Pedro: a configurable data entry tool for XML

Kevin L. Garwood1,*, Chris F. Taylor3, Kai J. Runte3, Andy Brass1,2, Stephen G. Oliver2 and Norman W. Paton1

1Department of Computer Science and 2School of Biological Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, UK and 3European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received on November 14, 2003; revised on February 13, 2004; accepted on March 16, 2004
Advance Access publication April 8, 2004

ABSTRACT
Summary: Pedro is a Java™ application that dynamically generates data entry forms for data models expressed in XML Schema, producing XML data files that validate against this schema. The software uses an intuitive tree-based navigation system, can supply context-sensitive help to users and features a sophisticated interface for populating data fields with terms from controlled vocabularies. The software also has the ability to import records from tab delimited text files and features various validation routines.

Availability: The application, source code, example models from several domains and tutorials can be downloaded from http://pedro.man.ac.uk/

Contact: garwood@cs.man.ac.uk

INTRODUCTION
The capture and annotation of data are important tasks for bioinformatics. Many scientists spend a significant amount of time on such tasks, to enable data to be accessed by other scientists or interpreted by software. The widespread use of high-throughput techniques is increasing the cost and importance of effective data capture, the latter being recognized, e.g. by the development of models for recording details of transcriptome (Spellman et al., 2002) and proteome experiments (Taylor et al., 2003). However, such models tend to be complex, which means that the capturing of such data is time-consuming, and in need of effective tool support.

This paper describes the Pedro data entry tool, which has been designed for capturing and annotating genomic data for storage or dissemination using XML. A key feature of Pedro is that its interface is generated from an XML Schema (Fallside, 2001, http://www.w3.org/TR/xmlschema-0/) definition. This feature means that the tool can be used with different kinds of data for which XML Schema definitions already exist, and also allows designers of XML schemas to conduct early validation of models with users. Additional features of the Pedro tool include: the provision of context-sensitive help on both the generic features of the tool and on the schema-specific data fields (including the provision of example acceptable values, where they are available); and the provision of facilities to simplify the process of stock ing data fields from controlled vocabularies.

Initially developed as a solution for the proteomics community, and first released in February 2003, the tool has already crossed domains into, e.g. the Grid middleware community. As of January 2004, the tool had over 300 downloads.

USING PEDRO
Figure 1 illustrates the Pedro user interface. The left-hand panel is a tree view that supports exploration of the hierarchical structure of the XML file that is being edited. A specific element in that structure, a SAMPLE, has been selected in the tree view, and the result is presented in detail in the right-hand panel. Single-valued elements within sample, such as sample_id, are represented as text boxes into which values can be typed directly. Multiple-valued elements, such as Sample-Origin, are represented using lists. The interface provides a range of editing functions, not all of which can be described here. However, as an example, the Keep, Cancel and Delete buttons at the bottom right corner of the SAMPLE panel, respectively: write changes made to the sample to Pedro’s representation of the document; undo the changes made to the sample, which is restored to the previously kept form; and remove the sample from Pedro’s representation of the document.

Pedro has been used in different ways by different user groups. Two use cases are described below to illustrate possible modes of operation.

Use case one
A data modeller denotes records and fields, using XML Schema, describing the data to be captured in a particular domain. Pedro is then run on the resulting schema, and generates appropriate data entry forms. Next, domain experts ‘field-test’ the model (schema) by attempting to fit their
domain-specific data into the forms. Responding to the comments of the domain experts during this activity, the modeller adapts the structure and content of the forms, by altering the structure of the XML Schema file and re-running Pedro. By this process, the schema is evolved iteratively until the domain experts accept its structure and content. Pedro thus allows (non-technical) domain experts to participate fully in an abstract modelling process, by framing that participation as a straightforward data entry activity. Additionally, such a group can document their model’s fields for their peers, by supplying appropriate web pages to the tool’s context-sensitive help feature. At the end of a data-modelling session, the scientists have a data model they can use for data storage or transport and, of course, the option to employ Pedro directly as a data capture solution.

Use case two

After following the kind of developmental cycle outlined above, Pedro is adopted for use as a data capture tool. Typical users working with Pedro for data capture will not be informatics specialists, and will benefit from a number of features of the tool. For example, the graphical rendering of the XML data file as a folder tree in the interface, simplifies the process of navigating around large datasets. Furthermore, the software automatically enforces the rules laid out in the schema describing which data must be provided and which can be absent, and whether data should be numeric or textual, fit a particular regular expression or come from a list of options. Meaningful textual reports and graphical indicators inform the user as to the nature of any error or omission—this feature allows users to create complete, consistently structured data files. Other benefits to the end user include the context-sensitive help, the facility to cut and paste subtrees between files and to save invariant data as a ‘template’ for re-use, and finally, Pedro’s ontology services, which enable the use of standard terminologies without causing significant inconvenience to the user. Where specific elements in the model are constrained to hold values from a controlled vocabulary or ontology, users are prompted to select values that are suitable.

SUMMARY

The growing use of XML in bioinformatics and widespread use of XML Schemas in a wide range of other domains, means that increasing numbers of users will be required to record complex information of many kinds as XML. Although, there is an increase in the number of XML tools, such as xmlspy® (www.xmlspy.com), such tools do not tend to emphasize data capture and annotation, and we also do not know about such a tool that provides comprehensive support for ontology-based annotation. For example, neither the Text nor the Grid views of xmlspy® seem particularly suitable for high-throughput data capture or annotation tasks on complex XML Schemas, and no support is provided for ontology-based annotation. As such, Pedro, an open source data entry tool, can be seen to fulfil the software requirements of a number of different communities.
Pedro: a configurable data entry tool for XML

The software possesses a range of features that both facilitate data capture and ensure the integrity of datasets, and as such represents an effective XML-based data capture solution. Additionally, Pedro dynamically generates data entry forms for data models expressed using a subset of XML Schema, enabling its use as a rapid data-modelling tool. It provides comprehensive facilities for creating, browsing and modifying data files, can import records from tab delimited text files, and features a sophisticated interface enabling the population of data fields with terms from multiple controlled vocabularies.

ACKNOWLEDGEMENTS

The screenshot in Figure 1 illustrates proteome data captured using the Pedro tool by Thomas McLaughlin from UMIST. Pedro has been jointly funded by the BBSRC IGF programme CoGeME grant and by the UK e-Science Programme through the North-West Regional e-Science Centre.

REFERENCES