Electronic Thesis And Dissertation Database Errors

Luke Schmader
Ryan Mestre
Client: Zhiwu Xie
• The Electronic Thesis and Dissertations mission:
  – “To preserve and provide access to the research and scholarship of Virginia Tech's graduate students”
• You can Browse through the theses on the library’s scholar page
  – Each record provides the type of document, author, URN, Title, Degree, Department, Advisory Committee, Keywords, Date of Defense, Availability, Abstract, and a link to the Document(s)
The Problem

- The ETD database filled with many different errors
- Parts of many papers are incorrectly displayed
- All of this must be manually fixed
The Mission

1. Find all the errors within the ETD database

2. Figure out patterns associated with each type of error

3. Make deletions and corrections if possible
Initial Setup

- Set up a SQL server in Linux along with Apache and phpMyAdmin

- Would allow us to query records with errors
  - See how errors displayed in database
Starting the Project

• Used a web browser to open an XML version of the database

• Take note of each error and figure out the pattern within the database

• Fix the error programmatically
 Errors

• Errors were found to be mostly control characters
  – Characters with an ASCII value of 0-31

• The errors were in many different contexts within the database

• Spanish characters were being displayed as due to XML encoding issue
The Solution

- Created a program that automatically finds and corrects all errors in ETD database

- Accepts error ridden XML, outputs usable and bug free XML

- Removes need for manual fixes of database
Design

- Hard code all error patterns into custom text file parser
- Replace Error cases with corresponding fix
- Output correct XML and Log File
The Expansion

- Client wanted a more versatile program

- Adjustments that make the program work for many circumstances and file types

- Needs… Configurability
Design

- Configuration file allows user to specify errors he wants to fix

- Toggle setting to auto fix according to Valid XML or other patterns

- Easy to use command line interface
Demo

FILEIN C:\Users\Luke\Documents\ETDdbFixer\etd_main.xml
FILEOT C:\Users\Luke\Documents\ETDdbFixer\etd_fixed.xml
FILELG C:\Users\Luke\Documents\ETDdbFixer\log.txt
FORMAT A
XMLFIX TRUE
BUGFIX Problem1 fix1 fix2
BUGFIX Problem2 auto
BUGFIX quantum fixer

ERROR IN LINE: 63566 at index: 260 with value: quantum
ries of relativity, quantum mechanics, c
Enter string to replace error or enter "!" to delete it. The choices you defined are:
1. fixer|
AUTOMATIC REPLACEMENT: fixer

ERROR IN LINE: 86656 at index: 73 with ascii value: 12
and Wind Tunnel Verifications of a Morphin
Enter string to replace error or enter "!" to delete it
Questions?