The CostGlue XML Schema

Francesco Furfari, Francesco Potortì and Dragan Savić

Abstract

An XML schema for scientific metadata is described. It is used for the CostGlue archival program, developed in the framework of the European Union COST Action 285: "Modelling and simulation tools for research in emerging multi-service telecommunications". The schema is freely available under the GNU LGPL license at http://wnet.isti.cnr.it/software/costglue/schema/2007/CostGlue.xsd, or at its official repository, at http://lt.fe.uni-lj.si/costglue/schema/2007/costglue.xsd.

Index Terms

XML schema, metadata, archival

I. The CostGlue program

CostGlue is a storage tool for big to huge quantities of data derived from simulation runs and measurements. It is intended as an aid to simulation and measurement pratictioners to efficiently store, access, filter and exchange data. Through the use of import and export modules written in Python, CostGlue is able to read and write specialised data formats.

As of 2008, CostGlue is still in the prototype phase. It is hosted at http://svn.ltfe.org/costglue, the SVN server (user:cost285, passwd:cost285) of the Laboratory for Telecommunications at the University of Ljubljana (SI), and can be freely accessed in read-only mode.

CostGlue was born within the European Cooperation in the field of Scientific and Technical Research (COST), a framework for scientific and technical cooperation, allowing the coordination of national research on a European level. COST Actions consist of basic and precompetitive research as well as activities of public utility. Specifically, the objectives of COST Action 285 have been the realization or improvement of simulation tools for telecommunications.

The CostGlue architecture is described in [1].

Data stored in Costglue is described via metadata in XML format, organised following the *CostGlue XML Schema*.

II. Metadata and processing data

In recent years increasing attention has been devoted to metadata for every application domain. The XML Schema language [2] provides a means for defining the structure, contents and semantics of an XML document and it is widely used to collect data about data, that is, metadata. The HDF5 [3] data format allows metadata to be associated with every object by using a series of predefined attributes in the form of *name=value* pairs. This mechanism is too simple for our requirements. Consequently, in order to insert metadata, we defined an XML Schema whose document instances can be saved together with the simulation data, so to be part of the logical data structure of the CostGlue HDF5 archive that is depicted in Fig. 1.

Metadata can be associated to every Data Group; metadata referring to the archive as a whole are saved together with the indexing table, while metadata for a single simulation run are saved in the related Data Group. Metadata can make reference to any kind of additional data, which are labeled in Fig. 1 as *post-processing objects*. Examples of post-processing objects are statistics on the raw data, charts, images, and any type of data that are produced from or relevant to the raw data.

F. Furfari and F. Potortì are with CNR - ISTI, via Moruzzi 1, I-56124 Pisa

D. Savić is with the Faculty of Electrical Engineering, Ljubljana (SI)

This work was partially funded by the European Commission under the COST 285 action.

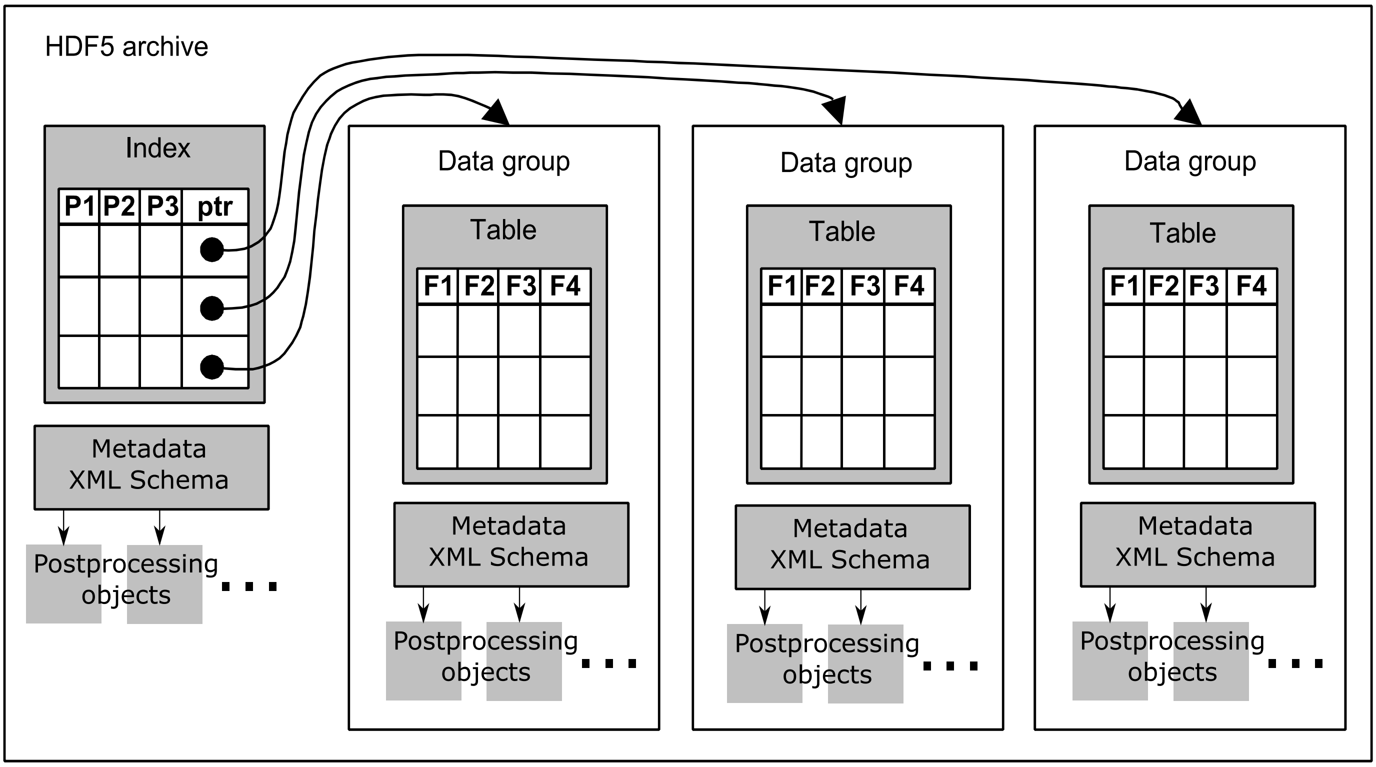

Fig. 1. Logical structure of the CostGlue HDF5 archive.

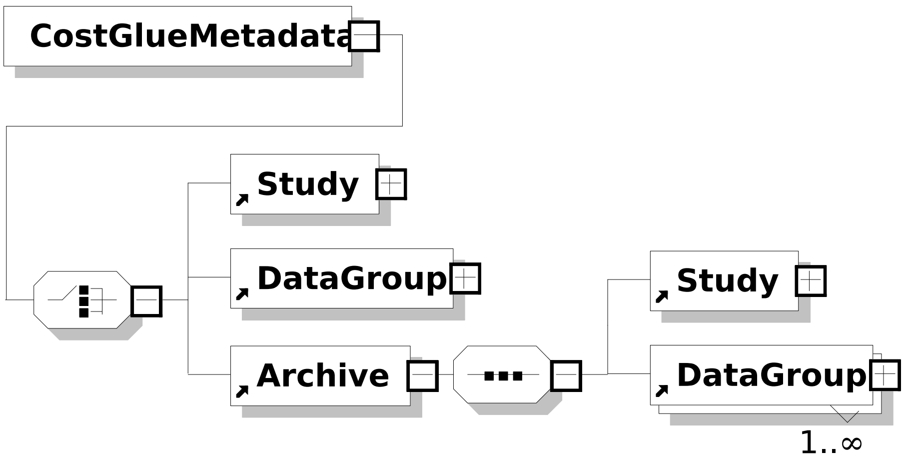

Fig. 2. The CostGlueMetadata element.

III. Schema description

The CostGlue metadata XML Schema, whose base element is depicted in Fig. 2, is inspired to the Reference Model for an Open Archival Information System (OAIS) [4] and uses parts of the CCLRC Scientific Metadata Model (CSMD) [5]. OAIS is a technical recommendation to provide permanent or indefinite long-term preservation of digital information. The objective of the CSMD model is to aid interoperability of scientific information systems among research organizations. The adoption of a common XML schema could facilitate further aggregation of telecommunication archives in catalogues to be published in repositories for the scientific communities such as CRAWDAD [6].

Three main elements are present, one for the metadata relative to the root (Study), one for those relative to a Data Group (DataGroup), and one that puts them together (Archive), as shown in Fig. 3.

Figure 3 illustrates the structure of the metadata relative to the root, which is stored in the root of the

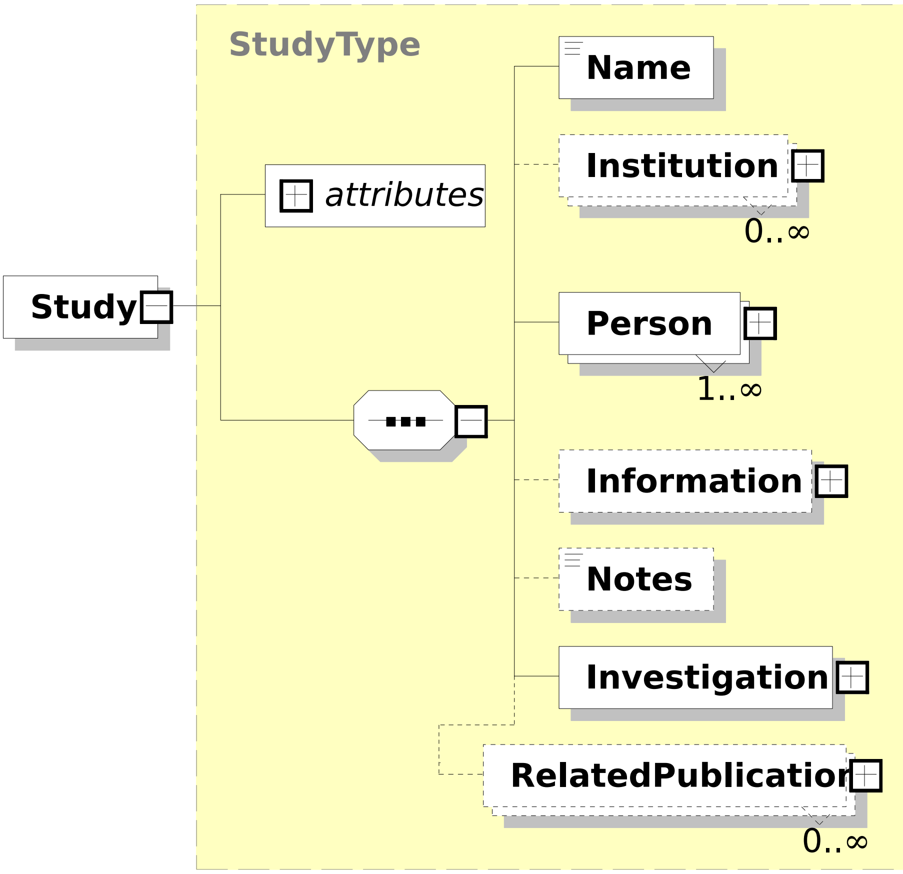

Fig. 3. The Study element.

archive together with the index table. The Institution, Person, Information and Notes elements are those defined by CSMD. The Study element and its embedded Investigation element reflect the taxonomy used in the CSMD model, where instances of an Investigation can be Experiment, Measurement or Simulation.

The Investigation element, which is thus central to the CostGLue schema, is depicted in Fig. 4. It is a close match to the one defined in CSMD, with a custom DataHolding element which is able to describe the data structure used in CostGLue. A notable addition is the PostProcessingObject element, whose puprose is allowing to embed arbitrary data inside the archive, such as graphs or statistical characterisation of data relative to the whole archive.

Figure 5 illustrates the metadata relative to each Data Group, which is stored in the Data Group together with the Data Table. One notable element is Software, extended from the CSMD one, which additionally includes the possibility of including the whole program source, the patches used and the input file. These are particularly relevant for a simulation or measurement environment where the simulator or the measurement instruments are heavily or totally software-based. PostProcessingObject elements can also be added to Data Groups.

IV. Schema usage

Most of the elements in the schema are optional. This choice is motivated by the need not to impose an excessive burden on the experimenter. The drawback of this choice is that it will be possible to have vastly incomplete metadata. However, if the objective is cooperation between experimenters, it is easy to envision that a CostGLue module can certify variable degrees of completeness of the metadata, so that repositories can accept only archives that comply to a certain degree of metadata completeness.

Some information in the schema requires verification during ingestion of new data and some can be automatically generated. One case is redundant information, such as the DataGroup/Parameters element,

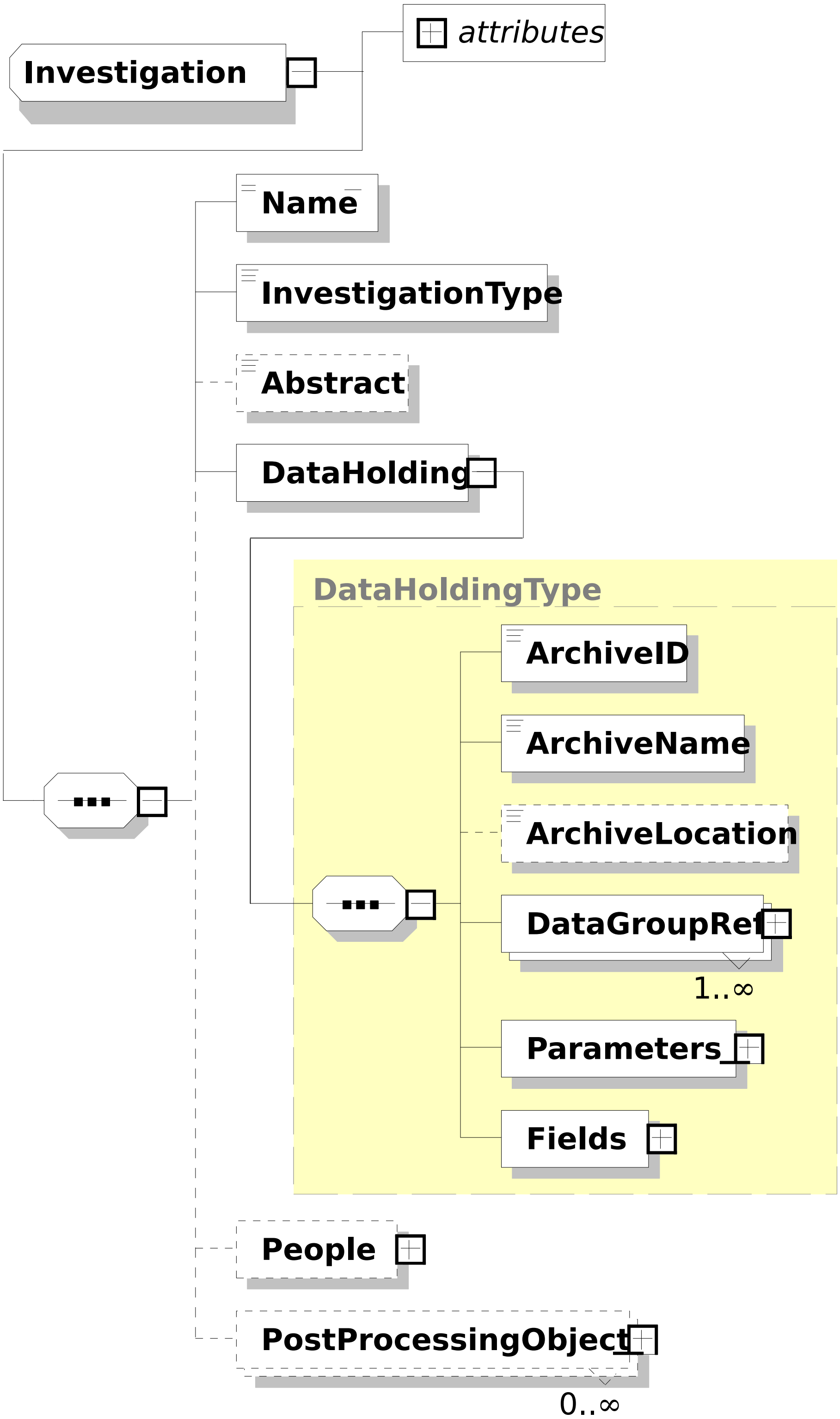

Fig. 4. The Investigation element.

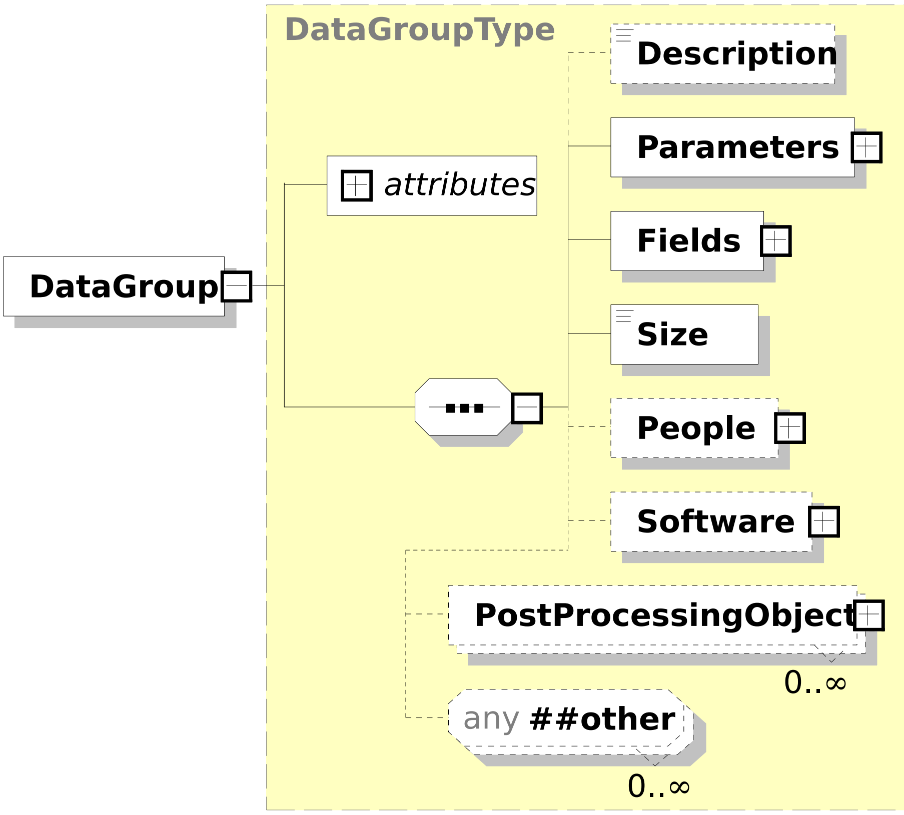

Fig. 5. The DataGroup element.

which is a copy of the Index row that points to the Data Group. Another one is elements referring to other elements, such as PostProcessingObject, which is a reference to an opaque object stored in the archive.

The metadata are constructed so as to be separable from data, so that a complete description of the archive can be easily produced, which requires little data storage and can also be made part of repositories, because metadata include pointers to the archive location. Datagroup elements can be extracted separately for inclreased efficiency and flexibility.

The schema does not include a way to serialise the archive data. This choice was made because we see the possibility of exporting the whole archive (data plus metadata) in XML format as a feature at a different level. Specif5Dically, HDF5 has an XML schema and a tool to convert a whole HDF5 binary file to XML, both available at http://www.hdfgroup.org/products/hdf5.

References

[1] D. Savić, M. Pustišek, and F. Potortì, "A tool for packaging and exchanging simulation results," in *proceedings of the International Conference on Performance Evaluation Methodologies and Tools (Valuetools)*. Pisa (IT): ACM, Oct. 2006, p. 60.

[2] W3C, "Xml schema," World Wide Web Consortium, Tech. Rep., Oct. 2004, version 1.1. [Online]. Available: http://www.w3.org/XML/Schema

[3] M. Folk, R. McGrath, and N. Yeager, "HDF: an update and future directions," in *International Geoscience and Remote Sensing Symposium (IGARSS'99)*, IEEE, Ed., vol. 1, 1999, pp. 273–275.

[4] CCSDS, "reference model for an open archival information system (OAIS)," Consultative Committee for Space Data Systems, Tech. Rep., Jan. 2002, cCSDC 650.0-B-1, Blue Book. [Online]. Available: http://public.ccsds.org/publications/archive/650x0b1.pdf

[5] S. Sufi and B. Matthews, "The CCLRC scientific metadata model: a metadata model for the discovery and exploitation of scientific studies and associated data," Science and Technology Facilities Council—(CCLRC), Tech. Rep., 2004. [Online]. Available: http://epubs.cclrc.ac.uk/bitstream/744/05_Brian_Matthews___csmd_core_grid_poland.pdf

[6] D. Kotz and T. Henderson, "CRAWDAD: A Community Resource for Archiving Wireless Data at Dartmouth," *IEEE Pervasive Computing*, vol. 4, no. 4, pp. 12–14, Oct. 2005. [Online]. Available: http://ieeexplore.ieee.org/iel5/7756/32928/01541962.pdf?tp=&arnumber=1541962&isnumber=32928