Tools for ILDG

Dr Chris Maynard
Application Consultant, EPCC
c.maynard@ed.ac.uk
+44 131 650 5077

WIKIPEDIA

A tool is a device that can be used to produce an item or achieve a task, but that is not consumed in the process

Wrong sort of tool can produce poor results, or not scale to larger problems

How do we access our data?

– In the same way we did a decade ago

– ssl terminal client (ssh) and copy protocol (scp)

• Data explosion

– Data volumes – Tbytes, Pbytes soon

– Data complexity

– many ensemble, many measurements

– Rise of the mega collaboration

– Globally distributed {machines, data, people}

We really need some tools!

Tools

- **Globus online (Monday)**
  - *Reliable Data Movement via SaaS* Raj Kettimuthu

- **Web2py (Poster)**
  - Poster *A new user interface for the Gauge Connection lattice data archive,* M. Di Pierro, J. Hetrick, D. Skinner, and S. Cholia
  - plus demo after this talk

- **LATFOR grid tools,** Dirk Pleiter *et al.* `ildg-get`, web client

- **UKQCD Ildg-browser**

- **JLQCD faceted web client**

  http://ildg.sasr.edu.au/Plone/ildg/ildg-clients

- **Metadata capture project**
  - EPCC and Tsukuba University
  - T. Amagasa, M.G. Beckett, C.M. Maynard, J. Perry, T. Yoshie

LATFOR tools

- ildg-get can access data, metadata, and ILDG services
  - need to know LFN, or markovChainURI of the metadata

- Metadata webclient

- http://www-zeuthen.desy.de/latfor/ldg/doc/swinstall.html

JLDG

- Faceted browsing

- http://www.jldg.org/facetnavi/

UKQCD ILDG-browser

- ## MDC GUI client
  - Self-contained Java application, runs on Windows/Mac/Linux.

- ## Allows users to:
  - GUI to construct queries to MDC
  - Search Metadata
  - Store queries
  - Retrieve metadata

- ## Does not have data access
  - use browser to find the Logical File Name (LFN)
  - Get data with ildg-get

UKQCD ILDG-browser demo
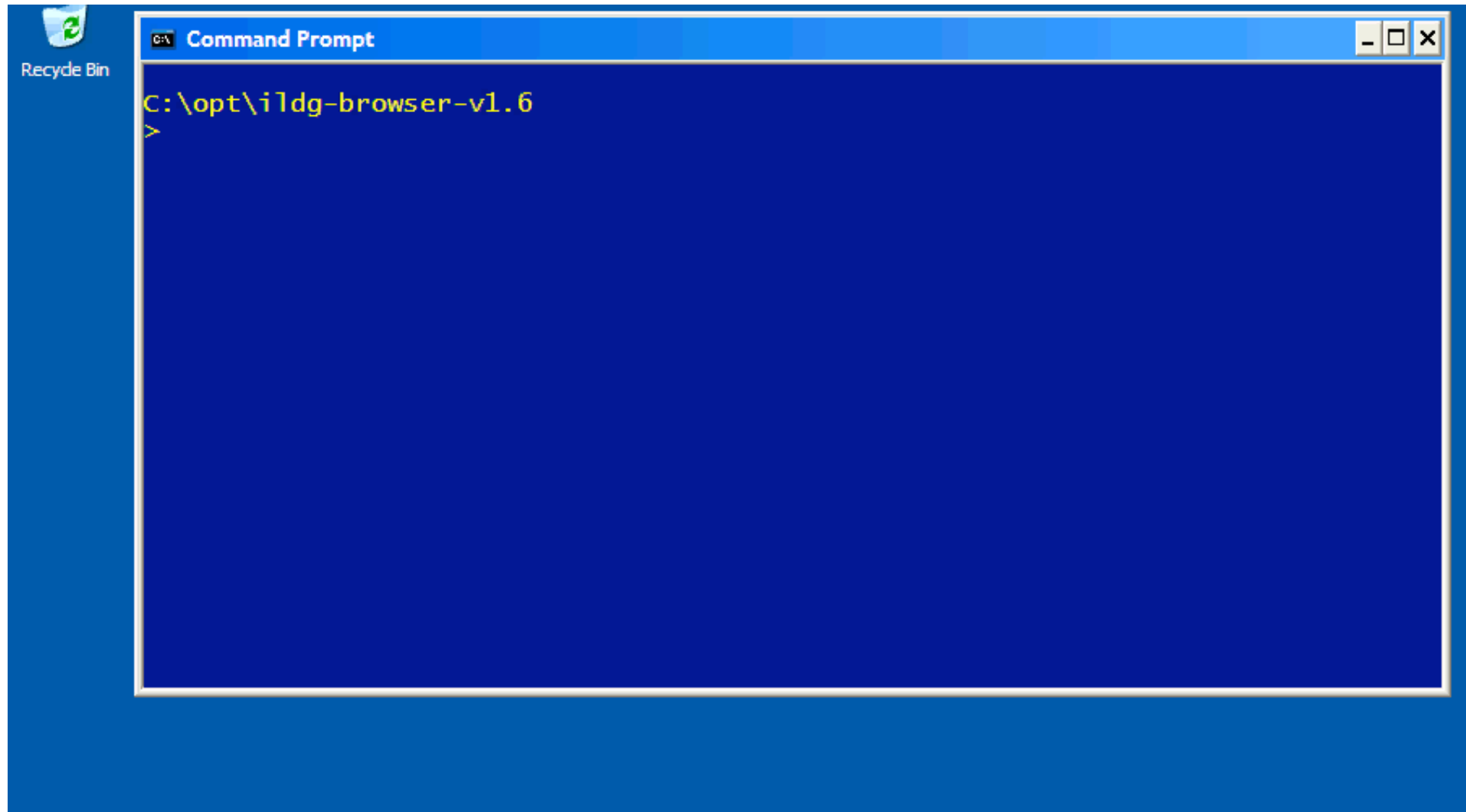
Metadata capture

- Tools thus described are for accessing ILDG services
    - they exist and are useful

- No tools for metadata capture
    - Ensuring data provenance is difficult
    - are there degrees of provenance?

- QCD production codes are highly optimised
    - run on highly diverse (and bespoke) architectures

- Require lightweight process to ease pain of post-processing data

Hard Work

ETMDC

- Edinburgh - Tsukuba Metadata capture project
  - T. Amagasa, M.G. Beckett, C.M. Maynard, J. Perry, T. Yoshie

- Explore workflow as a mechanism for MDC

- Edinburgh funded by
  - OMII-UK
  - Software Sustainability Institute
  - Edinburgh Global (UoE)

- End product
  - Demonstrator - universal metadata capture tool for ILDG
    - Linux/Unix environment
    - Python, XSLT, make
    - QCD utils
    - some hints from QCD code gen

MDC design criteria

- **Considered workflow tools**
  - Metadata generated and manipulated as part of data generation process
  - Examples: Kepler, Taverna, Ruby
  - QCD ConfGen Jim Simone's FNAL group

- **Complex tools with rich functionality**
  - Will they run in bespoke QCD environment

- **Lightweight is key criterion**
  - opted for simplest solution
  - build demonstrator out of most commonly available components
  - Used make to manage dependencies, but could upgrade to Kepler

- **Used two example codes**
  - JLQCD, CPS

Metadata

- ALL QCD codes output meaningful metadata
  - plus input parameter files
  - system size, physical parameters, quark, gluon couplings
  - algorithmic parameters, step size
  - measured quantities, plaquette, checksums etc
  - state information, user, code version, machine information
  - Gauge configuration file

- No scheme for organising this information
  - parse and process this information

- Add some minimal mark-up to information already produced
  - some hints for the tool

Hints

- Add simple markup to output
  - easy for user to implement – its just plain text
  - gives tool something to work with

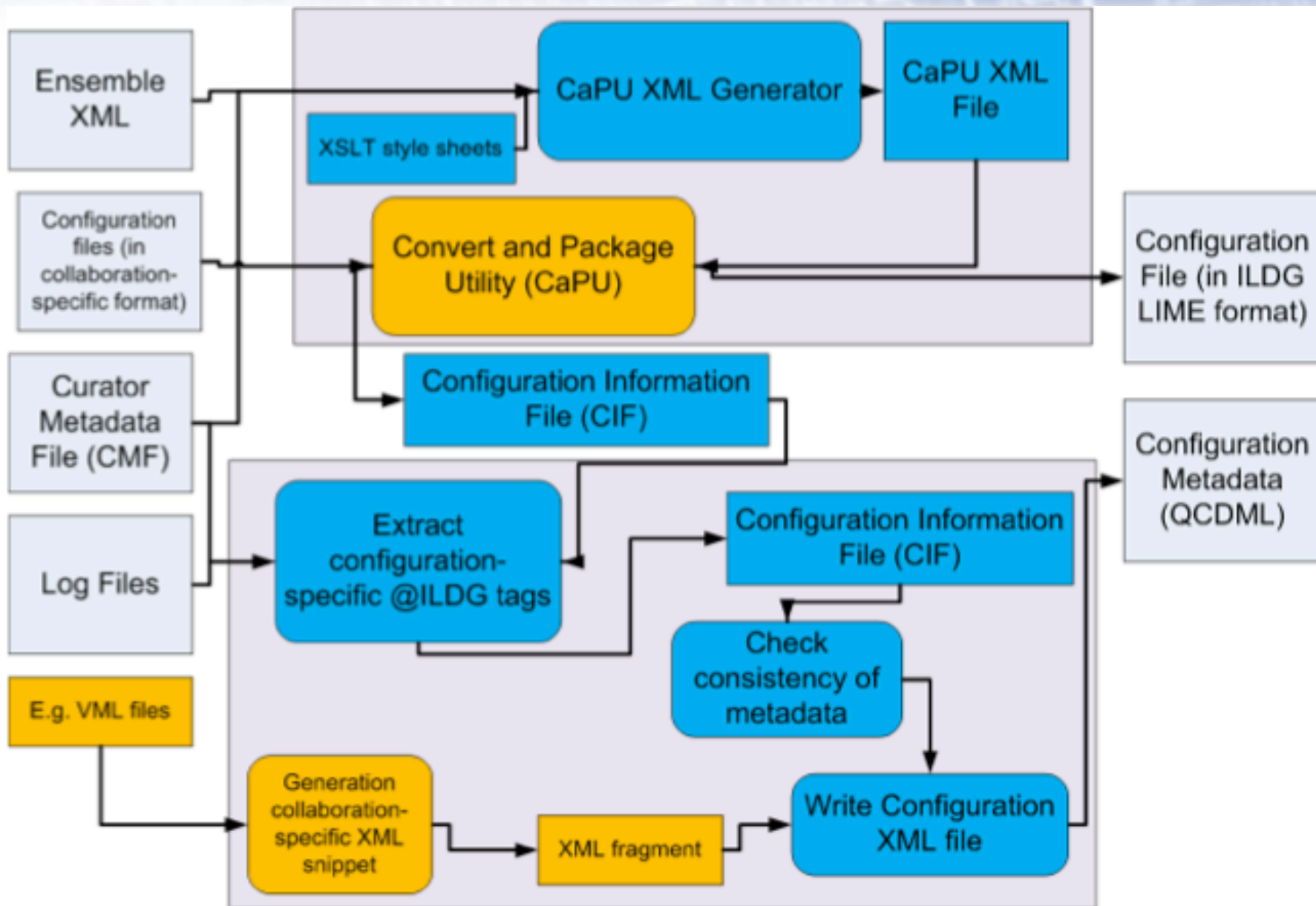- simple `@ILDG` tag for interesting information in plain text files

- Examples:

  `@ILDG:codeVersion "v4.0"`

  `@ILDG:checksum 475303070`

-

User input

- QCDml Ensemble ID [XML]

  – written by human once per ensemble

- gauge configuration files

- log files with hints

- Curator metadata file (CMF)

  – where are the data, log files etc

- MDC demonstrator will do the rest!

  – Two main components
  – Configuration File generator
  – Configuration XML generator

MDC architecture

Example CMF

```xml
<CMF>
<Ensemble>
  <EnsembleIDFileName>ensemble1.xml</EnsembleIDFileName>
</Ensemble>
<Configuration>
  <ConfigurationUpdateStart>1000</ConfigurationUpdateStart>
  <ConfigurationUpdateStep>10</ConfigurationUpdateStep>
  <ConfigurationUpdateEnd>1230</ConfigurationUpdateEnd>
  <ConfigurationFileName>config.%04</ConfigurationFileName>
  <ConfigurationILDGFileName>configILDG.%04</ConfigurationILDGFileName>
  <ConfigurationPrecisionILDG>64</ConfigurationPrecisionILDG>
</Configuration>
</CMF>
```

specify batch processing of configurations `@ILDG:UpdateStart` and `@ILDG:UpdateEnd` to delimit information in log file

format string-style pattern to specify file name
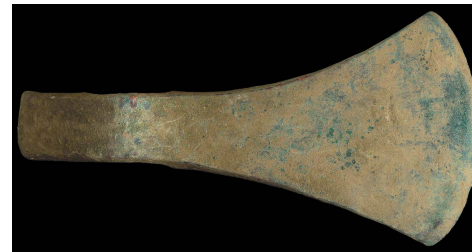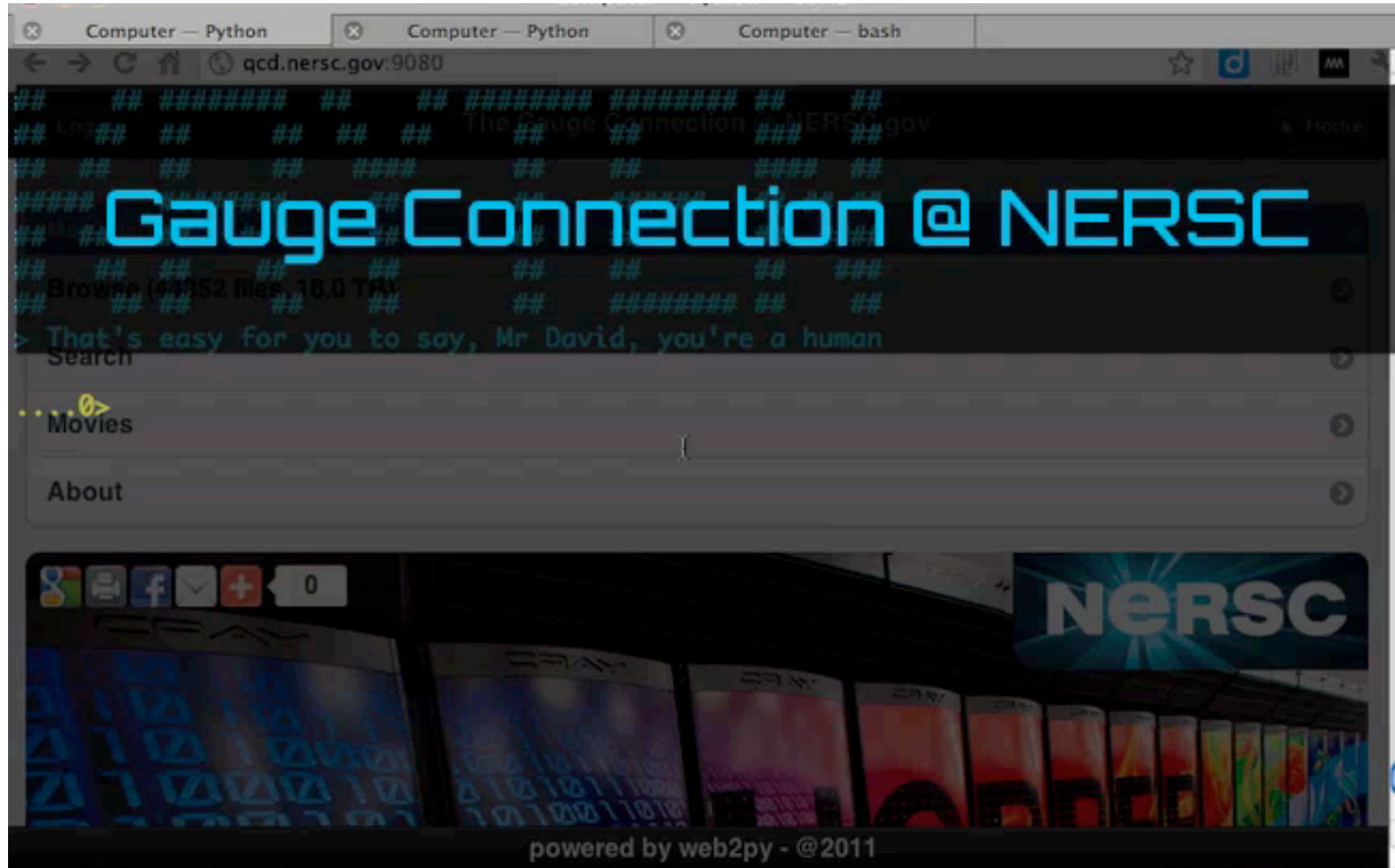
Configuration File Generator

- ## Two components
  - XSLT transform creates CaPU XML from
  - Ensemble XML ID
  - CMF

- ## Conversion and Packing Utility (CaPU)
  - specific to collaboration, but has common interface
  - converts data to ILDG format
  - measures plaquette, CRC checksum etc
  - writes Configuration Information File (CIF) (above + LFN)

- ## UKQCD based on qdp++ utility
  - if qdp++ can read your data, easy to modify the CaPU

- ## JLQCD is shell script + data conversion

Configuration XML Generator

- Creates the QCDml config ID

- Several components - Python

- Extract configuration specific information
  - from CMF, CIF and log files

- Consistency and completeness checker
  calculated plaquette = logfile plaquette
  - Do I have all the information I need?
  - Do the sources of metadata agree?
  - am I processing the data I think I am? Provenance

- Include collaboration specific information
  - e.g. VML from CPS

- Write the XML

Summary

- ## MDC Demonstrator
  - Using common linux/unix tools/software to build components
  - Can automatically post-process data into QCDml

- ## Others can use or adapt demonstrator
  - simple modifications to output of QCD code
  - simple modifications to CaPU

- ## Can be downloaded from
  ILDG web site

- ILDG – we need tools

- There are tools out there
  - useful!

- More groups are developing tools

- If you need help get in touch

- Share experiences

- Neolithic → bronze age
  - cross over or 1st order transition?

NERSC gauge connection

- http://tests.web2py.com/ildg/default/index