# An OO tag database for physics analysis at HERA

*A.Fox-Murphy*[1]

Deutsches Elektronen-Synchrotron, Hamburg, Germany

**Abstract**

A persistent object database based on Objectivity/DB was used for the first time in a running high-energy physics environment in the ZEUS experiment at the HERA electron-proton collider. An event tag database of 250 physical variables per event is created during offline event reconstruction and is later used to rapidly select events for physics analysis. Around 100 million events from 5 years of data are currently stored. The environment, the design choices and the tag database system are described, and the favourable users' experience is reported.

Keywords: Objectivity,tag database,offline,analysis

## 1 Introduction

A perennial problem in high-energy physics is how to access the events required for a physics analysis in an efficient way. Generally for a particular analysis it is only necessary to fully analyse a sparse subset of the full event set. A common solution to this problem is to create a tag database which contains summary information about each event which can be used to preselect those events which should be analysed in detail. In this way the CPU-intensive parts of the analysis are performed efficiently.

The ZEUS experiment at the HERA ep collider in Hamburg, Germany, has been operational since 1992. In 1996 a project was started to upgrade the existing offline event selection system to use an object-oriented tag database based on Objectivity/DB. This database, the ZEUS Event Store (ZES) has been in use since 1997. In this paper the design of the system and the experience obtained by physicists using it are reported.

## 2 ZEUS Offline environment

The current raw data rate passing the third level trigger at ZEUS is about 10 Terabyte per year, or 1 Megabyte per second. Events pass to offline storage at a rate of around 10 Hz, each event having a size of 100kB. Offline reconstruction of these events is performed by a farm of 17 Linux PCs, following which the reconstructed events are available for physics analysis. Both the raw and reconstructed data are stored in an ADAMO database and accessed from user analysis jobs written in FORTRAN. Currently the ZEUS data set accumulated since 1995 totals around 100 million events. Batch analysis of reconstructed events is done using the ZEUS central batch processing facility which is based on 3 SGI Challenge XL machines. In the near future this will be migrated to a farm of 30 Linux PCs. See [1] in this conference for full details.

## 3 The ZEUS Event Store

### 3.1 Context

Typically a physics analysis must process only a small subset of the full data set. Before the introduction of the ZES database preselection of the events was done using 128 precalculated boolean quantities (DST bits). These are generally simple logical combinations of predetermined loose physical cuts. Event selection for a user job was then a sequential test of the DST bits set for each event. The intention was to design a system which would improve access times to the reconstructed event data, and to provide a more flexible and intuitive user interface.

### 3.2 Objectivity database

ZEUS decided to use a commercial object-oriented database product, Objectivity/DB. This was believed to be convenient and flexible for our requirements, and to be sufficiently scalable for all future storage requirements. It has been extensively investigated in the HEP community. It is notable that the ZES database is thus separate from the ADAMO event database, and of a different type and using a different programming language (i.e. Objectivity/C++ rather than ADAMO/FORTRAN). Thus the advantages expected from the OO approach were offset against the investment required to integrate it with the existing software.

The choice of variables stored in ZES was made collectively by the ZEUS physics groups, and includes information from a wide range of detector components, as well as from the trigger and from the offline reconstruction, such as event kinematics, tracks and jets.

The set of tag variables is calculated just once (during reconstruction) from the full event data and then is used thereafter to select the events by means of a software selection filter which examines the tags.

The database schema is as follows. ZES consists of a federated database in which each database within the federation holds a number of containers, or *Run* objects, (one per HERA run), containing many (order 10,000) *Event* objects per run. Each *Event* object currently holds about 250 floating point and integer variables. Objectivity variable-length arrays (`ooVarray`) are used for flexibility: at any time the number of stored tags can be increased if required.

The file structure of the ADAMO database is reflected in ZES: in addition to the tags each *Event* object contain a reference to a *File* object which contains the details of the location of the event in the ADAMO files containing the full event data. In this way, for an event passing the tag cuts, the full data can be fetched and transparently passed to the user's analysis job.

Each database corresponds to a file of 200-450MB, and several tens of databases are required to hold each year's data. Currently data from 1995 onwards is stored in ZES, totalling around 100 million events, and occupying around 100 GB on disk.

The structure is largely transparent to the ZES user who simply uses control cards to select the run range for analysis and to apply the tag selection cuts in a simple intuitive way using the predicate string selection facility of Objectivity.

### 3.3 Integration

Around 18000 lines of C++ code integrate the OO ZES database with the existing ADAMO/FORTRAN environment and provide administration tools. This code includes

- interface to ADAMO event database and analysis environment, including predicate string query formulation
- database loading and other administration routines
- consistency checking/troubleshooting

- user command-line database query tool (mostly for debugging jobs)
- user ntuple generation tool (for directly creating ntuples containing the ZES information)

The user interface is through control cards set in a text file supplied by the user. These cards specify the run range and the event selection using a predicate string. For example a user may specify the following;

```
(
(Ee>5)
and((Zvtx>-50)and(Zvtx<50))
and((Eminpz>35)and(Eminpz<65))
and(Yjb>0.04)
)
```

Clearly this is intuitive for the physicist and allows the event selection to be precisely tailored to the analysis in question.

## 4 User experience

The experience of physicists using ZES has been positive, and it has gained acceptance in a broad spectrum of analyses. The flexible predicate selection mechanism may be used to greatly reduce the numbers of events which are passed to the user job for full analysis, with a corresponding decrease in the time required. Typically reductions of a factor 3-10 are seen, depending on the selection used. Even for null selections an improvement in performance is seen. Particularly sparse event selections show the most striking improvements.

## 5 Conclusion

The ZEUS Event Store was successfully introduced to ZEUS and represents a major improvement to the offline analysis environment. Due to demand the number of tag variables has been increased over the period that ZES has been in operation, from 92 originally to over 250 today. In the near future it is planned to add some important Monte Carlo sets to the ZES database. Further expansion of the scope of the tag variables is also expected.

In the longer term there may be advantages which result from the fact that ZES is an object-oriented database and is written in C++. Development and maintenance of code in C++ may be considerably easier than the equivalent in FORTRAN. Also it is hoped to be able to take advantage of some of the work done in the HEP community on OO analysis and visualisation tools.

## 6 Acknowledgements

Thanks to Tobias Haas, Rainer Mankel, Marek Kowal and Krzysztof Wrona for their help.

## References

1 M.Kowal et al., “A Linux PC Farm for Physics Analysis in the ZEUS Experiment”, CHEP 2000, February 2000.