Metadata Working Group Report

- People
  - **Convener**
    - Tomoteru Yoshie (Japan)
  - **Members**
    - Chris Maynard (UK)
    - Paul Coddington (Australia)
    - Jim Simone (USA - SciDAC)
    - Robert Edwards (USA - SciDAC)
    - Giuseppe Andronico (Italy)
    - Dirk Pleiter (Germany)
    - Balint Joo (UK)

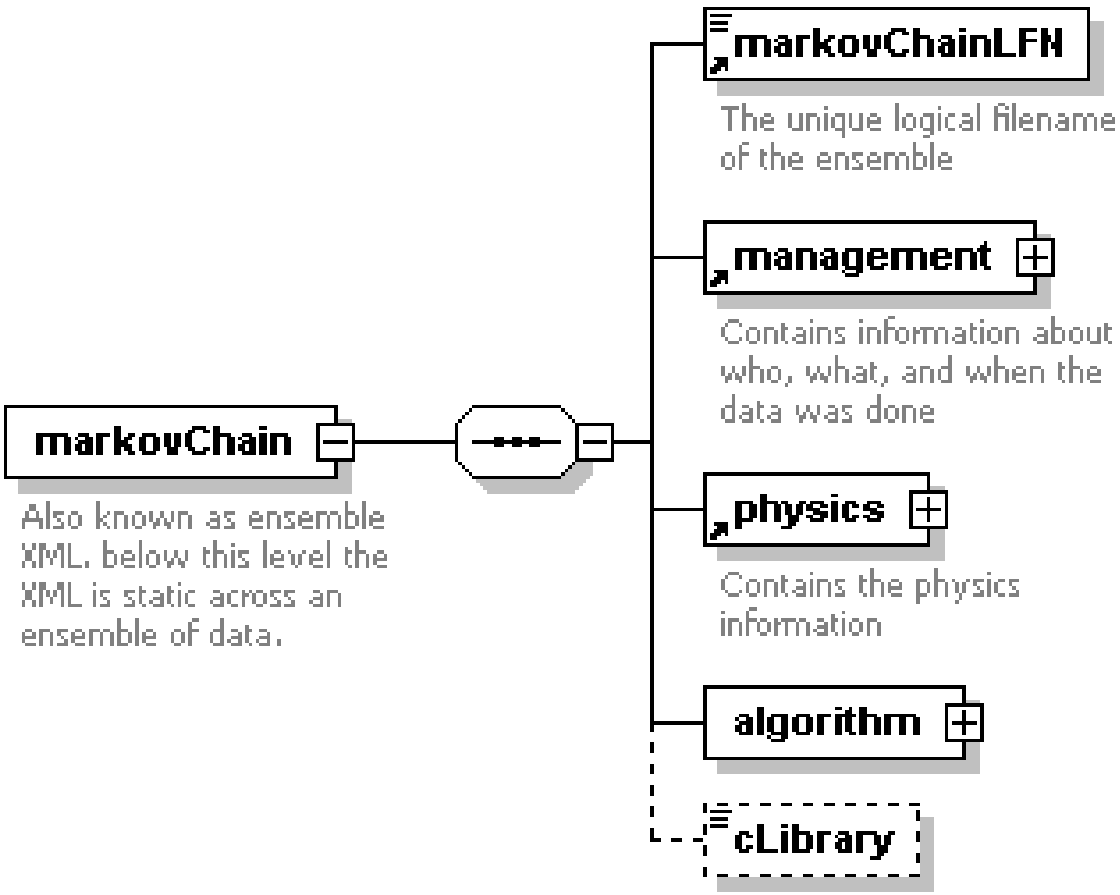
Contents

- QCDML0.4 design and schema
- Propose ILDG adopt this schema
  - QCDML1.0
- How we might proceed to extend QCDML
  - Derived lattice data
  - Gauge fixed cfgs
- BinX
  - Uses and examples

Ensemble and configuration

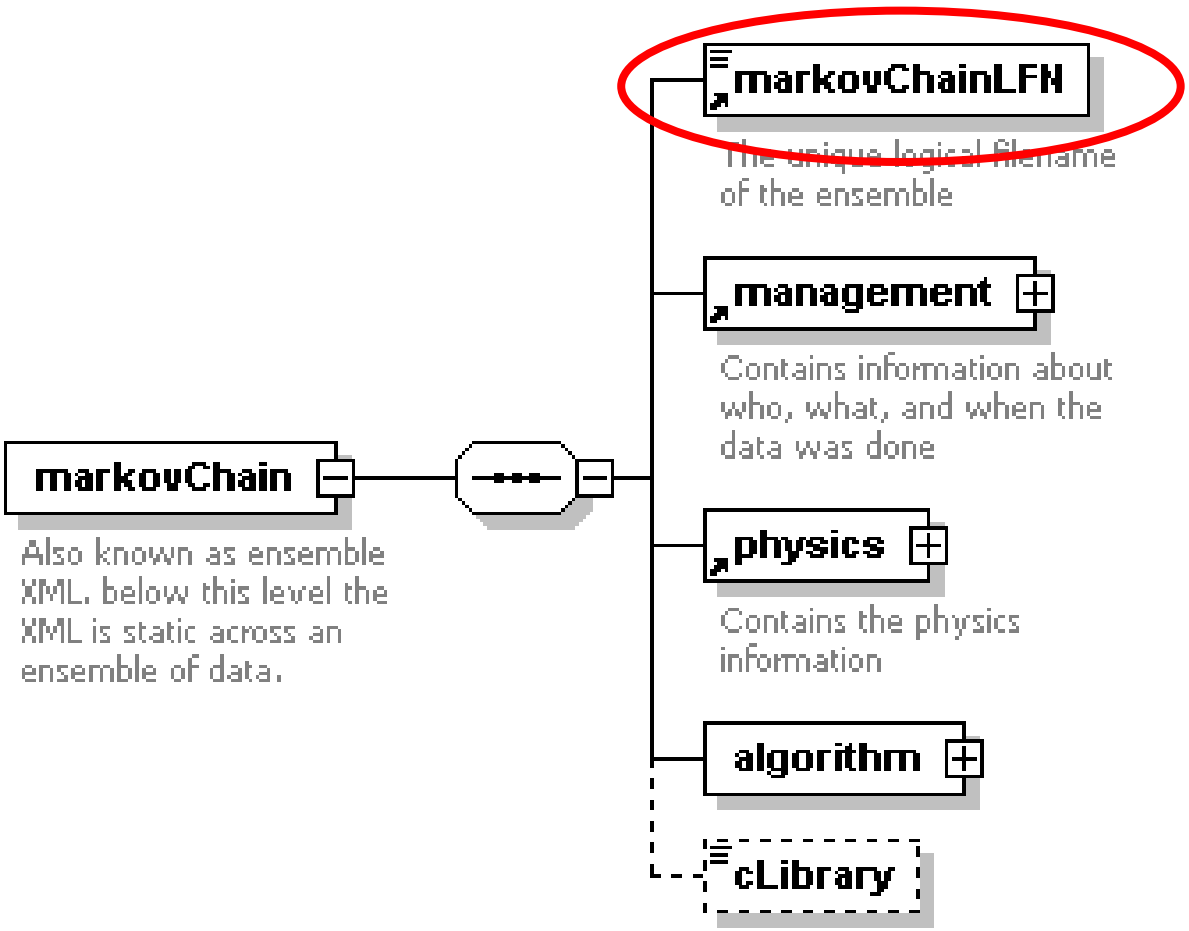
- Most metadata is common to all configurations in an ensemble
- Separate metadata into
  - Ensemble XML **<markovChain>**
  - Configuration XML **<gaugeConfiguration>**
- QCDML is made from two schemata
- Some metadata does not unambiguously belong to either namespace

Ensemble XML



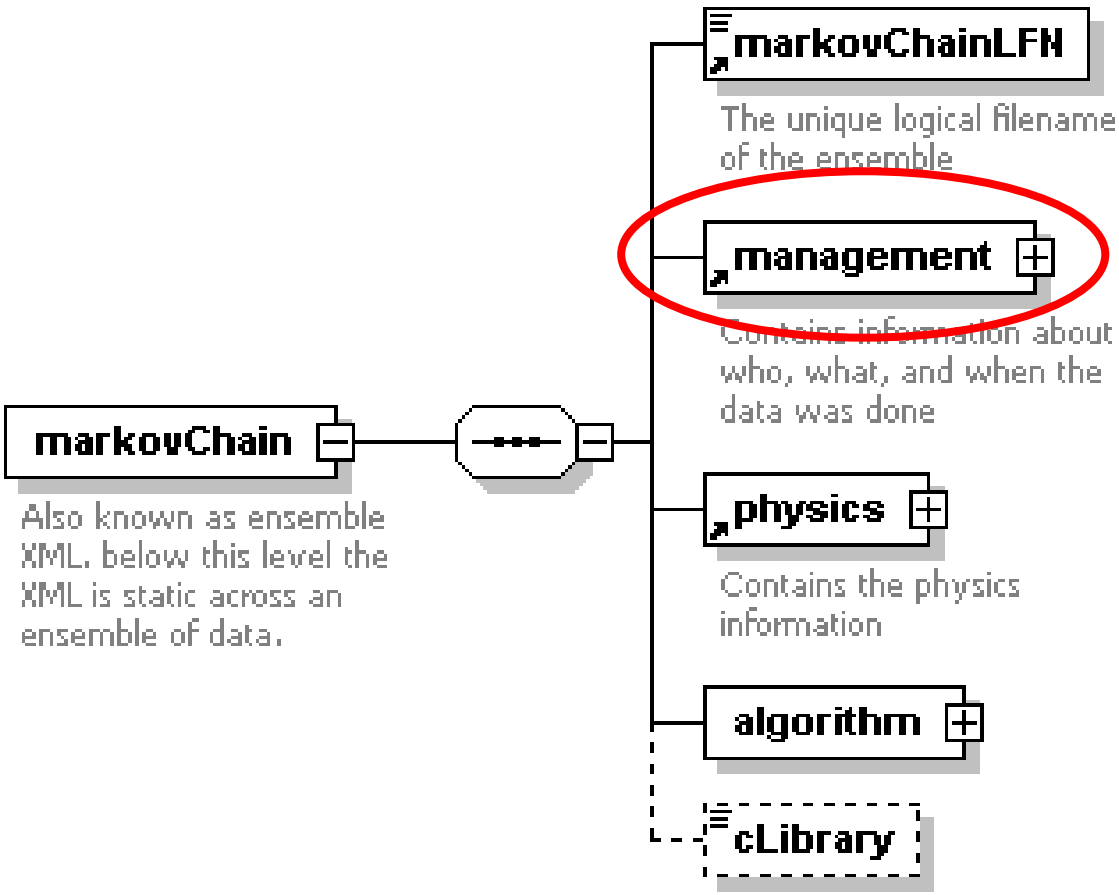
UML representation of XML schema

markovChainLFN



URI uniquely identifies the ensemble in the ILDG namespace

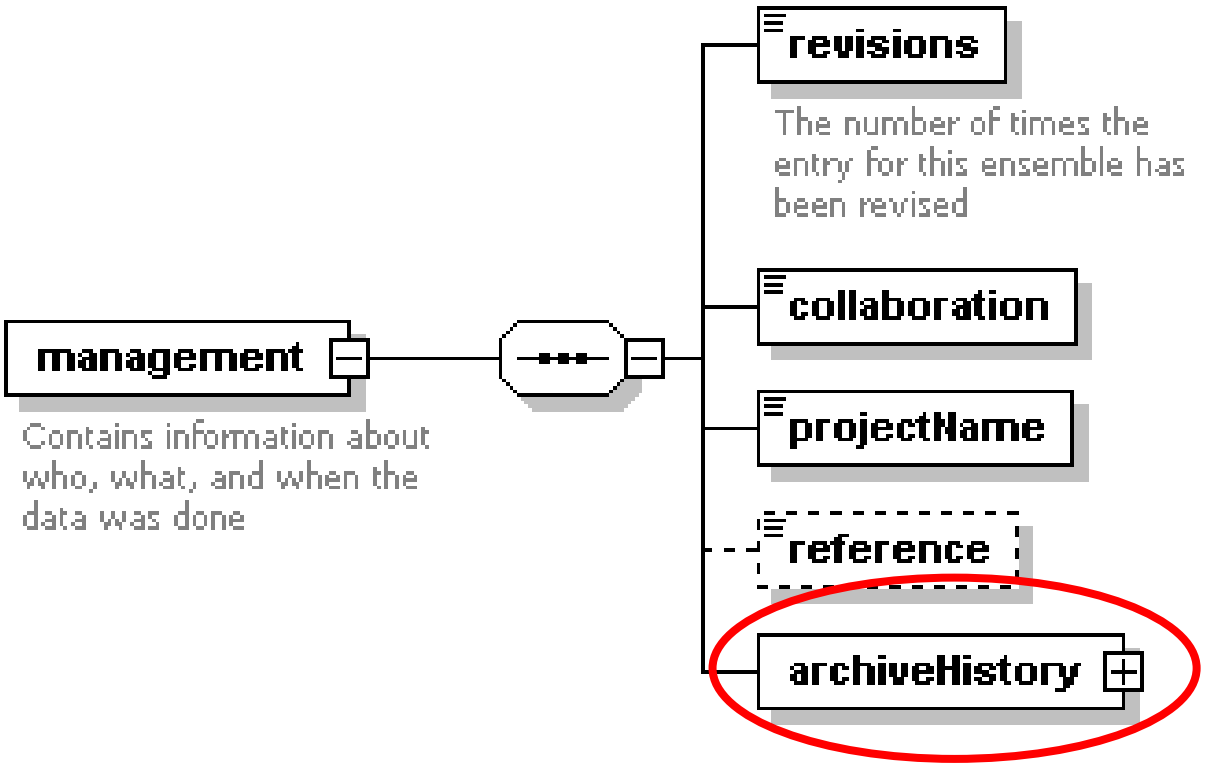
Management of the ensemble



Who, when, and what changes to the ensemble.

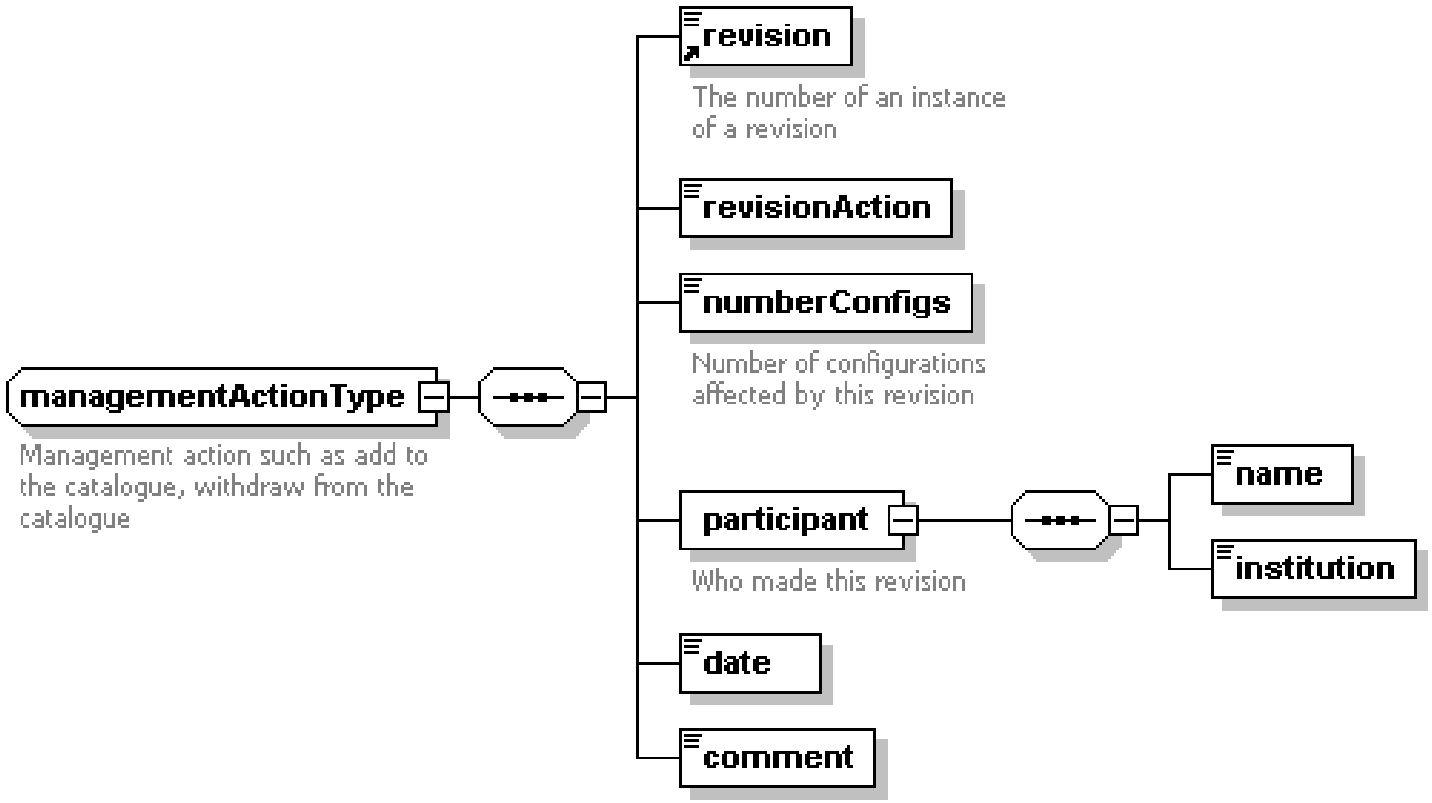
The management information is split between ensemble and configuration

Changing the ensemble



Archive history

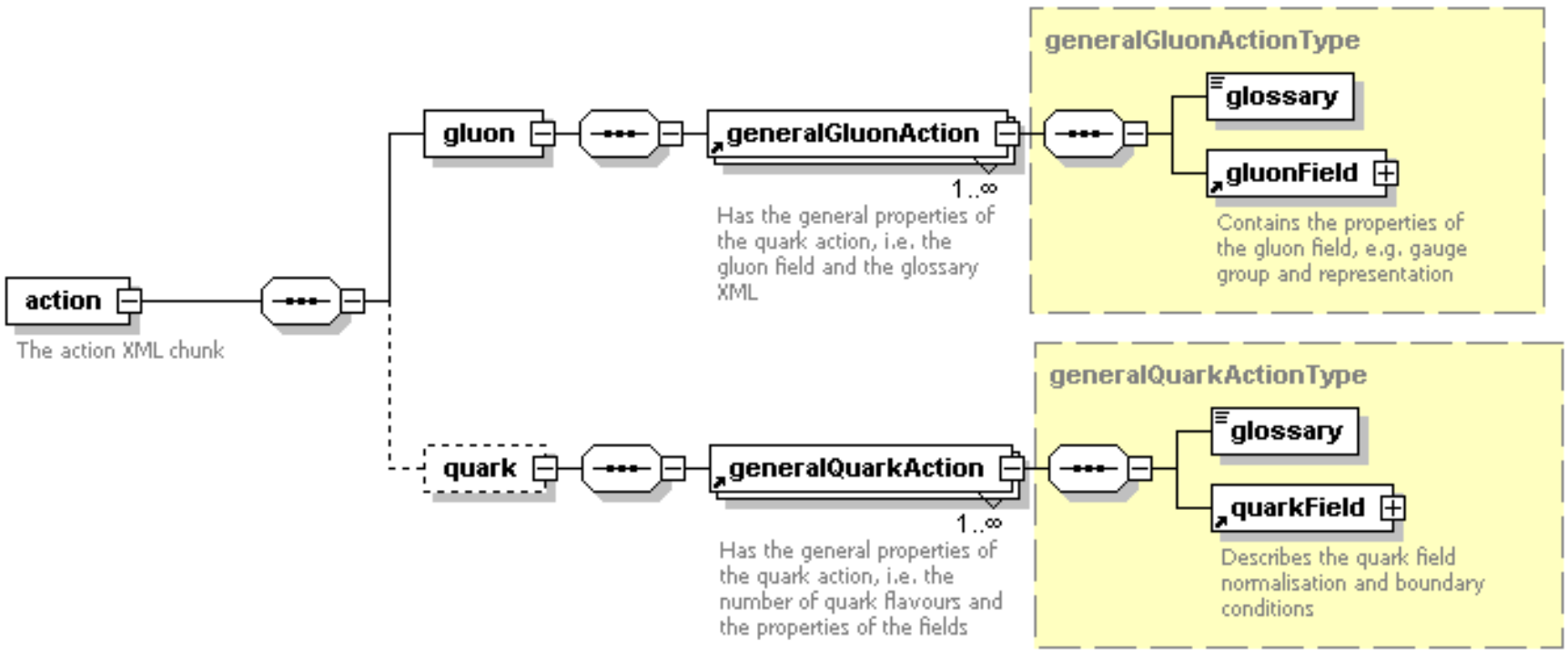
An array of …



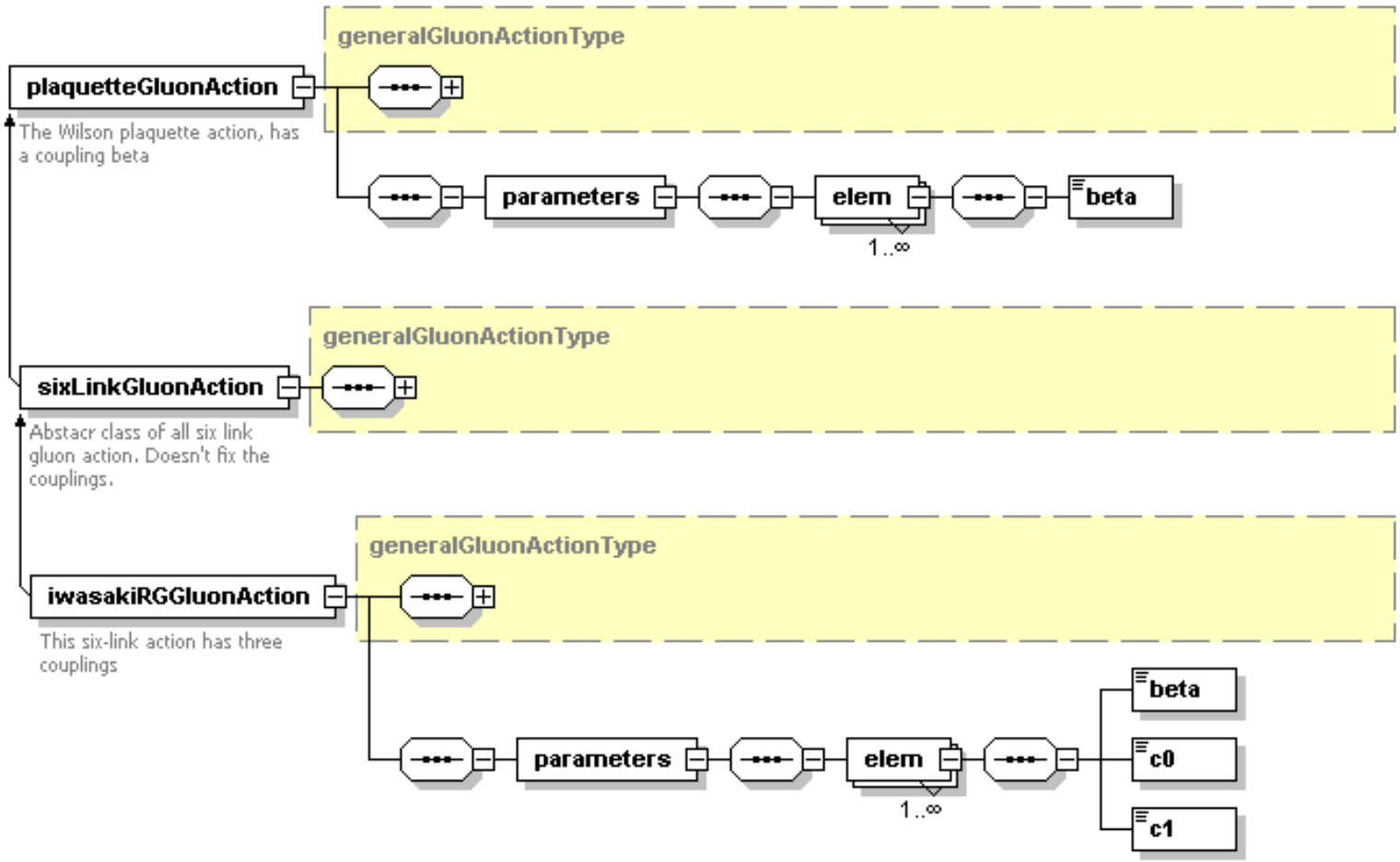
Action

- Most searched metadata
- Critical that data is …
  - Readily searchable
  - Easily extensible
  - Complete
    - All the information required to specify what a gauge configuration is
- Structure required
  - In the schema rather than XML ID

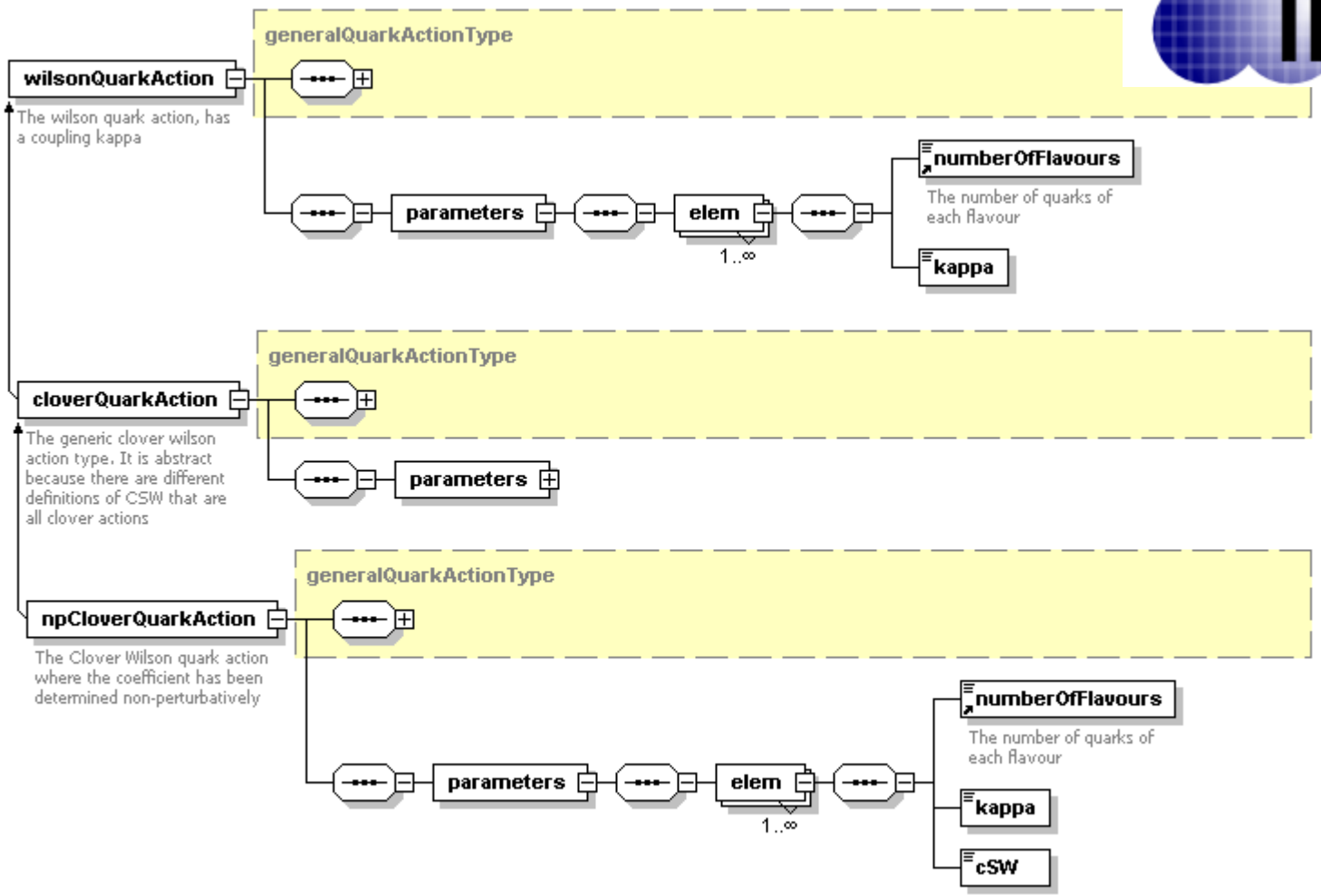
Generic action



Gluon inheritance



Quark Inheritance



Non-degenerate quarks

XML chunk from N_f=2+1 clover

```
- <elem>
    <numberOfFlavours>2</numberOfFlavours>
    <kappa>0.1350</kappa>
    <cSW>2.01752</cSW>
  </elem>
- <elem>
    <numberOfFlavours>1</numberOfFlavours>
    <kappa>0.1340</kappa>
    <cSW>2.01752</cSW>
  </elem>
```

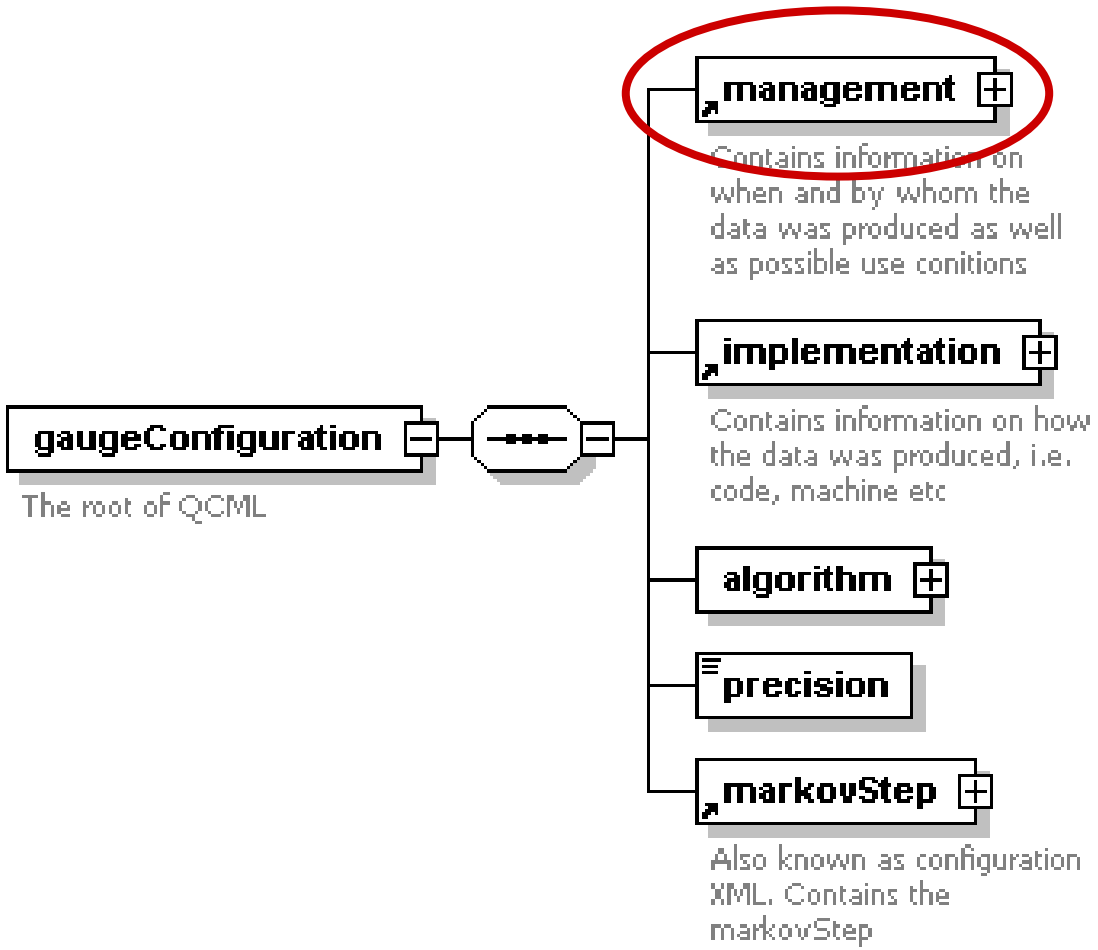
`<parameters>` is array valued

count `<numberOfFlavours>` with XPath query

Algorithm

- Algorithmic metadata split between ensemble and algorithm
- Most metadata is unconstrained parameter **<name/> <value/>** pairs
- Relevant information can be found
- Hierarchical structure for algorithms is
  - difficult to create
  - difficult to make extenisble
  - not that useful

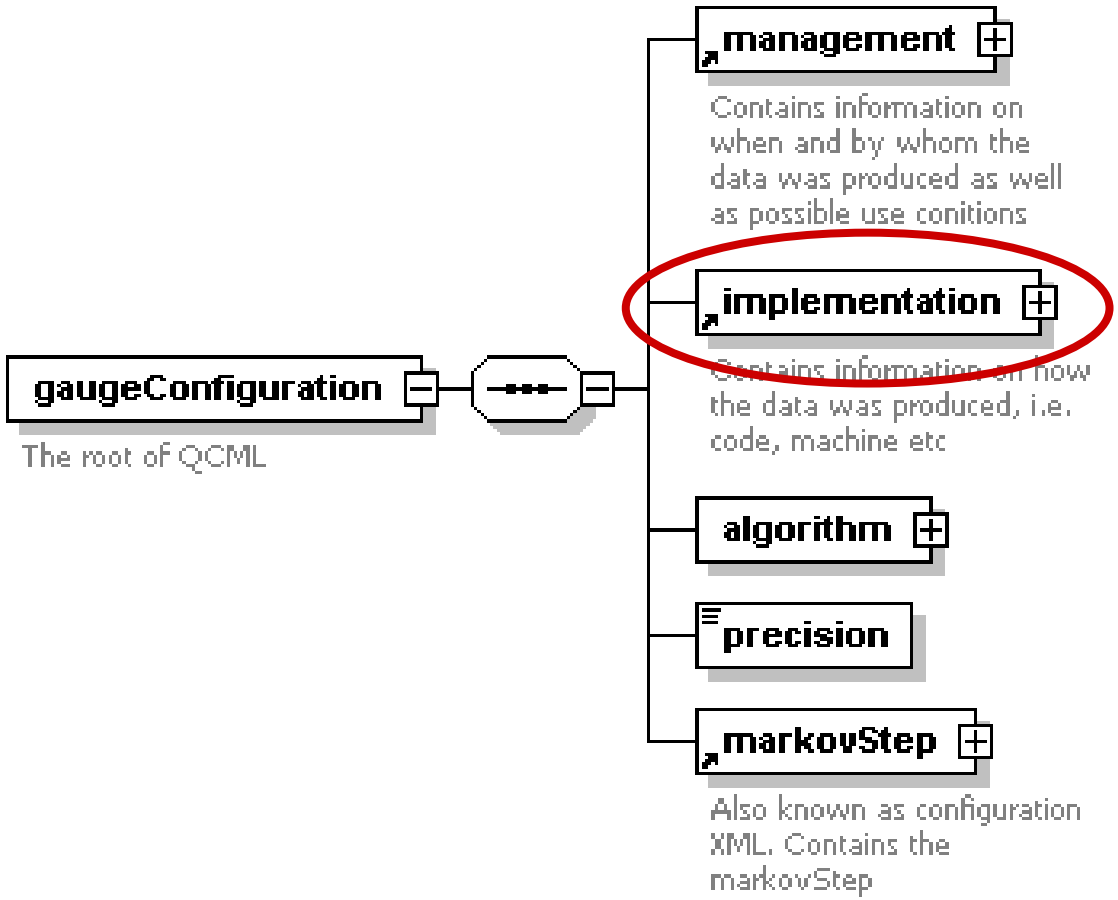
Configuration



Contains the management information for individual configurations

Same structure as the ensemble management

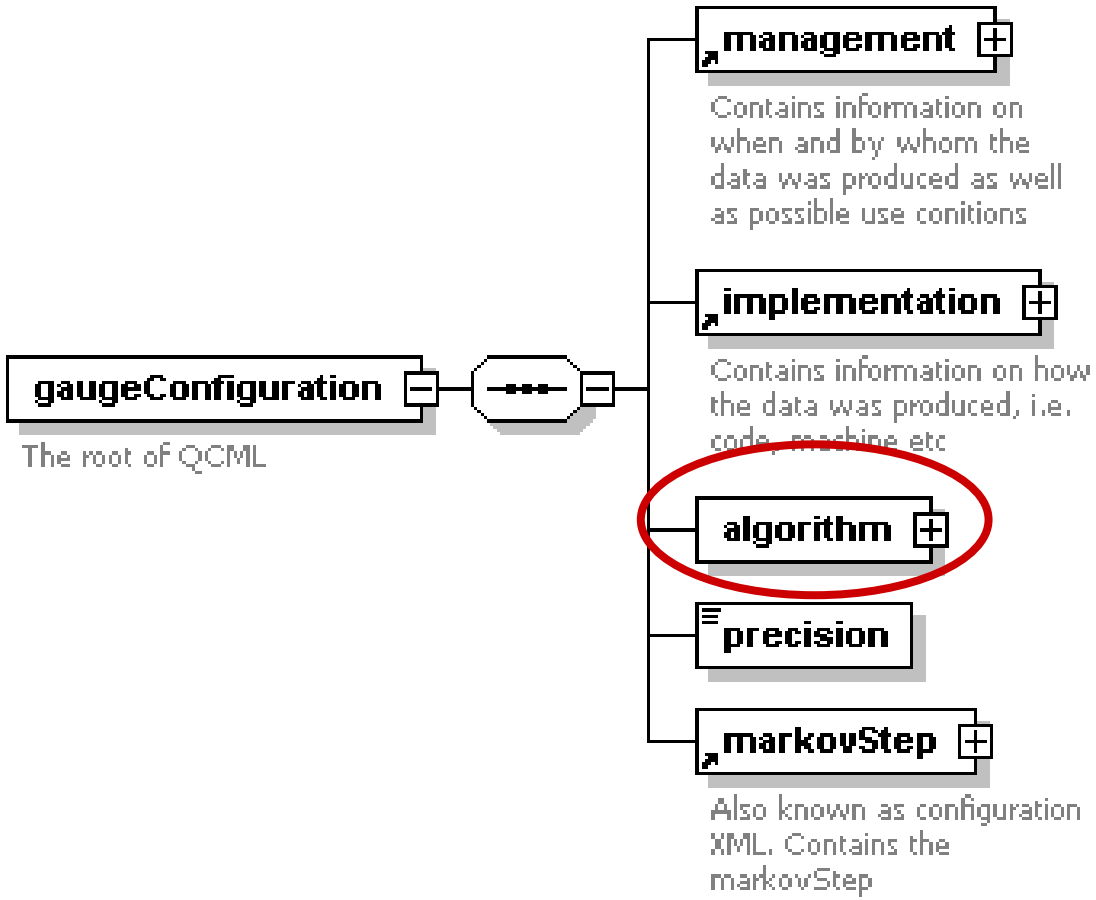
Implementation



Machine and code details

In principle these could be different for configurations in the same ensemble

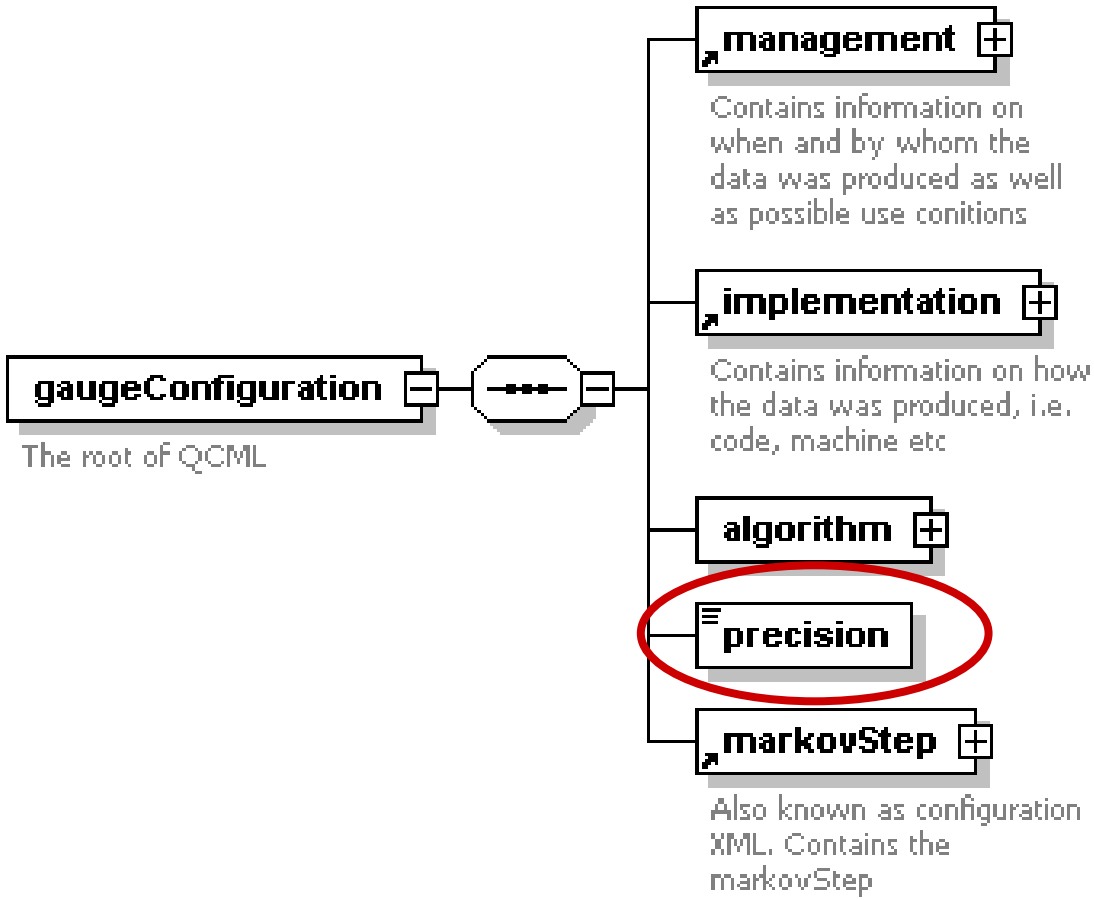
Algorithm



Algorithmic metadata specific to an individual configuration

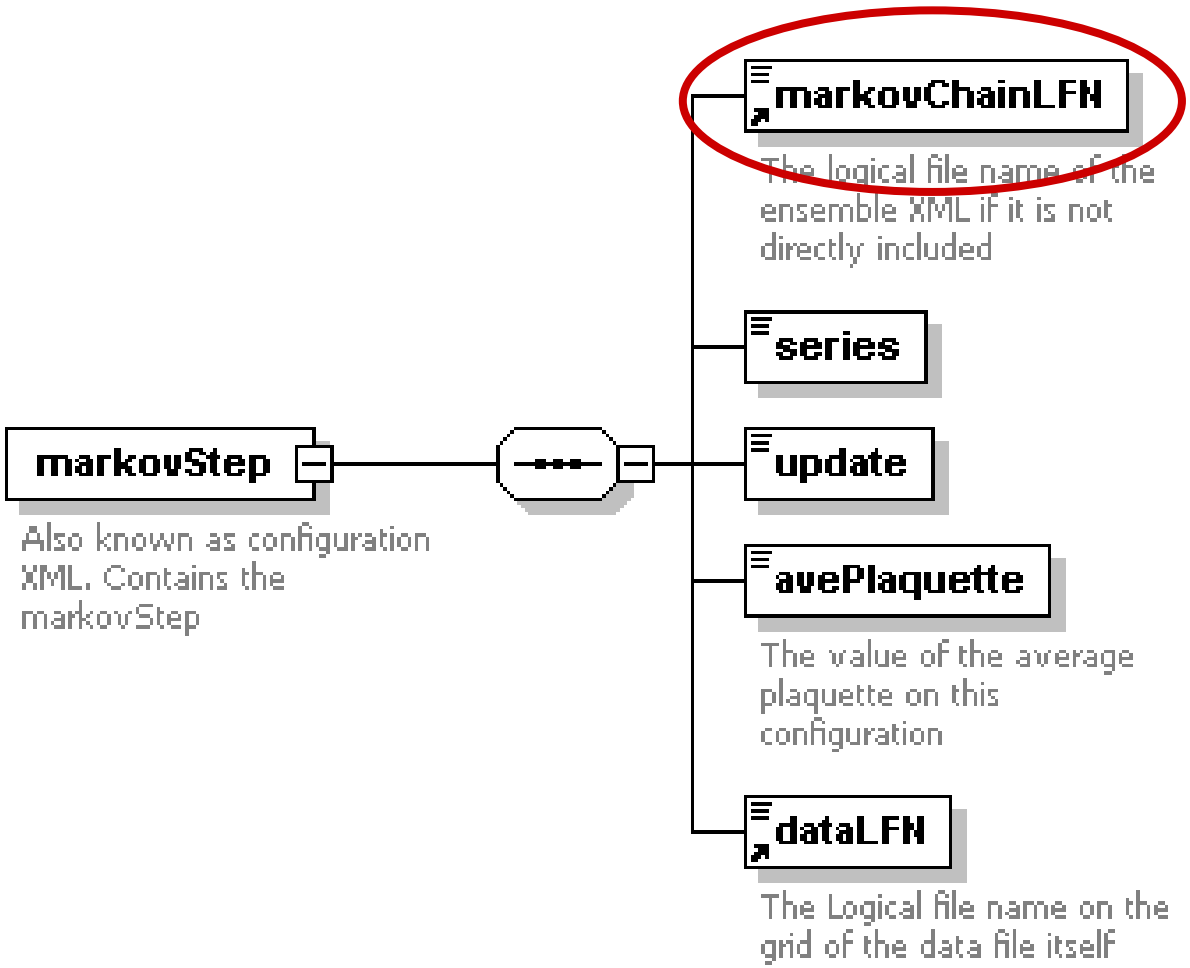
For instance, step size or solver residue

Precision



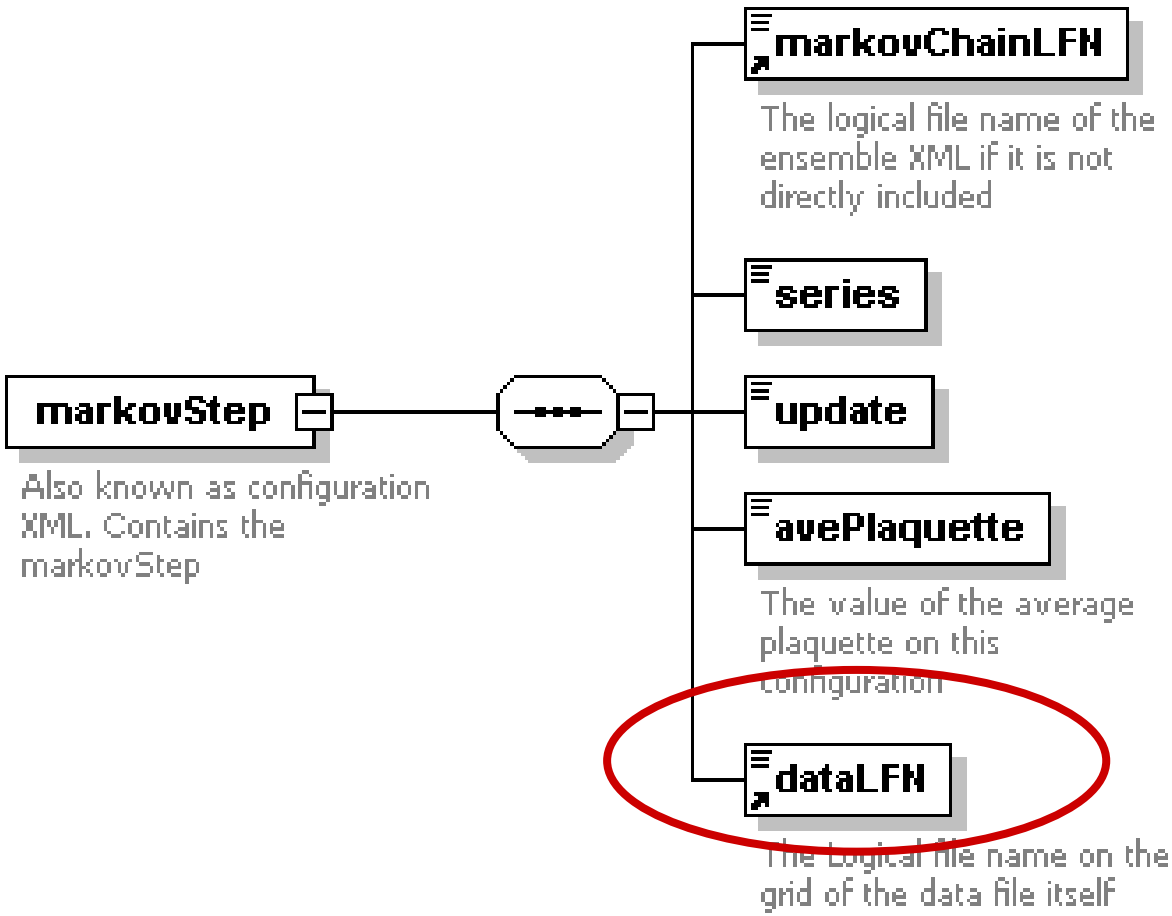
Debate as to whether an ensemble with configurations generated with different precision is valid

markovStep



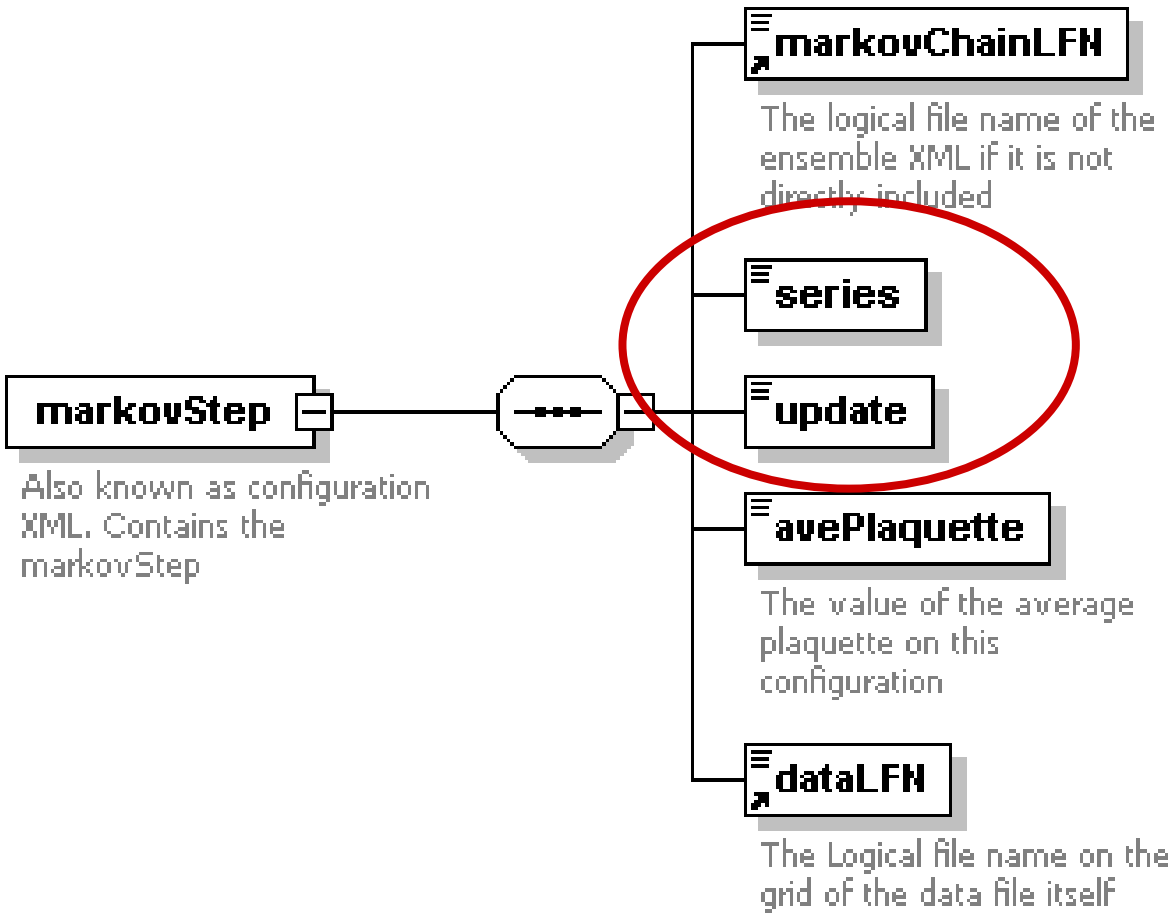
Logical File name of the ensemble in the ILDG namespace

dataLFN



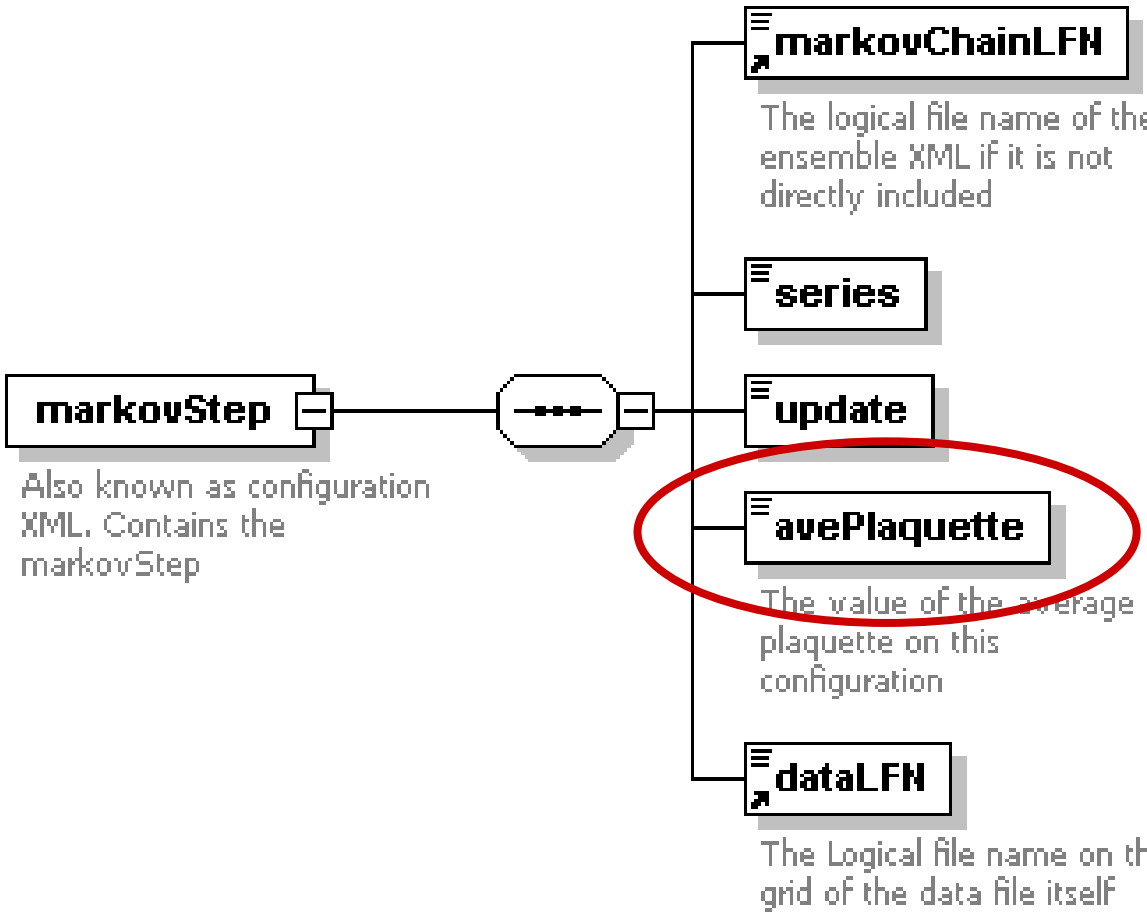
Logical File name of the configuration in the ILDG namespace

The markov chain



Where the configuration is in the trajectory of markov chain

avePlaquette



Very useful metadata, can be used to check data transformations are correct

QCDML1.0

- Schema marked up as version 0.4
  - Requires some tidying
- Remaining issues
  - Can a configuration for which a paper has not been published be part of ILDG?
- Remaining work
  - Inheritance trees for actions
- Move to QCDML1.0 and release

Extending QCDML

- Data format and packing of configs
  - See Yoshie talk
- Gauge fixed configurations
  - Should be fairly straightforward
- Propagators/correlators
  - Will need more work but basis laid in gauge configs

BinX

- XML markup for binary data
- Library for manipulating marked up data
- Production codes do not use BinX library
  - But easy to mark up data format in BinX style
  - ILDG middleware can use BinX for data manipulations
  - Gauge configuration format
  - correlators

Gauge config BinX

```
<!-- this is to show its a XML -->
<binx>
  <!-- this marks its a BinX doc -->
 - <definitions>
   - <defineType typeName="complexFloat">
     - <struct>
         <float-32 varName="Real"/>
         <float-32 varName="Imaginary"/>
       </struct>
     </defineType>
  </definitions>
 - <dataset src="D52C202K3580U025780T01" byteOrder="bigEndian">
   - <arrayFixed varName="gaugeConfigTimeslice">
       <useType typeName="complexFloat"/>
     - <dim name="z" indexTo="15">
       - <dim name="y" indexTo="15">
         - <dim name="x" indexTo="15">
           - <dim name="mu" indexTo="3">
             - <dim name="column" indexTo="2">
                 <dim name="row" indexTo="1"/>
               </dim>
             </dim>
           </dim>
         </dim>
       </dim>
     </arrayFixed>
  </dataset>
```

Small

Written once per ensemble

write code on top of BinX library

Change array order

2x3 → 3x3

average plaquette

ILDG BinX based gauge config manipulator?

Correlator data

```
- <binx>
  - <dataset src="D52C202K3500U010010_R10A200L3500X_L1300X_CMesonT00T31" byteOrder="bigEndian">
    - <arrayFixed varName="correlator">
        <float-32/>
      - <dim name="t" indexTo="31">
        - <dim name="channel" indexTo="35">
            <dim name="momentum" indexTo="10"> </dim>
          </dim>
        </dim>
      </arrayFixed>
    </dataset>
  </binx>
```

Compact. No standard shape to correlators

BinX will read in *any* shape

Array stripper

```
- <data>
  - <dim>
      <name>channel</name>
      <start>0</start>
      <finish>0</finish>
    </dim>
  - <dim>
      <name>t</name>
      <start>0</start>
      <finish>31</finish>
    </dim>
  - <dim>
      <name>momentum</name>
      <start>0</start>
      <finish>0</finish>
    </dim>
</data>
```

BinX + BJ's Xpath reader

Code reads this XML

Produces single slice array in text/XML

From *any* size/shape array

Schema for correlator channels

ILDG middleware extract channel from any correlator

Conclusions

- QCDML0.4 finished
  - Go to QCDML1.0
  - Start using
- Extend QCDML to other data
- CMM recommend BinX as an extremely useful tool
- Easy to create ILDG data manipulation based on BinX + schema