Metadata working group report

CMM’s ideas for where we go next

QCDml status

- Ensemble **`1.4.4`**
  - No change
- Config **`1.3.0`**
  - No change
- No updates required by community
- This is a *Good Thing*!
  - QCDml is stable and does it's job!
- What have the MDWG been doing for the past six months?
  - Resting

Propagator markup

- MDWG has had discussions on propagator format and description
- No appetite in the community for standard data format nor description
- USQCD has four formats to suit their needs
  - how many would everyone need?
  - Not everyone convinced there is sufficient storage
    - CMM doesn't think this is an issue
- Propagator metadata would be very difficult as everyone has a different favourite source
  - Possible interest in eigen values and vector and multigrid restriction/prolongation vectors

Review of QCDml

- Why do we need metadata?
- Extreme example: no metadata
  - Cfgs have random string names with no directory structure for different ensembles
  - Impossible to use
- Organise files
  - Into directories for ensembles
    - Give cfgs names with markov chain position
- Construct a scheme for the metadata
  - Rules for describing the data
  - Chose to construct scheme in XML

Why use XML?

- Semantic, eXtensible Markup language
- *XML was designed to carry data, not to display data*
  - *Cf.* with HTML, designed for displaying data.
  - Incompatible applications can exchange data wrapped in xml
- XML is just plain text
- User defined tags allow structure to be developed
  - Lattice QCD metadata is structured
- *XML does not DO anything*
  - You need an application for this
- XML schema
  - Defines a set of rules for the XML document
  - Applications can know types, parse and processes XML data
    - Could just be an XSLT style sheet to transform XML in HTML and render a web page e.g. LDG web-client

Problems with XML

- Lattice QCD (meta)data is really mathematics
- XML is not really ideal for storing this data
- For ensembles of gauge configurations can define common names for **<action/>** etc
  - Even **WilsonAction** has more than one common usage
    - Kappa versus mass
- Algorithm metadata is too complex for common names
  - Not really defined in the metadata
  - Unstructured parameter values included
- This is OK because an ensemble is defined by the action not the algorithm used to generate it
- Extending to propagators and correlators is hard for the same reason as defining the algorithm

We need an application

- *XML does not DO anything*
- For it to be useful we need to do something with it!
  - What do we want to do with it?
    - Is QCDml good for this purpose
  - QCDml design focused on searching the metadata catalogue
    - This was probably a good idea!
- **Xpath** used to query XML databases
  - Basic tools/APIs exist for constructing queries
    - *Cf*. UKQCD DiGS GUI browser, LDG web-client and JLDG *faceted navigation* application
- Metadata capture
  - How do we create XML IDs?
    - Does any application actually write QCDml?
    - UKQCD does post-processing
- Data provenance
  - Does QCDml provide this?

What next?

- QCDml seems to work OK
  - How much is it being used?
- We don't have many applications that DO something with it
- CMM's Questions for MDWG
  - What do we want to do with metadata?
  - Do we have the right sort of metadata for this?
  - What tools or applications do we need?
    - Someone then has to build them
    - if we don't ask, we don't get!

- Can we review QCDml usage to define what tools we need?