SimDB: mainly DM

thanks to usual suspects:
Claudio, Franck, Herve, Igor, Laurent,
Mireille, Norman, Patrizia, Rick, Ugo

Overview

- SimDB
- DM
  - domain model
  - latest version)
  - DM changes (wrt SVN 779)
- DM profile
- Serialisations
- TODO
- *Use by SimDAP (see tomorrow)*
- *Prototypes: SimDB browser (see Laurent)*
- SimDB – WG interaction (if time)

Apologies for not being there and hampered preparation

SimDB

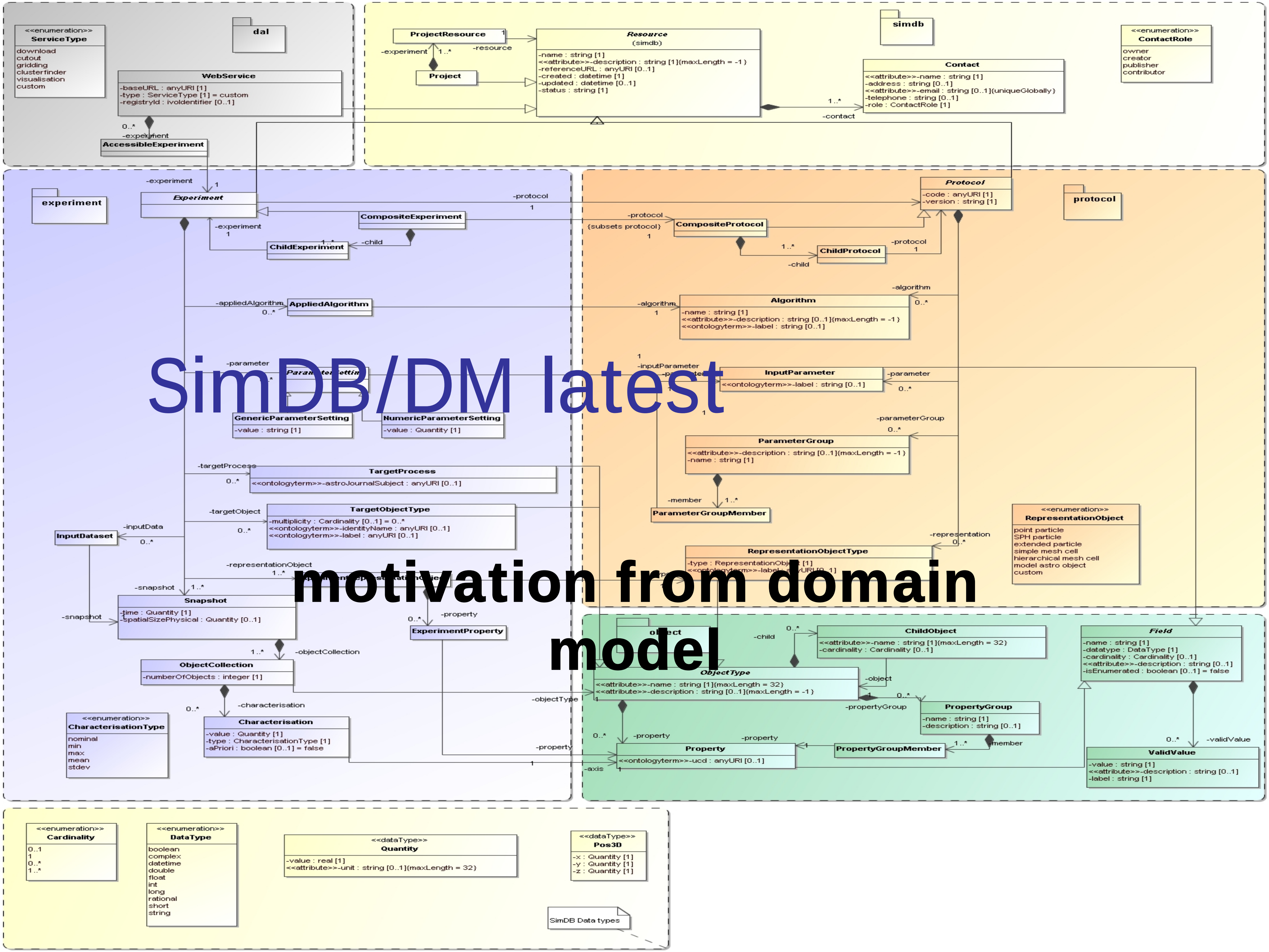
- Protocol for accessing a Simulation Database
- Built around a model for metadata describing *cosmological simulations*
  - i.e. 3+1D
  - not only LSS, Clusters also solar system.
- Used to be SNAP
  - DAL v2-like: based on DM
  - queryData/getData
  - After analysis phase realised things are not so simple:
    - data model complex
    - consequently hard to see simple, parametrised queryData in HTTP: ADQL/browse
    - data sets very large, requires remote filtering, requires interactive getData phase: which subvolume does one want ?
    - custom services important
- Ala registry, but more fine grained.
  - mixture S*AP and registry
  - complex model requires complex support, so maybe only few instances

SimDB as IVOA standard

- Specification for an online web service providing access to a repository storing metadata about numerical computer simulations of astrophysical systems and related resources.
  - Simulation *Registry*, or Simulation *Portal*. Currently the simulations are still
- A SimDB is supposed to be used to
  - *discover* simulations together with web services providing access to them.
- Normative aspects of SimDB:
  - SimDB is based on a (logical) data model, fully specified in UML2.
  - From the UML data model we derive physical models for use in their respective SimDB service contexts:
    - A relational database schema expressed according to the TAP specification.
    - An XML schema, defining valid XML documents containing SimDB meta data descriptions for use in messaging.
    - A set of UTYPEs identifying elements of the model in case this model is to be expressed in VOTables or other non-SimDB-standard representations, e.g. ADQL query results.
    - A human readable HTML document describing all the individual model elements in detail.
  - Physical representations are to be used in the service interface specification of SimDB instances. These are
    - An ADQL-based querying of the metadata repository as a relational database, following TAP.
    - A RESTful web service interface, using standard HTTP methods (GET, PUT, POST, DELETE, etc.) to provide mechanisms for maintaining the actual entries in a SimDB.
    - Possibly an OAI-PMH compliant publishing interface, to allow harvesting of SimDB records.

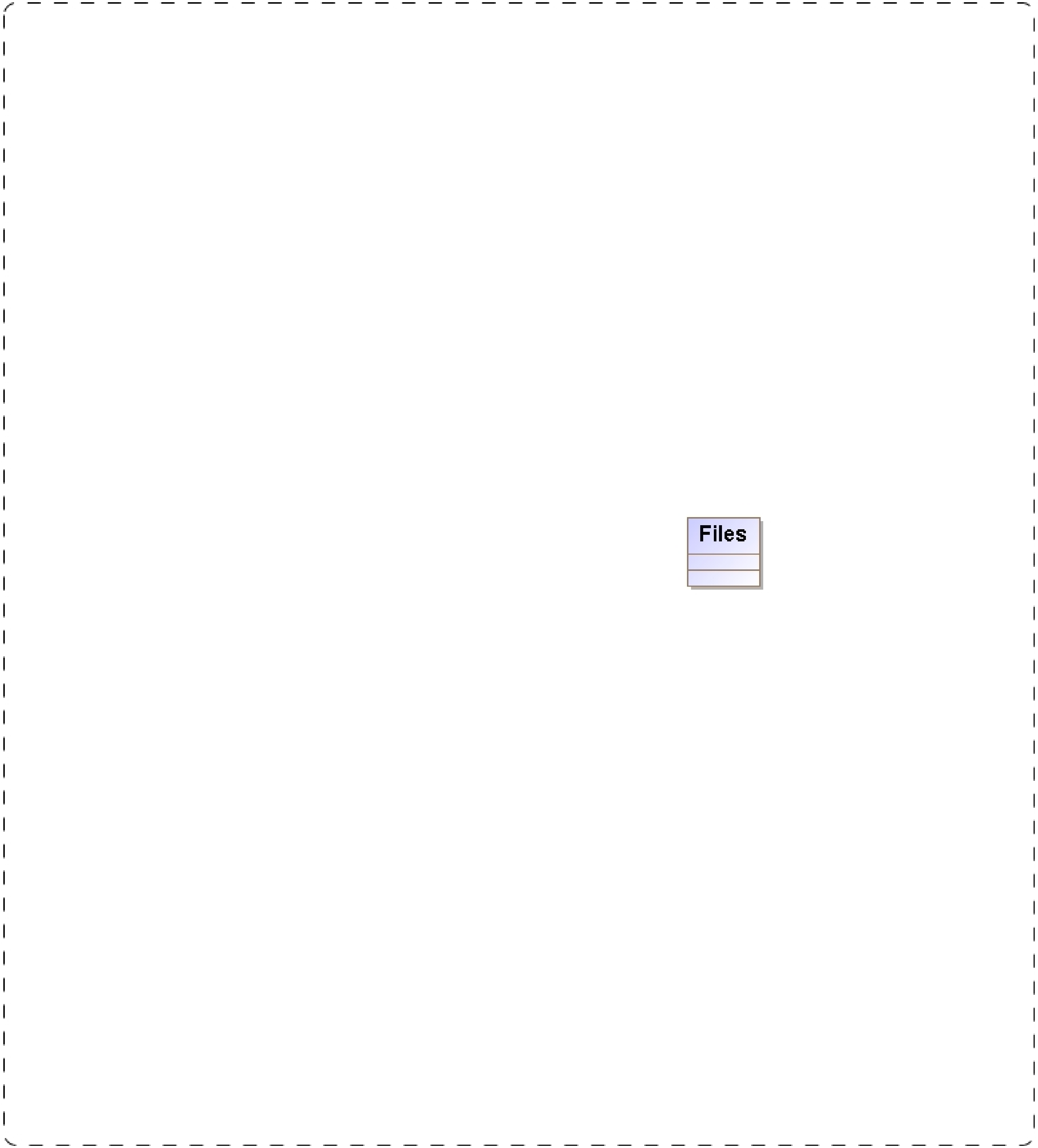
SimDB/DM latest

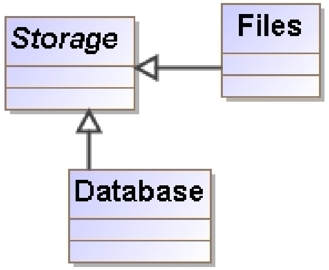
motivation from domain model

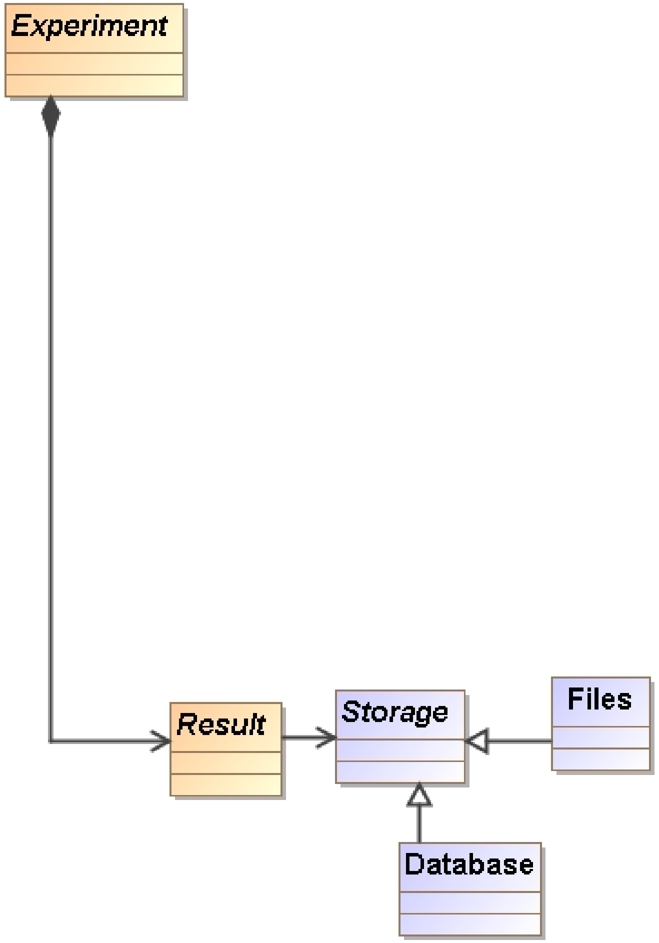


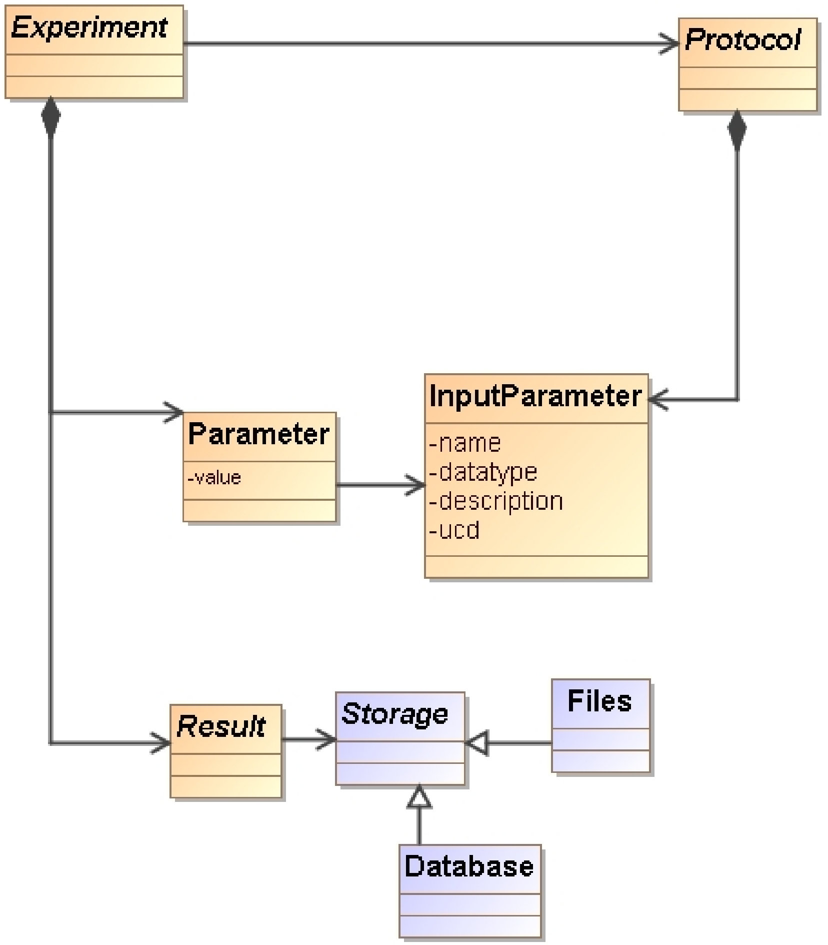
(links)

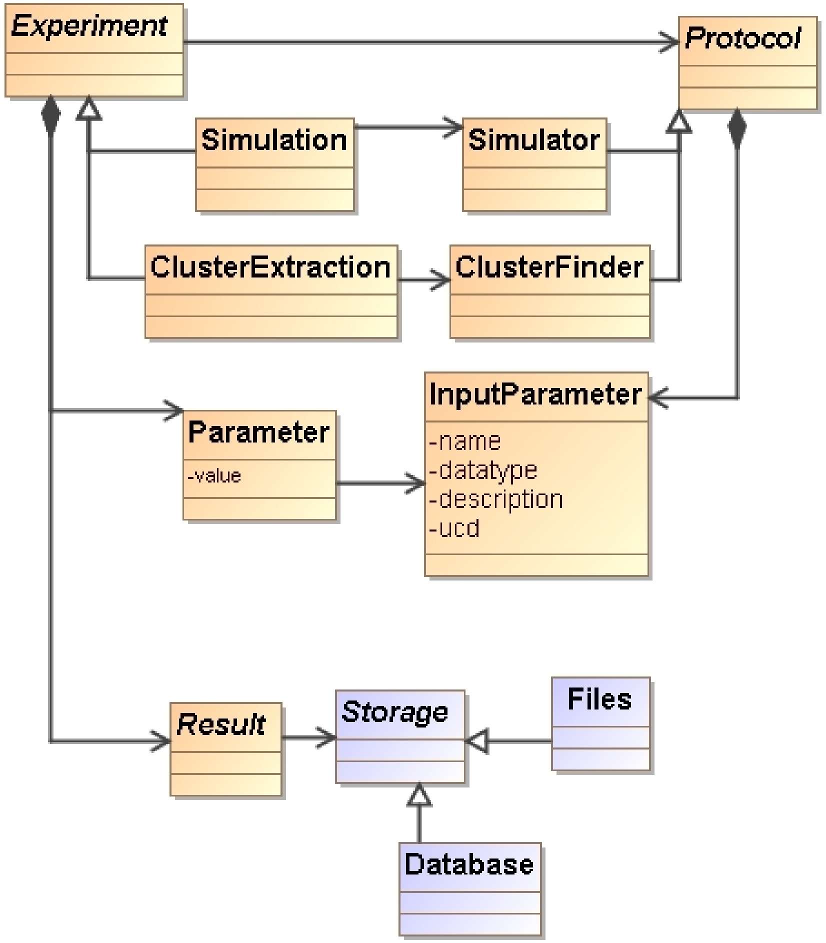
- SimDB on volute:
  http://volute.googlecode.com/svn/trunk/projects/theory/snapdm/
- model (XMI):
  http://volute.googlecode.com/svn/trunk/projects/theory/snapdm/input/SimDB_DM.xml
- generation scripts (XSLT):
  http://volute.googlecode.com/svn/trunk/projects/theory/snapdm/res/
- results (not always in synch):
  http://volute.googlecode.com/svn/trunk/projects/theory/snapdm/output
- HTML doc (generated)
  http://volute.googlecode.com/svn/trunk/projects/theory/snapdm/output/html/SimDB.html


Files


Storage
Files
Database


Experiment
Result
Storage
Files
Database


Experiment
Protocol
Parameter
-value
InputParameter
-name
-datatype
-description
-ucd
Result
Storage
Files
Database



Experiment

Protocol

Simulation

Simulator

ClusterExtraction

ClusterFinder

Parameter

-value

InputParameter

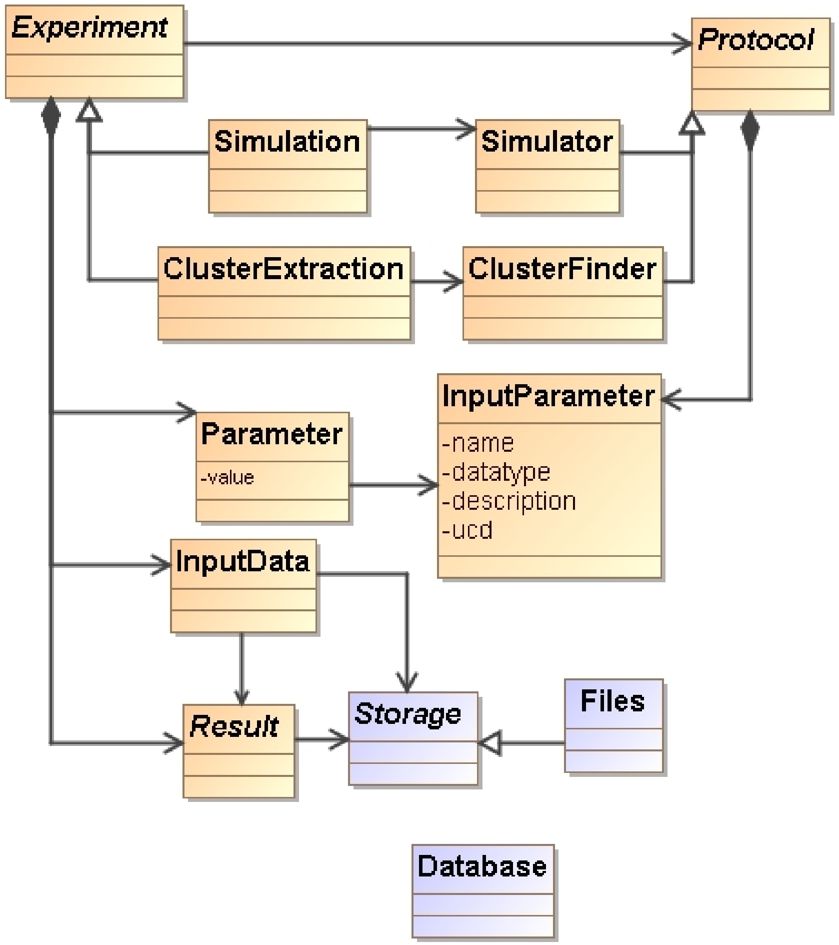
-name
-datatype
-description
-ucd

Result

Storage

Files

Database



Experiment

Protocol

Simulation

Simulator

ClusterExtraction

ClusterFinder

Parameter

-value

InputParameter

-name
-datatype
-description
-ucd

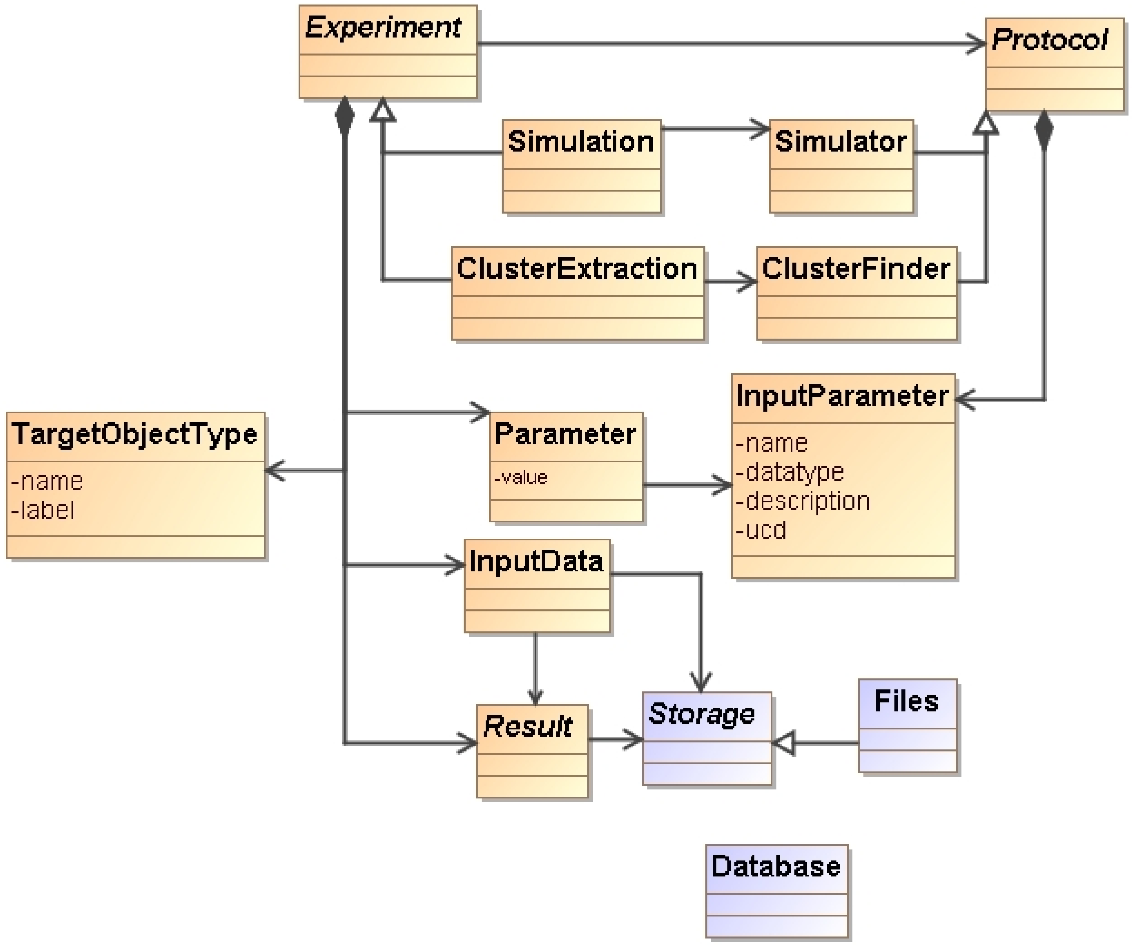
InputData

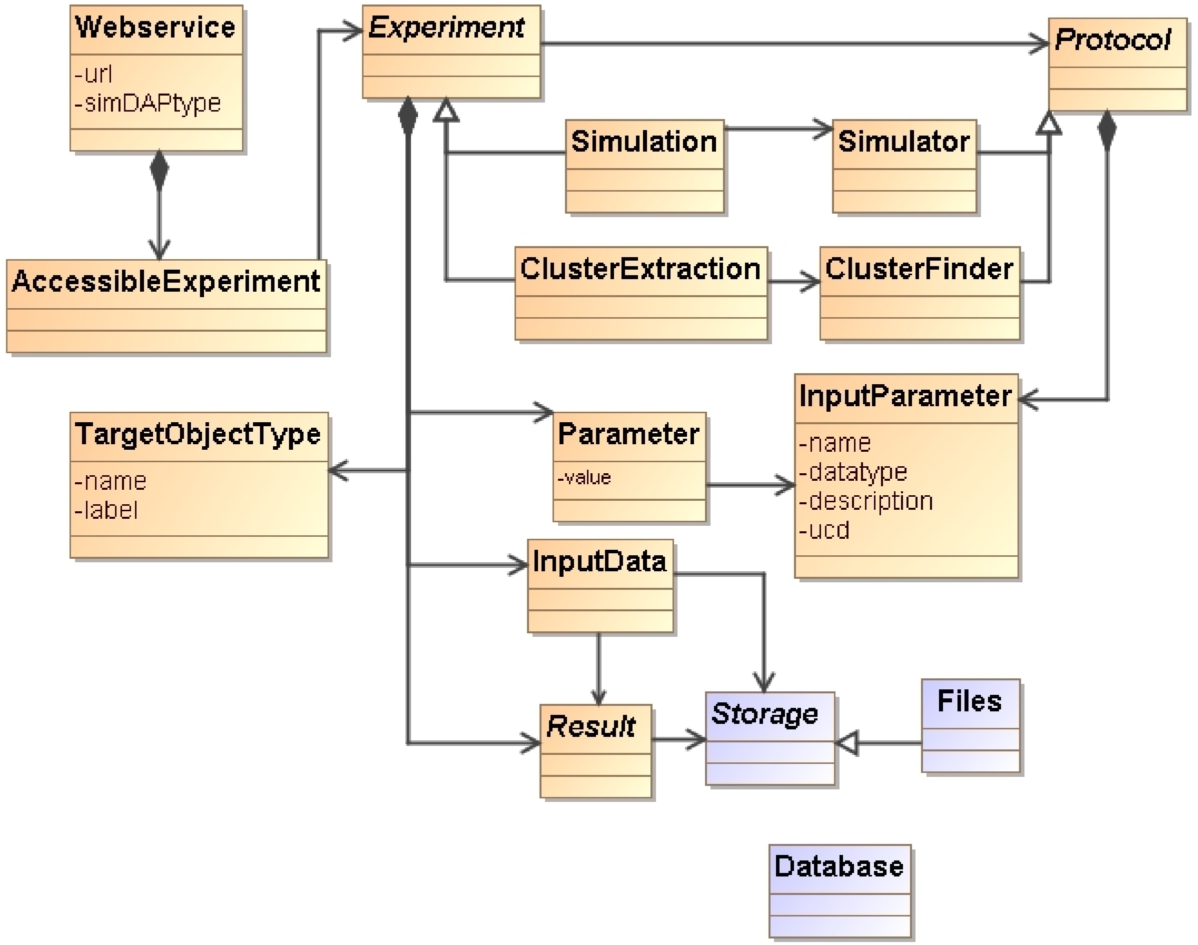
Result

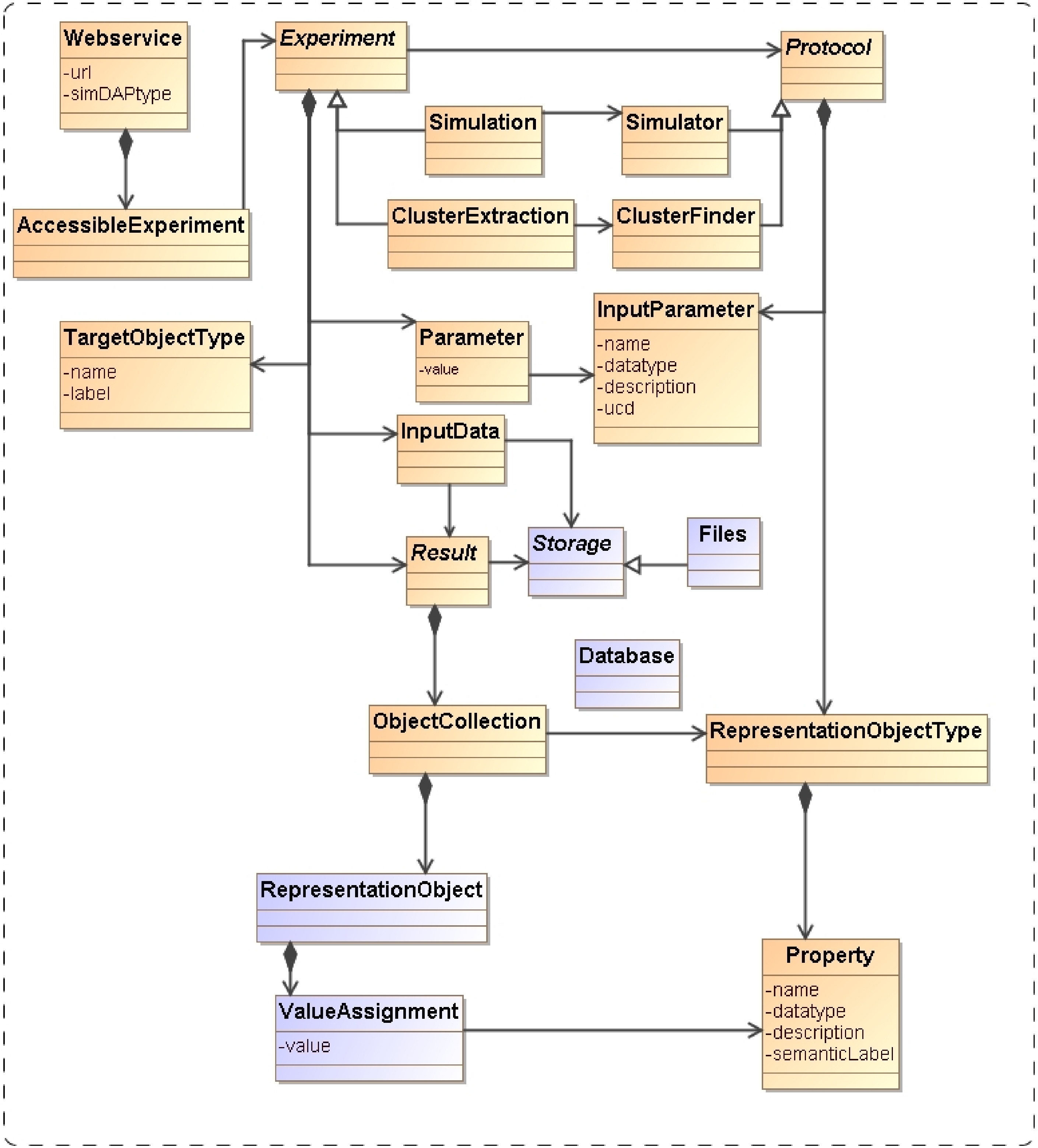
Storage

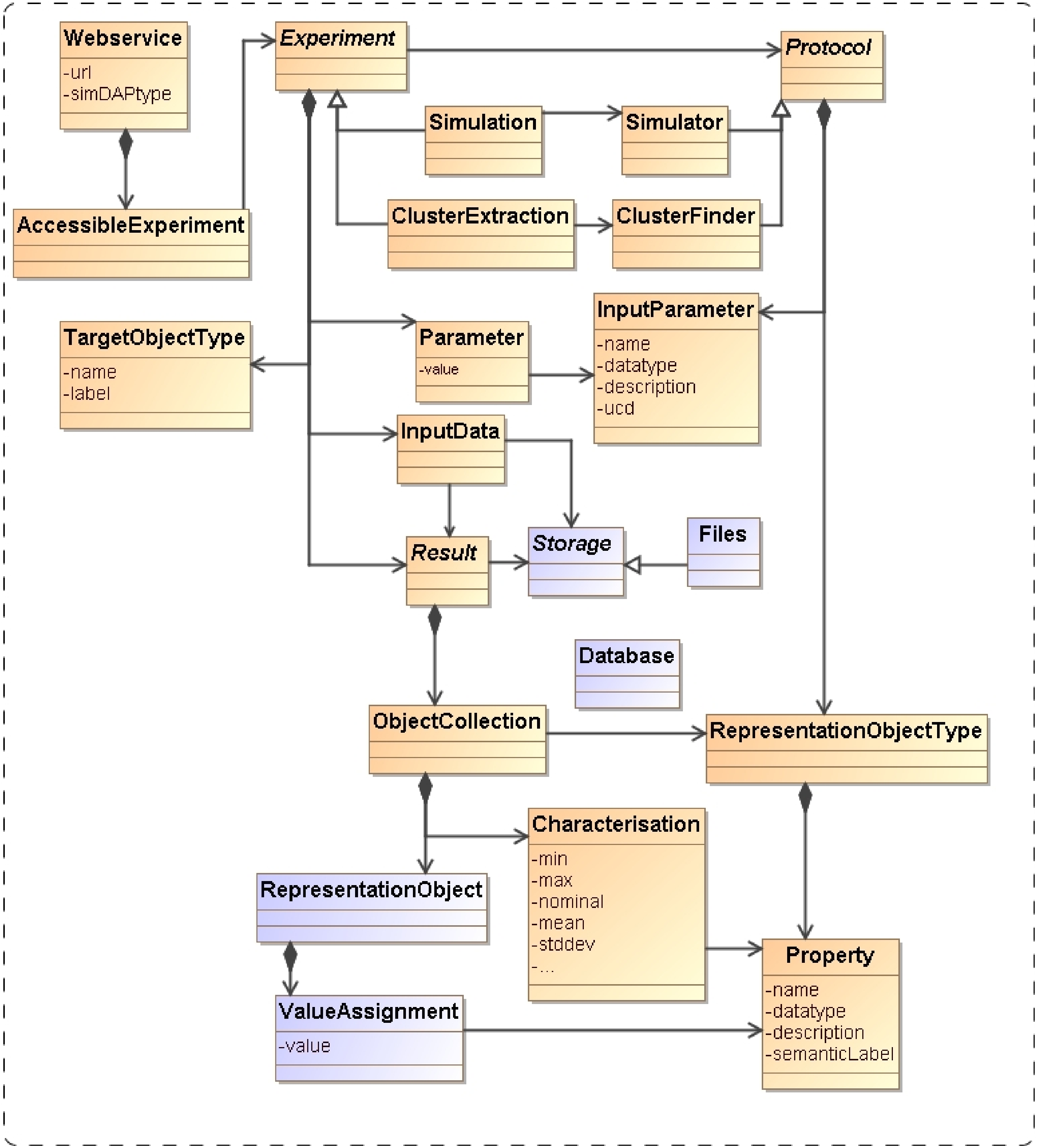
Files

Database


Experiment
Protocol
Simulation
Simulator
ClusterExtraction
ClusterFinder
InputParameter
-name
-datatype
-description
-ucd
TargetObjectType
-name
-label
Parameter
-value
InputData
Result
Storage
Files
Database


Webservice
-url
-simDAPtype
Experiment
Protocol
Simulation
Simulator
AccessibleExperiment
ClusterExtraction
ClusterFinder
InputParameter
-name
-datatype
-description
-ucd
TargetObjectType
-name
-label
Parameter
-value
InputData
Result
Storage
Files
Database


Webservice
-url
-simDAPtype
Experiment
Protocol
Simulation
Simulator
AccessibleExperiment
ClusterExtraction
ClusterFinder
InputParameter
-name
-datatype
-description
-ucd
TargetObjectType
-name
-label
Parameter
-value
InputData
Result
Storage
Files
Database
ObjectCollection
RepresentationObjectType
RepresentationObject
ValueAssignment
-value
Property
-name
-datatype
-description
-semanticLabel



Webservice
-url
-simDAPtype

AccessibleExperiment

Experiment

Protocol

Simulation

Simulator

ClusterExtraction

ClusterFinder

TargetObjectType
-name
-label

Parameter
-value

InputParameter
-name
-datatype
-description
-ucd

InputData

Result

Storage

Files

Database

ObjectCollection

RepresentationObjectType

Characterisation
-min
-max
-nominal
-mean
-stddev
-...

RepresentationObject

ValueAssignment
-value

Property
-name
-datatype
-description
-semanticLabel

DM changes

- Remove Party as root entity, just add Party's info under contacts
- Algorithm under protocol and "algorithm usage" under experiment?
- Web service a root entity class, referencing the experiments it handles? Or the protocol outputs?
- Project collects SimDB/Resources, not only experiments.

- Complete the selection of semantic vocabularies for the different skosConcept files.
- Collection of abstract classes relatively complex to deal with: e.g. ParameterSetting. Also, ParameterSetting with its subclasses harder to query on, need to know what type.
  - 2 collections?
- Get rid of composite protocol and experiment ? NOT YET DONE
- Get rid of Subvolumeextration-or and visualisation/visualiser as protocols.
  Are SimDAP services, latter not even producing snapshots.
  Do we need beyond cluster finder other postprocessing ?
- Discuss level of normalisation and consequences (indirect referencing) (see next)
- ...

Party

- Removed Party as root entity class
  - Every de-normalisation simplifies things
  - (When) will Party-s be reused
- Put info on contact
- Removed main contact
- issue: only 1 role per contact
  - could change model further, harder

Algorithm

- Moved Algorithm to Protocol
- Added AlgorithmUsed as associative class under Experiment
- Issue:
  - Do we want to keep Algorithm?
  - Yes: ClusterFinder can have different Algorithms

Webservice

- Made it a SimDB/Resource
  - Experiment should not have to change if a new service is provided.
- Aggregates Experiments it provides access to
  - But what if it is a query service and new experiments are added to the database, should it point to them as well, …
  - therefore also make possible to ….
- … point to a Protocol if it can handle all of its results
  - like Rick’s proposed Enzo visualisation service.
- But this is still not a complete model …
  - do we need to model some aspects of SimDAP services here?

More

- Removed SubvolumeExtraction/-or and Visualisation/-or.
  - Were place holders, are SimDAP services (and as such protocols etc etc, but not in model maybe.
- Remove CompositeExperiment and –Protocol.
  - what was their use again?

Normalisation

- Domain model was fully normalised
- First SNAP-dm way less so
- We made move back to almost complete normalisation again.
- We may want to move back:
  - parametersetting no reference, but use name
  - same for most other associative objects under experiment
- Issues:
  - redundancy (might just be use of name as foreignkey iso ID)
  - referential integrity not explicitly enforced (no reference!)
  - may need profile change: <<keyattribute>>, <<foreignkeyattribute>>

TODO

- Missing concepts?
- Sufficient attributes?
- Check all descriptions
  - correctness
  - do they make sense as descriptions in a standard doc (some clearly do not)

DM UML Profile

- “profile” includes all UML modelling elements
  - allows adding refinements: <<stereotypes>> with tag definitions
  - pre-defined primitive types (string, integer etc)
- Semantics (how to define values, fixed vocabulary, …) [see Norman’s email]
  - use skosConcept iso ontologyConcept etc.
- Constraints
  - Example: parametersetting→inputparameter as specialisation of experiment → protocol
    Generalise to new type of subsetting/specialisation
  - Other way to specify rules?
- Prescribe ordering of properties?
  only really to reproduce ordering in existing XML schemas (say Registry), in MagicDraw no mixing association-ends with attributes.
- Do we need something like <<xmlAttribute>> to indicate specific generation
- …

Semantics

- propose:
  - change reference to “ontology” to “skos”:
    <<skosconcept>> iso <<ontologyterm>>
    skosvocabulary iso *ontologyURI* tag
- *do* fix a single skosvocabulary for an attribute, i.e. do not leave it free as Norman suggests
  - motivation: ease of use and support
  - Users need to have some vocabulary in mind anyway (otherwise even equivalences won’t help), so why not the one we fix.
  - Leaving it free which vocabulary to use makes no sense as different vocabs have different meanings. Only way would be to still indicate one, but allow all equivalent ones, in which case there better be such and we need query support that can follow such equivalencies.
- We need to find appropriate vocabularies for the various attributes

Association specialisation

- Simulation.protocol → Simulator
  subsets
  Experiment.protocol → Protocol
- One can say that similarly
  ParameterSetting.parameter → InputParameter
  specialises
  Experiment.protocol → Protocol
- for the InputParameter is intended to be the one of the Protocol that the ParameterSetting's Experiment points at.
- Some way to indicate this specialisation might provide a way for this very common design pattern.

Serialisations

- Serialisations (= physical models) are required to be able to do something with a DM
- DM WG mandates UML+XML schema
- We believe that once a proper UML model is defined, the serialisations could (SHOULD) be generated automatically:
  - according to what rules is a task for the DM WG
- In SimDB, serialisations were created using “VO-URP” (UML representation pipeline): rules implemented in XSLT
- We generate:
  - XML Schema
  - DDL(tables and views)+TAP
  - HTML+UTYPE
  - JAVA (JPA+JAXB) (not normative)
- TODO:
  check the rules
  - check whether we may need to make explicit mention in certain cases how mapping was achieved
    - example: in TAP specs, columns are included from base class, these have sometimes descriptions that make sense in the context of the base-class, less when seen in the context of the sub-class. If we make this clear using some textual prefix this may help.

Serialisations (cntd)

- XML:
  - Generated according to rules similar to http://www.ivoa.net/internal/IVOA/VOResource010RevNotes/ModelBasedSchema.ppt
  - Introduce global unique identifiers (“ivoId”) *on all elements*
    - <simdb-ivoid>/<utype>#<id>
  - allow XML ID/IDREF when registering and reference inside XML doc
  - XML references outside of document: using ivoId of remote element
- TAP :
  - standard OR-mapping methodology
  - a view per class, containing all (also inherited) properties as columns
  - issue with foreign keys if references can go accross databases (i.e. SimDB implementations). The latter we so far exclude for SimDB v1.0
- HTML: human readable documentation
- UTYPE: <model>:[package>/]+<type>.<property>
- do we want others as part of the standard (e.g. JSON)

SimDB WG issues

http://www.ivoa.net/Documents/Notes/Theory/SimDBTrack-20080711.pdf

Need interqaction with WGs
2 motivations

- 1. We want to promote SimDB to standardisation track
  - Can not be done by IG (yet?)
  - We see four possibilities:
    - change above statement, i.e. allow a TIG to promote a standard
(good arguments for this!)
    - make TIG a TWG
    - move SimDB to a WG
    - allow creation of project specific focus groups that can promote a standard

2

- SimDB touches upon areas of multiple WGs (details later)
  - whatever solution, we believe the standardisation process should be followed by relevant WGs, possibly with explicit assistance.
  - not because we have no idea what to do, but because we want feedback on our ideas
  - form focus group with participants from each relevant (as judged by you after this presentation) WG.
  - Standardisation process formally requires OK of complete WGs before going to next version (CORRECT?).
  - This we want to either avoid, or at least parallelise.

Overlaps

- DM WG
- Registry WG
- Semantics WG
- DAL WG
- VOQL WG (smallest overlap)
- Theory IG

DM WG 1: Methodology + UML

- Data model is very important in SimDB
  BUT worry is that
  DM WG has really no experience with our approach.
  (what/who is “the DM WG” anyway?)
- Proper modelling methodology
  - Includes explicit goal and use cases for the model itself from the beginning (formally Victoria 2006).
- SimDB first rigorous use of UML in IVOA DM effort (even though decided on its use in Cambridge, 2003)
  - explicitly defined UML profile
  - *all* concepts are in UML
  - transformation rules from UML ⇒ serialisations
  - see
    http://www.ivoa.net/internal/IVOA/InterOpMay2004DataModel/dm-presentation20040528.ppt
    http://www.ivoa.net/internal/IVOA/IvoaDataModel/DomainModelv0.9.1.doc

DM WG 2: Serialisations

- All follow rules UML⇒ ..
  - implemented using XSLT (see …)
- XML schema (references!): follows rules laid out in http://www.ivoa.net/internal/IVOA/VOResource010RevNotes/ModelBasedSchema.ppt see also http://www.ivoa.net/internal/IVOA/InterOpMay2005VOTable/votableProposal.ppt http://www.ivoa.net/forum/votable/0504/0748.htm
- Relational schema
  - data models can be used to design relational databases, which make it easy to query them with ADQL
  - TAP interface easily generated in many forms
- UTYPE (in HTML and TAP)
  - when representing parts of model as a (VO)table, utype-s are useful
  - we propose a rule how to derive them from UML
- HTML : the humanreadable documentation of the DM
- Java (non-standard)

DM WG 3: Other collaborations

- Maybe DM WG can attempt our approach on other efforts …
  - observation
  - provenance (SimDB follows domain model has had provenance fully built in since end 2003 !)
  - characterisation in SimDB based on domain model for characterisation laid out in Beijing

Conclusions: DM

- happy to introduce DM WG to our ideas and participate in evaluating and hopefully formalising our approach.
- Contents of data model can best be evaluated by scientists, more a role for TIG than for DM.
- Links to existing data models are hard to enforce, as other data models are currently very hard to re-use
  - STC's RELOCATABLE is NOT useful for us
  - *Patterns* in Characterisation model has given motivation for SimDB's characterisation, but direct reuse is not useful.
- We need a discussion in DM how to do data modelling and in particular how to reuse models.
- Reusing data models by "import my XML schema" is naive and in general not useful.

Registry WG 1

- Some SimDB/Resources are or can be Registry Resources
  - but not equivalent
  - SimDB instance should be registered
  - SimDB/Project might be a Registry/Resource
  - SimDB/WebService should be registered as well
- Can we define automated SimDB/Resource->Registry/Resource transformation rules (XSLT?)
  - are we missing required metadata (curation?)

Registry WG 2

- SimDB is a fine grained meta-data repository of simulation results and associated SimDB/Resources
  - simulation registry?
  - what issues have you encountered with maintaining distributed registries
  - how about referencing across different SimDB instances? (Enzo simulation code registered in SD, Enzo simulations registered in Italy)
- Can a SimDB be turned into a harvestable registry
  - note, we propose ADQL for querying

Registry WG 3

- SimDB/DM has relatively high level of normalisation: many references (“shared binary associations”)
- Serialisation to XML can not always use ID/IDREF
- We propose use of IVO Identifiers
  - Must be resolved by SimDB, anyone else?
  - Have reference implementation working with this
  - Uses <simdb-ivoid>/<UTYPE>#<primarykey>
  - Can we just use syntax?
  - What other issues play a role?

Conclusions Registry

- Do we want to see SimDB as a fine grained registry? If so, need to evaluate data model to see it is compatible with Res DM.
- Need to get some new catagories in Reg-s to register SimDB-s, SimDB/Resources
- How about Identifiers?
- Possibly registry could follow VO-URP way, get ADQL for free!
  - order of elements, use of attributes as in Registry Schema currently not supported in VO-URP, so can not guarantee same schema, definitely equivalent one!
- feedback once we start going towards implementations.

Registry WG 4

- Any interest from Registry side in using our (LB, GL) VO-URP approach?

Semantics

(already feedback from Norman Gray)

- SimDB requires predefined semantic vocabularies to give valid values to “semantic concept” attributes.
  - eg. Property.ucd
- We propose a way to incorporate this in UML
  - Really technical issue between DM and Sem WG, but not proposed before. (so why not take our solution?)
- Questions remain
  - SKOS sufficient for our goals?
  - which vocabularies (create our own?)
  - how to use
  - a *fixed* vocabulary per semantic concept
    - easier to query and register
    - may be less ontological

Conclusions: Semantics

- problem quite well defined
- answers require knowledge of use cases
- may need an effort in … (where TIG, Sem ?) to define new vocabularies, say for algorithms, (computational) physics

DAL

- We allow querying using ADQL (well, SQL; see Laurent's demo)
- We can accommodate every TAP metadata standard
- We would like to simplify our life by not being forced to implement asynchronous querying.
- Please give feedback on our approach as a DAL effort:
  - build a data model
  - queryData using ADQL against TAP version of dm (maybe a standardised ParamQuery possible?)
  - queryData *response* serialisation using XML version of dm
  - getData separate standard, SimmDAP?
  - Could try this for newer versions of SSA, SIA; especially SCS with a Source data model could be tried!

VOQL

- We want to use ADQL for querying metadata databases
    - any problems foreseen?
    - no XPath support foreseen, could we support this?
- We propose a UTYPE serialisation
    - VOQL was interested in UTYPE

TIG

- Need feedback from scientists
- Promote uptake
  - prototypes with assistance to register resources
  - simulation/post-processing pipelines could produce appropriate documents

- Should remain owner of the project
  - act as client asking something of WGs.
  - should decide when they are happy.
  - so **TIG** should be in charge of proposing to exec that a standard should be promoted. Not nay of the WGs

Next steps

- Standardisation process: meeting late this afternoon of TIG with WG chairs.
- Finalise the DM
  - May require iteration with prototypes and test cases
- Work more on the protocol
- Currently 2 Notes
  - original one: DM+protocol
  - new one: DM+serialisation only
    - in progress,
    - being read (I hope),
    - being co-written (I hope !!)
- prototypes
  - LB, GL; RW; others?
- DISCUSSION !