccurs="0" maxOccurs="unbounded"/>
              <xsd:any namespace="##other" processContents="lax" minOccurs="0" maxOccurs=
                  "unbounded"/>
        </xsd:sequence>
        <xsd:anyAttribute namespace="##other" processContents="lax"/>
    </xsd:complexType>
</xsd:element>
```

XPDL profile used in meta-propagation (MetaPunT)



metadata for data and for processes

- ISO standards (data and services)

19115, 19113, 19114, 19135, 19138, 19119, (19139)

ISO 19113 - Quality principles, ISO 19114- Quality evaluation procedures, ISO 19115-Metadata, ISO - 19138 - Data quality measures and ISO - 19135 Registration,

19113 +19114 +19138 = 19157

- UncertML (OGC discussion paper)

encoding uncertainty measures

- for data … what about Geo-processes?

metadata for data quality

Table 1: Data quality elements and data quality sub-elements with definitions (ISO 19113)

Data quality element	Data quality subelement	Definition
completeness	commission	excess data present in a dataset
	omission	data absent from a dataset
logical consistency	conceptual consistency	adherence to rules of the conceptual schema
	domain consistency	adherence of values to the value domains
	format consistency	degree to which data is stored in accordance with the physical structure of the dataset
	topological consistency	correctness of the explicitly encoded topological characteristics of a dataset
positional accuracy	absolute or external accuracy	closeness of reported coordinate values to values accepted as or being true
	relative or internal accuracy	closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true
	gridded data position accuracy	closeness of gridded data position values to values accepted as or being true
temporal accuracy	accuracy of a time measurement	correctness of the temporal references of an item (reporting of error in time measurement)
	temporal consistency	correctness of ordered events or sequences, if reported
	temporal validity	validity of data with respect to time
thematic accuracy	classification correctness	comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset)
	non-quantitative attribute correctness	correctness of non-quantitative attribute
	quantitative attribute accuracy	accuracy of quantitative attributes

metadata for process quality

(proposal)

Table 5: Process quality elements and sub-elements

Process quality element	Process quality sub-element		Definition
conflation	information loss	O	loss in conflating input data sources
	information gain	O	gain in conflating input data sources
conceptual validity	semantic conformance	D	adherence to the semantic relations within the "disciplinary" domain
	domains integration	O	level of integrated modelling in relation to the "disciplinary" domains involved
logical validity	conceptual conformance	O	adherence to rules of the conceptual model
	domain conformance	O	adherence to the output data values of the domain
	computational format	O	degree to which the encoding format follows standards
	topological preservation	O	preservation of the explicitly encoded topological characteristics of the input data sources
		M	propagation of the uncertainty in the

(proposal)

positional error propagation	absolute error propagation	M	propagation of the uncertainty in the absolute positions of features in datasets
	relative error propagation	D	propagation of the uncertainty in the relative positions of features in datasets
	gridded error propagation	O	propagation of the uncertainty in the gridded data position values
	scale preservation	O	preservation of scale(s) of the input datasets
	spatial scale error propagation	D	propagation to outranging scale conformance of input datasets
temporal error propagation	time propagation	M	propagation of the uncertainty in the time measurement
	time scale propagation	D	error propagation due to outranging scale conformance for the input datasets
thematic error propagation	impact of classification correctness	M	propagation of uncertainty due to departure from accurate classification
	impact of non-quantitative attribute correctness	O	propagation of uncertainty due to correctness of non-quantitative attribute
	quantitative attribute error propagation	M	propagation of uncertainty of quantitative attribute

metadata for processes `ProcessQuality.xsd`

- encoding using the same structure as in ISO19139 for data quality

DQ_element

PQ_element

```
PQ_ConflationInformationLoss,
PQ_ThematicClassificationCorrectness,
PQ_QuantitativeAttributeErrorPropagation
PQ_SemanticConformance,
PQ_TemporalErrorPropagation,
PQ_TopologicalPreservation
...
```

- `scope` contain `From` and `To` attributes
- `result (PQ_Result_PropertyType)`

  can be a WPS reference

uncertainty / accuracy /sensitivity

- uncertainty analysis

  what is the output uncertainty?

- **and sensitivity analysis**

  where output uncertainty comes from?

for each atomic process

$$S_{X_i}(Y) = \frac{V(E(Y|X_i))}{V(Y)}$$

1. **collect quality metadata** about inputs (distribution, variance, ...)
2. sampling design accordingly
3. look at output distribution, variance, ... and compare with inputs

Workflow level

A. using the model

B. using an emulator (see UncertWeb project)

C. can we do a simple estimation without 2 and 3?

yes: meta-propagation

- workflow on the **meta information**
- combining metadata about quality of data and processes to derive quality

  of the outputs … of the workflow itself

- error propagation main aspect

  (but not the only one)

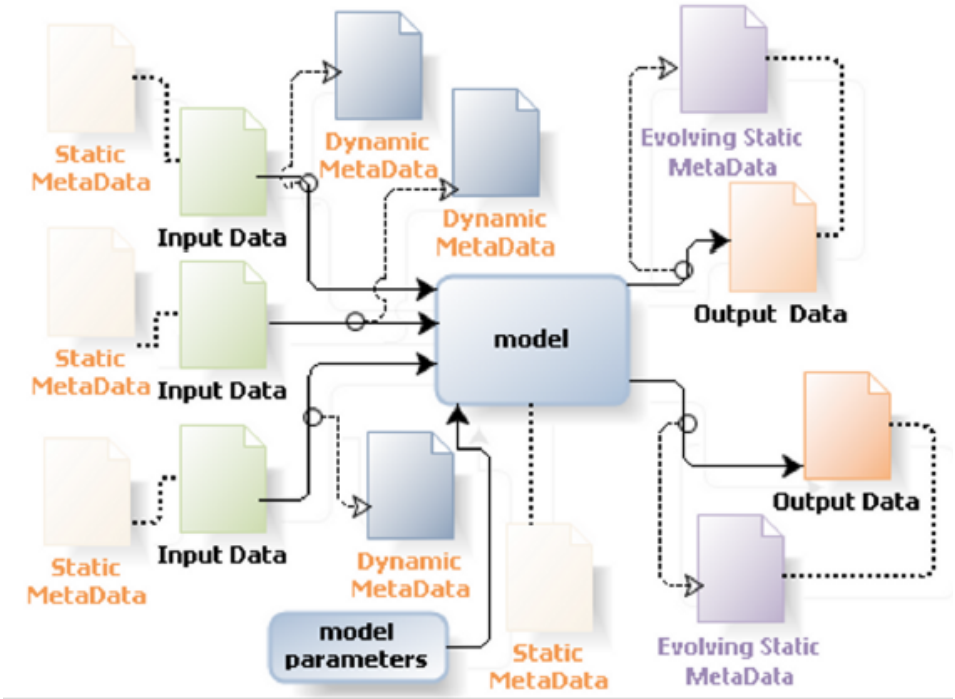
  about quality of a workflow

propagating thematic uncertainty

$\hat{X}1$

$\hat{X}2$

$\hat{X}3$



$\hat{Y}$

$\hat{Z}$

?

$V(\hat{Y})$

variance

DQ_QuantitativeAttributeAccuracy

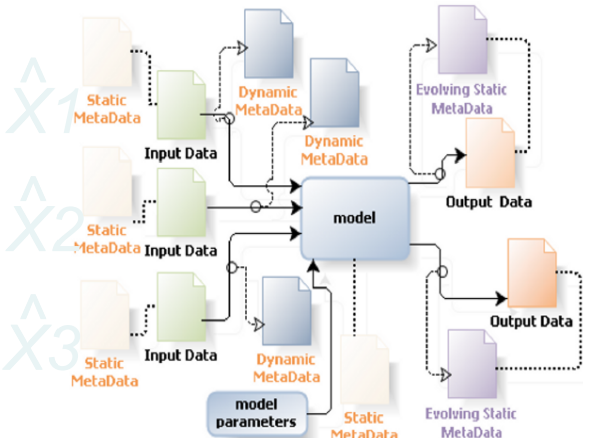
PQ_QuantitativeAttributeErrorPropagation

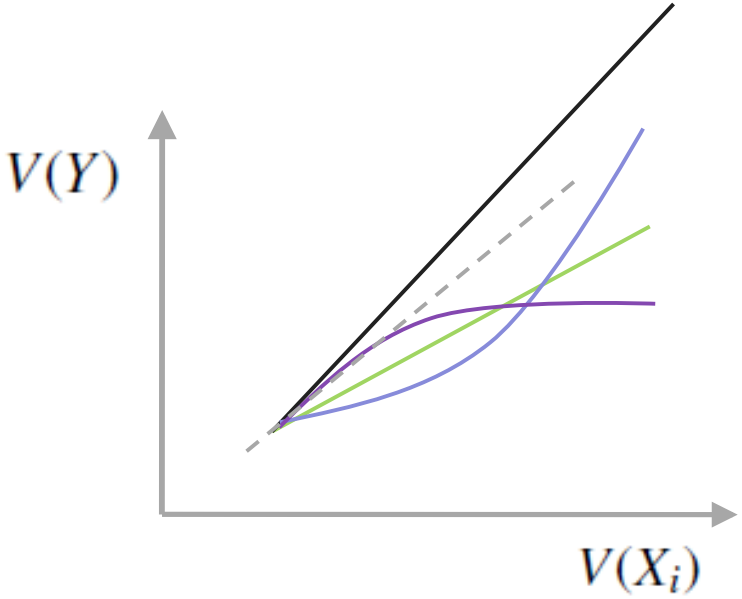
Centre for Geospatial Science

The University of Nottingham

European Geosciences Union General Assembly 2011
Vienna | Austria | 03 – 08 April 2011

propagating thematic uncertainty





$$V(Y) = T_{fX_i}(V(X_i))$$

<< > > ~ =

more than

Sensitivity Information

need a kind
Of "meta"-sensitivity
***i.e.* for various**
sampling
Variances
a variance transfer function

meta-propagated estimate

$$\widetilde{V(\hat{Y})} = \max_{X_i} T_{fX_i}(\hat{V}(\hat{X}_i))$$

Centre for Geospatial Science

The University of Nottingham

European Geosciences Union General Assembly 2011
Vienna | Austria | 03 – 08 April 2011

metadata for processes / basic measures

Table 6: Measures for quantitative attribute propagation: *V(.)* is the variance, *E(.|.)* is the conditional expectation and the subscript "–*i*" means "all expect i", (see Saltelli et al. (2008) for the sensitivity indexes)

Processing quality subelement	Measure		Description	Definition
quantitative attribute propagation	analytical sensitivity	O	$S^{\sigma}_{X_i}(Y) = \frac{\delta(Y)}{\delta(X_i)}(\sigma_{X_i}/\sigma_Y)$ (1)	standardised analytical sensitivity of the output *Y* to the *i*th input variable *Xi*
	partial sensitivity	M	$S_{X_i}(Y) = \frac{V(E(Y\|X_i))}{V(Y)}$ (2)	first order sensitivity index of the output *Y* to the *i*th input variable *Xi*
	total sensitivity	D	$S^{T}_{X_i}(Y) = 1 - \frac{V(E(Y\|X_{-i}))}{V(Y)}$ (3)	total sensitivity of the output *Y* to the *i*th input variable *Xi*
	partial variance transfer function	M	$T_{f\|X_i}$ such that $V(E(Y\|X_i)) = T_{f\|X_i}(V(X_i))$	model of linking a a range of values of variance for *Xi* to the partial variance of Y knowing *Xi*
	total variance transfer function	M	$T_{f\|X_{-i}}$ such that $V(E(Y\|X_{-i})) = T_{f\|X_{-i}}(V(X_i))$	model of linking a a range of values of variance for *Xi* to the variance of Y knowing all but *Xi*
	variance transfer function	M	T_{fX_i} such that $V(Y) = T_{fX_i}(V(X_i))$	model of linking a a range of values of variance for *Xi* to the variance of Y

functionals to be used for quality evaluation

- encoding the variance transfer function

  a WPS as encapsulating the measure

- spatial /a-spatial

  local, pixel, neighbourhood, …

- separablity:

  one to one or multivariable (inputs, outputs)

towards Web Workflow Service?

- WPS for the meta-propagation analysis (XPDL in out)

  MetaPunT 1.0 standalone java

- WPS **acting** alike a workflow service

  WPS GetCapabilities:

  . specific operations stored as available processes (Op)

  . list of the workflows processes (Wkf)

the principle is the Ops informed on a wkf by returning an enriched XPDL file representing the workflow

1. OpShow Id_Wkf returns the XPDL (enriched) of a Wkf
2. OpSet data/processes (modifiable entries of Wkf) returns the updated XPDL file with the updated metadata (particularly propagated metadata)
3. OpExecute, same as OpSet but runs the Wkf as an"aggregated process", returns an XPDL containing as well the links for the outputs.
4. OpStatus returns the status per node of the Wkf in an XPDL file

towards Web Workflow Service?

- WWS
- GetCapabilities OGC generic request
- DescribeProcess (Workflow) request to retrieve the definition of a workflow in a number of standard formats, in which XPDL is the primary choice. It corresponds to OpShow.
- DefineWorkflow like OpSet allowing to set/modify a workflow (fixed workflow witih user's input, partially modifiable workflow with user's inputs and swaps of internal processes or data, or user's workflow)
- Execute (Workflow) as OpExecute launch the execution in "instant" or "delayed" mode, as in WPS and requests the execution status as XPDL or "other workflow format".

Parameters to manage the

- different levels of aggregation/hierarchy (*e.g.* an erosion model may have precipitation model and a run-off model (among other sub-models).

- uncomplete but published conceptual workflows (collaborations)

summary

- integrated modelling /scientific workflow

  model building / reusing / user's perspective /rescaling / quality assessment

- uncertainty / sensitivity analyses for workflows

  error propagation / uncertainty analysis / emulator ("metamodelling") / use of metadata

- metadata for data and for processes

  quality metadata / UncertML / quality principles & measures for processes

- metamodel for workflows

  notation/ encoding/ enrichment

- towards Web Workflow Service?

  WPS / WWS / requirements for workflow assessment

EuroGEOSS
A EUROPEAN APPROACH TO GEOSS

FP7 European project

Centre for Geospatial Science

The University of Nottingham

European Geosciences Union
General Assembly 2011
Vienna | Austria | 03 – 08 April 2011