Developing and Accessing Scientific Databases with the Object-Protocol Model (OPM) Data Management Tools

I-Min A. Chen, Anthony S. Kosky, Victor M. Markowitz and Ernest Szeto
Information and Computing Sciences Division
Lawrence Berkeley National Laboratory, Berkeley, CA 94720
{IAChen, Anthony_Kosky, VMMarkowitz, E_Szeto}@lbl.gov

The *Object-Protocol Model* (OPM) data management tools provide facilities for rapid development, documentation, and flexible exploration of scientific databases. The tools are based on OPM, an object-oriented data model which is similar to the ODMG standard, but also supports extensions for modeling scientific data [4]. Databases designed using OPM can be implemented using a variety of commercial relational DBMSs, using *schema translation tools* that generate complete DBMS database definitions from OPM schemas [5]. Further, OPM schemas can be retrofitted on top of existing databases defined using a variety of notations, such as the relational data model or the ASN.1 data exchange format, using OPM *retrofitting tools* [2].

Several scientific databases have been designed and implemented using the OPM tools, including the Genome Database (GDB) at Johns Hopkins University School of Medicine,[1] the Primary Database of the German Human Genome Research Center in Berlin,[2] the Electronic Notebook for the Spectro Microscopy Collaboratory at the Advanced Light Source Beamline 7 at the Lawrence Berkeley National Lab,[3] and the Event/STAR database at the Relativistic Heavy Ion Collider at the Brookhaven National Lab.[4] Other scientific databases, such as the Genome Sequence Database (GSDB), [5] have been retrofitted with semantically enhanced views using the OPM tools.[6]

Tools for querying (native or retrofitted) OPM databases include a *query language translator*, which interprets queries expressed in the OPM query language (OPM-QL) and translates them into the languages supported by the underlying DBMS [1]. OPM-QL is similar to the ODMG standard for object-oriented database query languages, with extensions supporting the unique features of OPM. Web-based schema browsing and query tools provide access to OPM databases via the Web [3]. The OPM *schema browser* allows exploring OPM schemas represented graphically, while the OPM query interface allows graphical construction of ad-hoc OPM queries, and dynamic generation of Web (HTML) query forms. A variety of scientific databases are available on the Web for browsing and querying with these tools.[7] OPM *schema publishing tools* allow documenting OPM databases in a variety of formats and notations, including HTML and Postscript.

Multidatabase OPM tools have been developed as an extension of the core OPM toolkits, with support for: (1) assembling heterogeneous databases into an OPM based multidatabase system, while documenting their schemas and inter-database links; (2) processing ad-hoc multidatabase queries via uniform OPM interfaces; and (3) assisting scientists in specifying and interpreting multidatabase queries [6].

Incorporating a database into an OPM multidatabase system involves constructing one or more OPM views of the database and entering information about the database and its views into a *Database Directory*. The Directory stores information necessary for accessing and formulating queries over the component databases, including: general information required for accessing the database; structural information on the schemas of each database; and information on known links between databases, including semantic descriptions of the links, and data manipulations necessary in order to traverse such links. An example of a database directory is the Molecular Biology Database Directory available on the Web.[8]

Queries against an OPM multidatabase system are expressed in an extension of the OPM query language, that includes additional constructs necessary for accessing multiple databases. Multidatabase queries are processed by generating queries over individual databases, and combining

[1] http://gdbgeneral.gdb.org/gdb/
[2] http://www.rzpd.de/
[3] http://www-itg.lbl.gov/~ssachs/notebook/project.html
[4] http://gizmo.lbl.gov/jopmDemo/star.html
[5] http://www.ncgr.org/gsdb/
[6] http://gizmo.lbl.gov/jopmDemo/gsdb10.html
[7] http:http://gizmo.lbl.gov/jopmDemo/demoDbs.html
[8] http://gizmo.lbl.gov/DM_TOOLS/OPM/MBD/MBD.html

the results using a local query processor. The stages of generating individual database queries and manipulating data locally can be interleaved depending on the query evaluation strategy being pursued. The OPM multidatabase tools have been used experimentally on a number of applications, including a federation of molecular biology databases involving GDB and GSDB.[9]

Current OPM work includes further development of the OPM multidatabase tools, extending the OPM retrofitting tools to cover additional data models, and extending the OPM toolkit to support complex data types, such as DNA sequences and 3-dimensional crystallographic data.

Comprehensive information regarding the OPM tools is available at http://gizmo.lbl.gov/opm.html.

Acknowledgement. This work is supported by the Office of Health and Environmental Research Program and the Mathematical, Information, and Computational Sciences Division of the Office of Energy Research , U.S. Department of Energy under Contract DE-AC03-76SF00098.

References

[1] I. A. Chen, A. S. Kosky, V. M. Markowitz, and E. Szeto. The opm query language and translator. Technical Report LBL-33706, Lawrence Berkeley National Laboratory, 1996. Available at http://gizmo.lbl.gov/opm.html.

[2] I. A. Chen, A. S. Kosky, V. M. Markowitz, and E. Szeto. Constructing and maintaining scientific database views in the framework of the object-protocol model. In this proceedings, 1997.

[3] I. A. Chen, A. S. Kosky, V. M. Markowitz, and E. Szeto. Exploring databases on the web. Technical Report LBNL-40340, Lawrence Berkeley National Laboratory, 1997.

[4] I. A. Chen and V. M. Markowitz. An overview of the object-protocol model (opm) and the opm data management tools. *Information Systems*, 20(5), 1995.

[5] I. A. Chen and V. M. Markowitz. Opm schema translator (opm version 4). Technical Report LBL-35582, Lawrence Berkeley National Laboratory, 1996. Available at http://gizmo.lbl.gov/opm.html.

[6] V. Markowitz, I. Chen, and A. Kosky. Exploring heterogeneous molecular biology databases in the context of the object-protocol model. In S. Suhai, editor, *Theoretical and Computational Genome Research*. Plenum, 1996. Also available as Technical Report LBL-38181, http://gizmo.lbl.gov/opm.html.

[9] http://gizmo.lbl.gov/jopmDemo/gdbs_mqs.html