XMLHEP: proposal for a structure of partonic event files

Alexandre Sherstnev

on behalf of the CompHEP collaboraton

- Why do we need the event files?
- XMLHEP - standard for events files: general conception
- Simple example: XML and HTML representations of an event file header
- XMLHEP interface libraries
- XMLHEPpyth-1.0: interface to PYTHIA 6.2
- Monte-Carlo DataBase - common place for event file storage. CMS MCDB and LCG MCDB.
- Conclusions

Many, many MC generators...

- Modern complexity of MC generation has resulted in separation of 2 stages of the events generation:
  - Matrix Elements - ME (AlpGen, CompHEP, MadGraph, etc.)
  - Showering and Hadronization - SH (HERWIG, PYTHIA, Sherpa, etc.)
- We have to organize an interface between generators of these stages (Solution: Les Houches Agreement I).
- Two possible solutions for the data transfer:
  - Events are transfered 'on-fly' (Les Houches Agreement I).
  - Events are stored in events files.
- Many experiments wants to store events in files on both generation stages: partonic and particle level.

Data storage standard

Standardization is necessary for the events files.

- Initial requirements:
  - Platform independence;
  - An uniform syntax for any types of information stored in the event files.
  - The syntax should be extensible (allow to add some new information to files with minimal changes in files).
  - Simple I/O and parsing by already existed and well-maintained software tools.
  - Possibility to browse the files via the Web (at least, some general information about processes).

XMLHEP - standard for event files

General conception:

- All data in an event file are divided to 2 parts:
  - Header: describes general information about the event flow:
    - Information about author, the file itself (date and place of creation, etc.), supported specifications, collider description;
    - General description of subprocess stored in the file: particles description, cuts, physical parameters, generator specific information (generation conditions, etc.);
    - Data for event records parser: formats of event records for each subprocess;
  - Events records: variable data of events written in some compact format to one string (particle momenta, color chains, etc.).
- Header is stored in a text file with the XML syntax.
- Event records are compressed (by the zip algorithm) and attached to the header file.

Why XML?

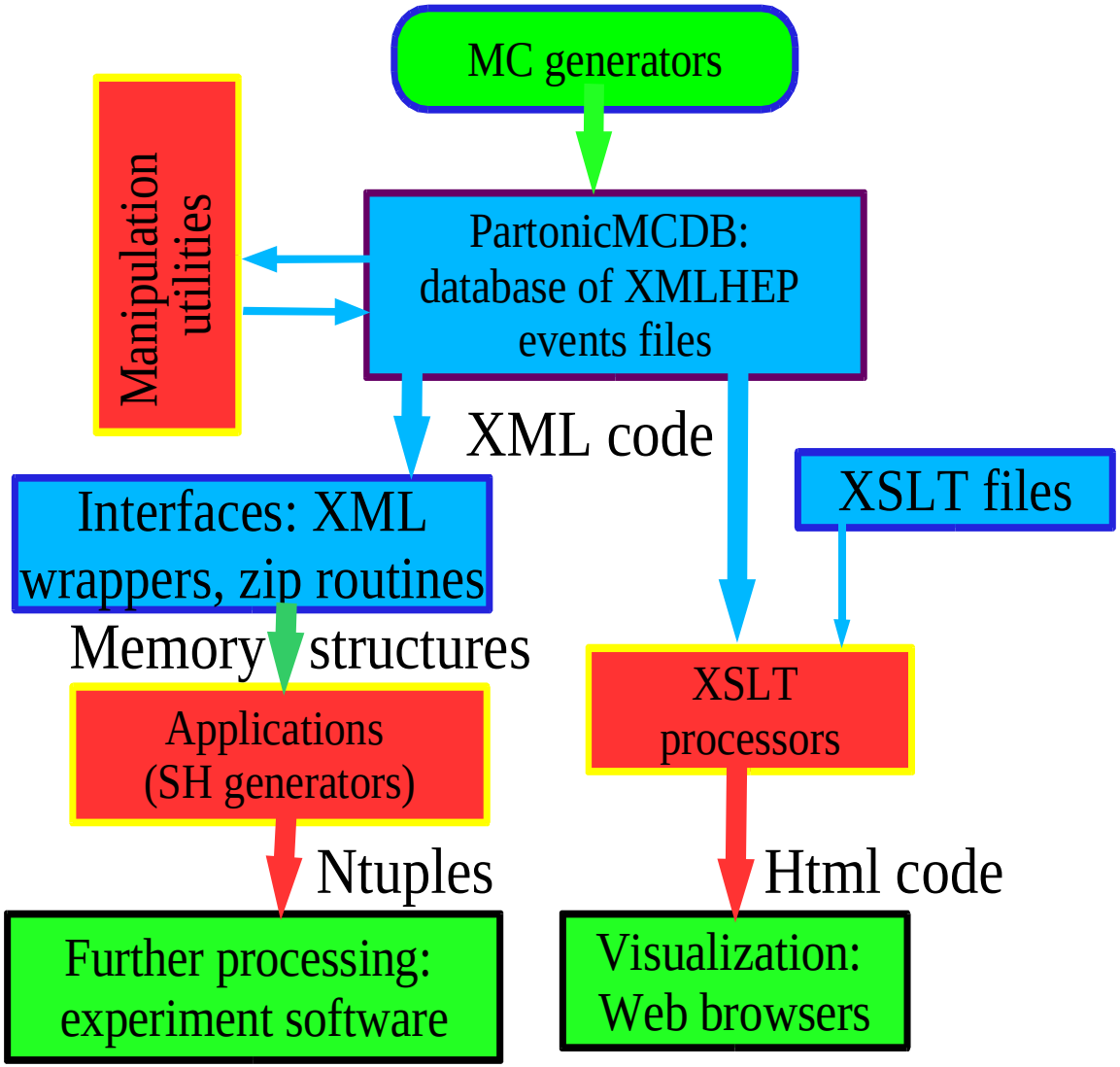
- XML is a sufficient: tree structures allow to store all necessary information in event files.
- XML is a eXtansible: it defines just rules for the document syntax and allows to develop problem-specific sets of elements. In particular, one can develop an elements set specially for High Energy Physics (Namespace: hepns).
- XML is a widespread: there are many reliable and well-maintained packages for XML documents parsing for different programming languages (C, C++, FORTRAN with C-wrappers).
- XML is a legible: XML syntax can be read by users (by eyes).
- XML is an Internet oriented: good support in many Web browsers; standard tools for visualization via Web browsers (by XSLT/HTML techology) .

XMLHEP: a Simple Example

```
<Machine>  LHC
  <Beam id="1" energy="7.00000E+03">
    <Particle name="proton" KF="2212" mass="9.38000E-01"/>
    <Strfun name='CTEQ' version='5l'> <PDFLIBinfo id="46" gr="4"/>
  </Beam>
</Machine>
<Task>
  <Process ID="1">     u, D -> n, e, E, b, B
    <GeneralInfo>
      <CrosSection value="1.11227E-03" error="1.90987E-03" Nevent="100"/>
      <Generator name="CompHEP" version="4.2.0"/>
    </GeneralInfo>
</Task>
```

XMLHEP standard presentations

The XML/XSLT thechnology allows to transform XML files (which are suatable to data transfer) to HTML code (for vizualization)



HTML Example (1)

- After applying of a XSLT transformations to orginal XML document one can obtain HTML files, suitable for presentation in Web browsers.

File Information:
Author: sherstnv (sherstnv@theory.sinp.msu.ru)
Path:
thep14.sinp.msu.ru:/scratch/sherstnv/development/cpyth_project/cpyth-2.1.0/tag-library
Date: 02-Oct-03

Specifications

1. XMLHEP 1.0
2. XML Namespace: hepns 1.0
3. Les Houches Agreement 1.0 (MAXPUP=100 MAXNUP=500)

Machine: The LHC collider

Description

Subprocesses

Subprocess: id: 1 (details)
Name: u, D -> n, e, E, b, B
Cross section: 1.11227E-03 ± 1.90987E-03 % (50.000%)
Number of Events: 10000
Subprocess: id: 2 (details)
Name: D, u -> n, e, E, b, B
Cross section: 1.11227E-03 ± 1.90987E-03% (50.000%)
Number of Events: 10000

Results

Number of Subprocesses	Total Cross Section (pbn)	Total Cross Section Error (%)	Number of Events
2	2.22454E-03	2.70987E-03	20000

HTML example (2)

Description of a subprocess

Process (id: 1)

Name: u, D -> n, e, E, b, B
Generator: CompHEP, Version: 4.2.3
File Information:
Author: sherstnv(sherstnv@theory.sinp.msu.ru)
Path:
thep14.sinp.msu.ru:/scratch/sherstnv/development/cpyth_project/cpyth-2.1.0/tag-library
Date: 02-Oct-03

Cross Section (pbn)	Cross Section Error (%)	Number of Events
1.11227E-03	1.90987E-03	10000

Cuts:
Cut Name Scheme: CompHEP, Version: 4.2

Variable	Variable Name	Minimum Value	Maximum Value	Status
Transverse Momentum of Particle id=3	T3	1.00000E+01		regular
Transverse Momentum of Particle id=4	T4	1.00000E+01		regular
Rapidity of Particle id=3	Y3	-4.00000E+00	4.00000E+00	regular
Rapidity of Particle id=4	Y4	-4.00000E+00	4.00000E+00	regular

---

Process particles:
Number of Particles: 6
Particle Name Scheme: CompHEP, Version: 4.2

Name	KF	ID	Mass	Status	Mother 1	Mother 2
u	2	1	0.00000E+00	in	0	
D	-1	2	0.00000E+00	in	0	
ne	12	3	0.00000E+00	out	1	2
E	-11	4	0.00000E+00	out	1	2

HTML example (3)

Description of a subprocess

b	6	5	4.85000E+00	out	1	2
B	-6	6	4.85000E+00	out	1	2

Generator specific information:
Generator: CompHEP

1. MC Generation Parameteres: RandomFinal: 55A9508BA2EE
2. Physical Parameters:
   EE=3.134500E-01; SW=4.807600E-01; s12=2.229000E-01;
   s23=4.120000E-02;s13=3.600000E-03; MZ=9.118760E+01;
   wW=2.027980E+00;Mtop=1.743000E+02; wtop=1.54688E+00;
   Mb=4.85000E+00
3. Kinematical Scheme:
   12 -> 6 , 345
   345 ->5 , 34
   34 ->3 , 4
4. Phase Space Regularizations:

Momentum	Mass	Width	Power
34	MW	wW	2
345	mtop	wtop	2

5. QCD Information: NL: 1
   Nflavour: 5
   QCD Lambda (GeV): 1.65200E-01
6. PDF Factorization Scale: 9.118760E+01

Interfaces for XMLHEP files

- Any file format requires parsing to load data for applications.
- For SH generators we plan to develop a specail interface libraries with external XML parsers (for FORTRAN/C/C++) and zip routines:
  - expat library: C/C++ parsers for XML code (for FROTRAN one needs to develop special wrappers for C functions).
  - A special parser for events records (with using of a libzip to uncompress the data).
- First version of the interface package has to maintain, at least, the standard 'Les Houches Agreement I' structures.

XMLPyth-1.0

- First goal: an interfacing package of XMLHEP files to PYTHIA 6.2 (events flows as an external processes):
- Main points:
  - Main program will be a FORTRAN PYTHIA program;
  - The interface will use expat parsers for the XML code parsing with C-FORTRAN wrappers;
  - All front-end routines will be collected to C and FORTRAN libraries;
  - Transparent structure of routines, associated with every XML element/attribute. The interface will contain a some set of standard XML elements. If an user will want to add new elements/attributes (with a new functionality), he/she can to do it according to a simple standard procedure.

XMLPyth-1.0 in LCG MCDB

The LCG Generators services project is developing a Monte-Carlo DataBase - common place to store events file which are necessary of the LHC collaborations.

- The MCDB software should have an interface to every S/H generator, used in LHC collaborations. XMLPyth will be first package of this type.
  - In addition to interface routines XMLPyth has to contain module for automatic search of requested events files in MCDB.
  - Special programming modules, which will allow to build the interface to the collaborations software environments.

MCDB - common place for event files

- Content of the MCDB is parton level event files. In ideal, these files should be stored under an uniform standard (XMLHEP).
- MCDB should have interfaces of 2 different types:

  - Web interface: a web server with a simple access to the full description of all available event samples in the server and to the samples themselves.

  - Programming interface: users will install the interface on local machines and set just some parameters in the interface. After that, the interface will find automatically the required event files in MCDB and attach SH programs (PYTHIA, etc.).

- Web part of MCDB has to be divided onr 2 zones:

  - Public area: for all users who are interesting in MC events. There they may find all information about the event samples, download event files, and attach any comments/questions about the events.

  - Authors area: authorized users (authors) can change dynamically the content of MCDB – create/edit arcticles, described events, upload event files, and reply to users questions/comments.

CMS MCDB (main developer: Lev Dudko)

- The CMS MCDB location is a cmsdoc.cern.ch server:
  http://cmsdoc.cern.ch/cms/generators/mcdb/
- All data are stored in afs (now on three 2Gb disks).
- The MCDB is based on Berkley DBM with a Perl interface: Administration toolkit of CMS MCDM is based on Sanitarium WebLoG (free soft) - system of Web publishing (perl scripts).
- Authors of event files have a write access to the disks via special Web interface. The access is given by a MCDB administrator. The authors use afs passwords to come to Authors Area (SSL).
- Currently the CMS MCDB has 4 administrators: A. De Roeck, L. Dudko, A. Sherstnev, and S. Slabospitsky.
- Now MCDB contains events generated by CompHEP only. We have a program to transfer the events to PYTHIA. The interface sources also are available in MCDB.

LCG MCDB

For LCG MCDB we plan:

- MCDB will be divided on 2 parts: PartonicMCDB (for partonic events) and ParticleMCDB (for events after SH). PartonicMCDB will use XMLHEP standard (?), ParticleMCDB - ntuples/rtuples with the HepMC interface package (?). In addition to event files, MCDB will contain HTML pages with info about event files, created by XSLT trasformations from XML files.
- Simple interfaces for users: an effective search of event files, Web access to all available information about event files, simple access to the files themselves via the Web and programming interfaces.
- The effective toolkit for authors: the toolkit should allow to work easy with articles; also the author have to fill some necessary minimum of information about the uploaded event file.
- Simple administration of a future MCDB server with possibility of a mirroring of the site.
- In future, we plan to migrate from Berkley DBM to a relational DB.

Conclusions

1. XMLHEP is a standard good for partonic events storage in files.
2. XMLHEP standard gives advantages over the 'fixed formats':
   - necessary format flexibility;
   - well-supported parsings tools;
   - XSLT technology for different transformations (for example, to HTML for visualization)
3. XMLHEP can be discussed as candidate for of file storage format in the LCG PartonicMCDB.
4. One needs to develop programming interfaces to SH generators (1 per the generator), in case of XMLHEP too.
5. All groups (especially, MC generators authors/experts groups) interested in the standard are invited to discuss XMLHEP in frames of the LCG Generator Servces project.