Building Quality into a Digital Library

http://purl.org/net/repository

Hussein Suleman, Edward A. Fox, Marc Abrams

Department of Computer Science
Virginia Tech
Blacksburg, VA 24061, USA
Tel: 1-540-231-3615
E-mail: {hussein, fox, abrams}@cs.vt.edu

ABSTRACT

The Web Characterization Repository contains a collection of internet log files used by researchers to analyze and improve on the architecture of the Web. This repository improves on prior collections by thoroughly testing the log files for format to assure a degree of data quality. Instituting quality control into the digital library addressed many complex issues including technical support for quality assessment, the definition of a workflow to achieve quality control, the assignment of tasks to different people and the definition and automation of quality assessment for log files. By reaching realistic compromises on these issues it was possible to build quality control as an integral part of the digital library.

KEYWORDS: certification, quality, XML, repository

INTRODUCTION

As the World Wide Web has grown in popularity, so has research characterizing its usage. The ultimate goal is to improve the architecture to better support the activities of users. This research was initially driven by sets of trace files (commonly referred to as log files) that were collected by individuals as needed. As time progressed, some organizations like the Internet Traffic Archive [1] established repositories for public use while others made data available for verification of research results. These data sets were used in many studies but some researchers expressed reservations because the data sets were not sufficiently general and in many cases contained errors. To address this issue, the Web Characterization Activity working group of the World Wide Web Consortium (W3C), resolved to set up a repository of trace files which are checked for errors and therefore of a higher quality than unchecked files.

REPOSITORY ARCHITECTURE

The repository is composed of a database of metadata and a set of tools and procedures used to update the database and ensure data quality. This database contains metadata related to trace files, publications and progams used by researchers. The latter two types of entries do not follow the same quality control guidelines as for the trace files, so they are not considered further in this paper. The metadata for trace files includes fields that indicate the results of validation tests on the files. Since the files may be stored at remote physical locations and/or owned by other organizations, achieving quality control requires a well-defined procedure, as discussed and illustrated below.

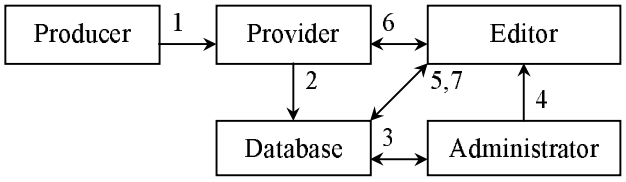

Figure 1: Workflow of submission process

Submission and Certification Process

The data is first generated by a Producer, typically the owner of a website or a network (in the case of network-level packet traces). The Provider of the data gets the trace files and submits them to the repository by entering the associated metadata into the online interface of the repository (http://purl.org/net/repository) and making the trace files available/accessible to the Administrator of the repository. The Administrator then allocates one or more Editors to the task of validating the files. An interaction between the Editor and Provider is initiated to understand the format of the files and create a formal Extensible Markup Language (XML) specification of the format. This specification is then used as an input to a validation program (executed by either the Provider or Editor) which confirms that each record in the trace file conforms to the expected format in terms of data types and ranges. The Editor studies the report from this validation program and resolves major errors with the Provider. Finally, if the data is deemed to have only minimal errors, the Editor flags the trace files as certified in the repository.

XML Format Specification

The validation process is driven to a large degree by the XML specification of the format of the trace files. XML is widely used as a markup language to describe structure in a text file. In order to maintain the original form of trace files, the format specification is stored separately. It is assumed that any trace file is a text file where each line is a space-delimited set of fields with a common format. The XML specification then indicates a list of fields and their composition in terms of either well-defined trace file field names (e.g., "size") or generic data types (e.g., "number"). Including parameters with the data types makes it possible to define the type of data contained in a field much more precisely and this supports a more accurate validation process (see Figure 2 for sample XML and trace file data).

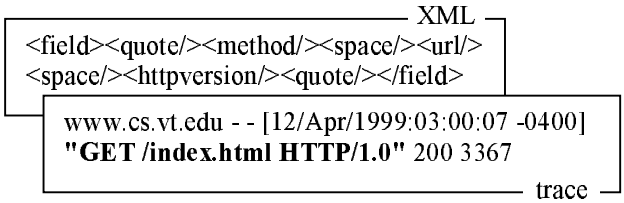

Figure 2: XML specification for one field of a trace file and a sample trace file entry with the field highlighted

IN SEARCH OF WORKABLE CERTIFICATION

Evolution of the System

When the repository project was initiated, the emphasis was on storing and retrieving of metadata. The digital library used for this purpose was the Mantis system from OCLC [2]. In setting up this digital library, the metadata set (the specific fields included in the database) and user interface had to be agreed upon. When the certification procedure was introduced, it was apparent that the original system could not be easily adapted to support this procedure so a custom-made digital library was constructed.

What is Certification?

There was much debate about the meaning of certification. The main concern was that any person using the data would trust a certification label and thus the repository ought to enforce that its data was of a high quality. Most importantly, the data ought to adhere to its stated format. It was also deemed desirable to ultimately perform basic statistical analysis on the data so that users of the data could know to what degree the trace files were representative of general internet usage.

Why XML ?

Before any validation could take place, the format of the original trace file needed to be known. To support multiple formats, Providers of data needed to be able to select a format for the files they submitted. Since most trace files are generated and/or processed by specialized software, it is not possible to specify a complete list of possible formats in advance. Instead, when the data is being submitted, there needed to be a mechanism to specify the format as a list of data types. Rather than define a new language for this, XML was chosen because of its emergence as a standard specification language. The additional benefits of XML are that its syntax is widely understood (due largely to its similarity with HTML) and it can be created and its syntax verified using readily available tools. The next challenge was to simplify the use of XML so that Providers of data who did not know XML could submit trace files. Firstly, all format specifications were entered into a database so that future submissions would not require new definitions. Secondly, an interactive interface was constructed using JavaScript to allow the user to specify the list of fields by type and then have the server generate an XML specification.

How Validation Works

The validation program parses the trace files and checks that the field structure corresponds to that indicated by the XML format specification. A comprehensive report lists occurrences of errors for the Editor to take note of when deciding if the quality of the trace files is acceptable or not. At the same time, the validation program can convert the file to a different format as specified by a second XML specification. Errors in the original file will then appear as missing entries (indicated by "-") in the new file so as not to interfere with the global statistics of the trace file.

CONCLUSIONS

Building quality into a digital library requires negotiation and compromise to achieve a workable balance between high quality and usability. In particular, a formal XML specification supported quality control procedures by making automatic quality assessment tools feasible.

OPEN ISSUES

The question of certification for trace files has been addressed. Other forms of entries in the database (e.g., links to papers) could also be certified in a manner analogous to peer review. Also important is the problem of maintaining quality of data when merging metadata from external databases.

DEMONSTRATION

A separate submission has been made to DL2000 for a demonstration of the interface and system interactions.

ACKNOWLEDGMENTS

The project was steered by the Web Characterization Activity working group of the W3C and the Network Research Group and Digital Library Research Laboratory at Virginia Tech. National Science Foundation grant NCR-9627922 partially supported this work.

REFERENCES

1. The Internet Traffic Archive, http://www.acm.org/-sigcomm/ITA/

2. Shafer, Keith E., Mantis Project: A Toolkit for Cataloging. *Annual Review of OCLC Research*, 1998. Accessible at http://www.oclc.org/oclc/research/-publications/review98/shafer/mantis.htm