Database Creation and Information Extraction from ETDs

CS 4624

Virginia Tech

Blacksburg, VA

Spring 2013

Client: Venkat Srinivasan

By

Lamont Banks, Joseph Luke

1 CONTENTS

2 ABSTRACT

This project was done at the behest of the Computing Research Association (CRA). The main point of the project was to collect data associated with electronic theses and dissertations (ETDs) to allow determination of why graduate students in computing go into computing research. The deliverables include a database of the ETDs analyzed and a framework for manual approaches to this data extraction.

To accomplish these objectives, ETDs from North Carolina State University (NCSU), Florida State University (FSU), Auburn University (AU), Wake Forest University (WFU), and Virginia Tech (VT) were analyzed and inserted into the database. Extensible Markup Language (XML) was decided upon as the structuring format for the ETDs, and a tag structure was created utilizing biographical, educational, and institutional data from each ETD. Some of the tags included author name, title of the paper, year published, undergraduate institution of the author, etc. XML was chosen because of its prevalence in the ETD field, its structural properties, and ease of use. These tags were used to create the attributes for each entry in the database in Microsoft Access. Access was chosen mostly because of convenience and easy porting of tags into the system. However, the database could be moved into another system quite easily. In order to move the database, it would be converted to XML and then imported into a MySQL database or Oracle. Challenges that arose included missing data or insufficient information in various areas. For instance, many papers lacked information about source of funding, country of origin, and information about the author.

The second deliverable took the form of instructions (pg. 4) to an Amazon Mechanical Turk user on how to extract information. These instructions were created and provided in order to increase speed and decrease errors in manual data extraction. It was found that the basic structure of most ETDs is similar and is normally in this approximate order (dependent on institution of origin): title page, table of contents, abstract, actual content, biography, acknowledgements, and resume (not normally present). In these, all but the table of contents and the paper itself contains required information for the database. The instructions provide the most common locations for each tag/attribute and alternate locations (if any were found). They also instruct the Mechanical Turk user on what to do in case of missing data for each attribute.

3 USER MANUAL

3.1 AMAZON MECHANICAL TURK PROCEDURE

This section assumes that the user is familiar with Amazon Mechanical Turk and is using the following instructions to extract data from ETDs by hand. For each tag, the user should find the information in the standard location, or the alternate location(s) if necessary. All notes and examples should be followed and referred to. Unless otherwise stated, if an attribute cannot be found within the listed standard or alternate locations, the attribute should be left blank.

Table 1 ETD Tag Descriptions

Tag	<title>
Description	Standard Location
Records the title of the ETD.	Cover Page
	Alternate Location
	Notes/ Examples

Tag	<research _category>
Description	Standard Location
An integer ID representing one of the fields of computer science research that best categorizes this paper.	Extrapolated
	Alternate Location
	Notes/ Examples
	The categories and their respective IDs are listed in **Figure 1.** Left to the cataloguer's discretion.

Tag	<author>
Description	Standard Location
Records the author of the paper.	Cover Page
	Alternate Location
	Notes/ Examples
	All formatting included.

Tag	<institution>
Description	Standard Location
	Cover Page
	Alternate Location

Name of the school for which the ETD is filed.	Notes/ Examples Currently, all variations on institution names are retained. **Example:** "Virginia Tech" is recorded as well as the longer "Virginia Polytechnic Institute and State University".

Tag	<submission_year>
Description	Standard Location
Year the ETD was submitted for review.	Cover Page
	Alternate Location
	Notes/ Examples

Tag	<advisor>
Description	Standard Location
The primary advisor of the author.	Acknowledgements
	Alternate Location
	Cover Page
	Notes/ Examples
	Nearly all of the authors in the ETDs filed for this project make an explicit reference to their advisor. All name formatting is retained. In the case of multiple advisors, the first name listed is tagged. If there is no acknowledgments section or advisor mentioned by the author, this tag is left blank.

Tag	<advisor>
Description	Standard Location
The primary advisor of the author.	Acknowledgements
	Alternate Location
	Cover Page
	Notes/ Examples
	Nearly all of the authors in the ETDs filed for this project make an explicit reference to their advisor. All name formatting is retained.

	In the case of multiple advisors, the first name listed is tagged. If there is no acknowledgments section or advisor mentioned by the author, this tag is left blank.

Tag	<advisor>
Description	Standard Location
The primary advisor of the author.	Acknowledgements
	Alternate Location
	Cover Page
	Notes/ Examples
	Nearly all of the authors in the ETDs filed for this project make an explicit reference to their advisor. All name formatting is retained. In the case of multiple advisors, the first name listed is tagged. If there is no acknowledgments section or advisor mentioned by the author, this tag is left blank.

Tag	<num_acknowledged_colleagues>
Description	Standard Location
Records the number of explicit references to professors (including the advisors), fellow students, and colleagues made by the author.	Acknowledgements
	Alternate Location
	Biography
	Notes/ Examples
	Persons are only counted once; no duplicates.

Tag	<num_acknowledged_friends>
Description	Standard Location
Records the number of nun-professional explicit references made to friends and family of the author. The references counted for this tag should not overlap with those in **<num_acknowledged_colleagues>**.	Acknowledgements
	Alternate Location
	Biography
	Notes/ Examples
	Group terms such as “family” or “friends” should not be counted, **unless** there are no references to specific persons.

Tag	<undergraduate_institute_name>
Description	Standard Location
Name of the author's undergraduate institution.	Biography
	Alternate Location
	Notes/ Examples

Tag	<undergraduate_institute_degree>
Description	Standard Location
Name of the author's undergraduate degree.	Biography
	Alternate Location
	Notes/ Examples

Tag	<acknowledge_funding_org_1>, <acknowledge_funding_org_2>, <acknowledge_funding_org_3>
Description	Standard Location
Names of any explicitly mentioned funding organizations.	Biography
	Alternate Location
	Cover Page, Abstract
	Notes/ Examples
	Numerical suffixes should be dropped, and the general name of the funding body retained. **Example:** *"US Department of Energy Grant No. 005-D341"* ➔ *"US Department of Energy"* The tags are filled in the order in which the funding organizations are encountered; there is no tagging "hierarchy.

Tag	<prior_workplace_1>, <prior_workplace_2>, <prior_workplace_3>
Description	Standard Location
Names of any organization the author explicitly states having previous experience	Biography
	Alternate Location
	Acknowledgements
	Notes/ Examples

in **after** completing their undergraduate degree.	Our schema currently does not record the author's title/position while working.

Tag	<prior_research_area_1>, <prior_research_area_2>
Description	Standard Location
Any stated research/expertise of the author before entering graduate school.	Biography
	Alternate Location
	Notes/ Examples
	Example: (excerpt from an ETD biography): *[…]spent her senior year working for Dr. Michael Young on Liquid Narrative projects[…] continued to work for LNG in pursuit of a Master's degree through the summer of 2001[…]* The research area of the "Liquid Narrative" projects (Media and Visualization in this case) would be recorded in the tags.

Tag	<country>
Description	Standard Location
Home country of the author, if listed.	Biography
	Alternate Location
	Notes/ Examples
	Do not assume the country if not listed.

Tag	<city>
Description	Standard Location
Home city of the author, if listed.	Biography
	Alternate Location
	Notes/ Examples

Tag	<province_state>
Description	Standard Location
Home state/province of the author, if listed.	Biography
	Alternate Location

	Notes/ Examples

3.2 DATABASE

The database provided was made in Microsoft Access 2013 and should be very simple to use. The database will also open with no problems on a Microsoft Access 2010 installation, although previous versions were not checked. Once the database is open, double-click on the "etd revised" entry under the Tables tab on the far left side of the screen. You will find the attributes listed across the top of the table, starting with "num_pages." To edit an existing entry, scroll to find the attribute for the entry that you want to change, and click anywhere in that cell. To add a completely new entry, scroll to the bottom of the entries to a row marked with an asterisk (*) on the far left and begin entering information for each attribute in this row. You will notice that the asterisk changes to a pencil icon and a new row is created below with the asterisk next to it.

To delete an existing entry, click on the far left box in the row you want to delete in order to select it, then right-click and select "Delete Record." In order to export the database as a formatted XML file, click on the "External Data" tab at the top, in the same menu as file. There will be three panels in the new toolbar; the one on the left starts with "Saved Imports" and the one next to it on the right starts with "Saved Exports." Click on "XML File" in this second panel and follow the prompts to indicate location and name of the file, and what should be exported (data, schema, presentation). Leave the default option, which is to export data and schema (first two check boxes), selected unless you know that you need the third option.

Also included in the database file is another table entitled "Research Fields." This can be edited in the same manner to add, remove, or modify existing research fields to use in the tag "Research_Field"

4 DEVELOPER'S MANUAL

4.1 XML TAG DESCRIPTION

There were several XML tagging structures that were developed and tested throughout the project. The technical descriptions of the current tagging structure are described below.

1	Software Engineering
2	High Performance Computing
3	Media and Visualization
5	Human Computer Interaction
6	Education
7	Data Management and Information
8	Networking
10	Biologial Computing
11	Theoretical Research
12	Security
13	Artificial Intelligence

Figure 1. Research Fields

```
<title>Analyzing Software Artifacts Through Singular Value Decomposition to Guide Development
Decisions</title>
<Primary_Category>1</Primary_Category>
<Author>Mark Stephen Sherriff</Author>
<Institution>North Carolina State University</Institution>
<submission_year>2007</submission_year>
<Advisor>Dr. Laurie Williams</Advisor>
<Num_Acknowledged_Colleagues>22</Num_Acknowledged_Colleagues>
<Undergraduate_Institute_Name>Wake Forest University</Undergraduate_Institute_Name>
<Undergraduate_Institute_Degree>Computer Science</Undergraduate_Institute_Degree>
<Acknowledged_Funding_Org_1>IBM</Acknowledged_Funding_Org_1>
<Acknowledged_Funding_Org_2>Center for
Advanced Computing and Communication</Acknowledged_Funding_Org_2>
<Acknowledged_Funding_Org_3>National Science Foundation</Acknowledged_Funding_Org_3>
<Prior_Workplace_1>IBM</Prior_Workplace_1>
<Prior_Research_Area_1>1</Prior_Research_Area_1>
<Num_Acknowledged_Family>7</Num_Acknowledged_Family>
<Country>United States</Country>
<City>Salisbury</City>
<Province_State>North Carolina</Province_State>
</etd>
```

Figure 2. Sample Completed XML Structure

4.2 RATIONALE FOR XML TAGS

The overall goal of the XML structure was to capture information that would be useful in determining why students enter graduate research. The following is the description of the tags used:

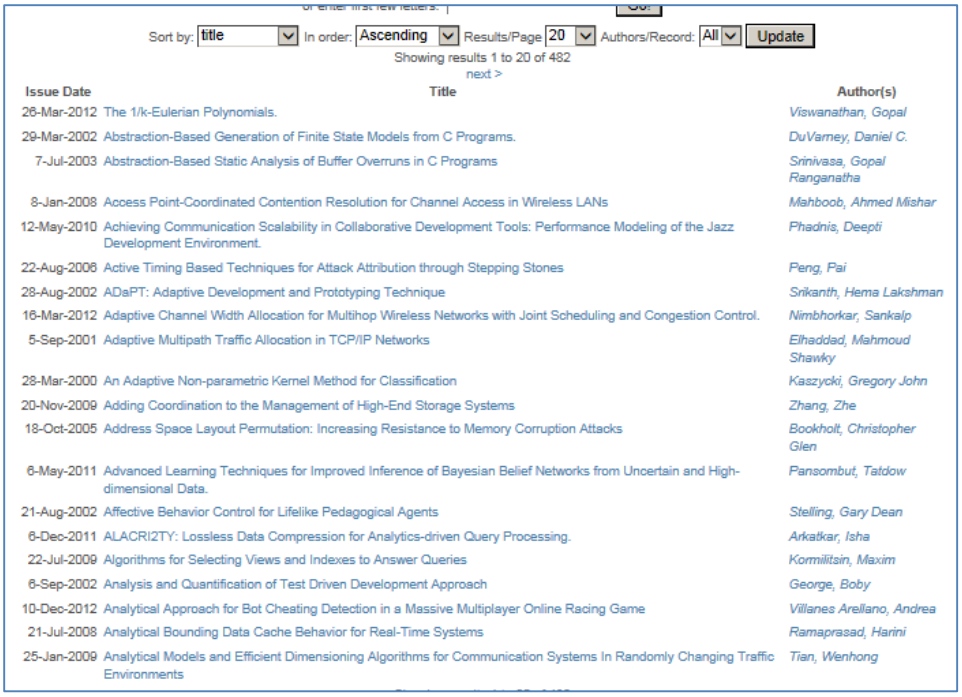
<title>, <author>, <institute>, <submission_year>

The ETD database should contain basic information about the dissertation; these tags were created to track these basic details.

<research_category>

Data mining of the collected data could reveal interesting trends in what types of research students pursed. Any useful data that data mining would yield is highly dependent on a balanced representation of the various fields. For this project, ETDs were selected randomly to the best of the compilers' abilities. Specifically, a search for Computer Science related papers was done on various university ETD databases, and 4-5 papers were chosen from each page of results.

Figure 3 shows a sample search results page:



Sort by: title In order: Ascending Results/Page 20 Authors/Record: All Update

Showing results 1 to 20 of 482

next >

Issue Date	Title	Author(s)
26-Mar-2012	The 1/k-Eulerian Polynomials.	*Viswanathan, Gopal*
29-Mar-2002	Abstraction-Based Generation of Finite State Models from C Programs.	*DuVarney, Daniel C.*
7-Jul-2003	Abstraction-Based Static Analysis of Buffer Overruns in C Programs	*Srinivasa, Gopal Ranganatha*
8-Jan-2008	Access Point-Coordinated Contention Resolution for Channel Access in Wireless LANs	*Mahboob, Ahmed Mishar*
12-May-2010	Achieving Communication Scalability in Collaborative Development Tools: Performance Modeling of the Jazz Development Environment.	*Phadnis, Deepti*
22-Aug-2006	Active Timing Based Techniques for Attack Attribution through Stepping Stones	*Peng, Pai*
28-Aug-2002	ADaPT: Adaptive Development and Prototyping Technique	*Srikanth, Hema Lakshman*
16-Mar-2012	Adaptive Channel Width Allocation for Multihop Wireless Networks with Joint Scheduling and Congestion Control.	*Nimbhorkar, Sankalp*
5-Sep-2001	Adaptive Multipath Traffic Allocation in TCP/IP Networks	*Elhaddad, Mahmoud Shawky*
28-Mar-2000	An Adaptive Non-parametric Kernel Method for Classification	*Kaszycki, Gregory John*
20-Nov-2009	Adding Coordination to the Management of High-End Storage Systems	*Zhang, Zhe*
18-Oct-2005	Address Space Layout Permutation: Increasing Resistance to Memory Corruption Attacks	*Bookholt, Christopher Glen*
6-May-2011	Advanced Learning Techniques for Improved Inference of Bayesian Belief Networks from Uncertain and High-dimensional Data.	*Pansombut, Tatdow*
21-Aug-2002	Affective Behavior Control for Lifelike Pedagogical Agents	*Stelling, Gary Dean*
6-Dec-2011	ALACRI2TY: Lossless Data Compression for Analytics-driven Query Processing.	*Arkatkar, Isha*
22-Jul-2009	Algorithms for Selecting Views and Indexes to Answer Queries	*Kormilitsin, Maxim*
6-Sep-2002	Analysis and Quantification of Test Driven Development Approach	*George, Boby*
10-Dec-2012	Analytical Approach for Bot Cheating Detection in a Massive Multiplayer Online Racing Game	*Villanes Arellano, Andrea*
21-Jul-2008	Analytical Bounding Data Cache Behavior for Real-Time Systems	*Ramaprasad, Harini*
25-Jan-2009	Analytical Models and Efficient Dimensioning Algorithms for Communication Systems In Randomly Changing Traffic Environments	*Tian, Wenhong*

Figure 3. Sample Search Results page

In this figure, results in descending alphabetical order, but the search parameters were varied while tagging documents. Meaning, ETDs were added after sorting alphabetically, sorted by year, or by author's name.

A more optimal selection method would should be used to get the most value out of this tag.

<advisor>

Any algorithms tracking this tag may find links between specific professors within a given university and the inspiration for the author's move into research. For example, an analysis of the ETDs may find that a particular professor is frequently cited as the motivation for students going into research or perhaps a large proportion of professors from a particular university is responsible for graduate student enrollment.

<num_acknowledged_colleagues>, <num_acknowledged_friends>

The intent behind this tag was to help identify the prevalence of (presumably) strong networks of supports in completing research projects.

<undergraduate_institute_name>, <undergraduate_institute_degree>

The schools that send significant numbers of students into graduate schools could possibly be investigated and studies to determine what programs and curriculum they have implemented to achieve their results. The degree of an overwhelming majority of the catalogued ETDs were in computer science (or extremely similar fields), but this cannot be assumed for all students, so this data has been tagged.

<acknowledged_funding_org_1>, <acknowledged_funding_org_2>, <acknowledged_funding_org_3>

Only about a fifth of the recorded ETDs' authors explicitly mention the source of the funding. The rationale for tracking this information is that if the author felt their funding sources to be worth noting, it may be important in their decision to begin research.

<prior_workplace_1>, <prior_workplace_2>, <prior_workplace_3>

The author's industry experience may influence what kind of research is done or how it is conducted. Analysis could be conducted to see which jobs move back into research more than others, if any, among other things.

<prior_research_1>, <prior_research_2>

Previous research experience is very likely to lead to more research, and analysis of this attribute could provide information about changes in research habits with age or other factors, or how research interests are distributed across the population of ETD authors.

<country>, <city>, <province_state>

This information is useful to track where the author comes from, lending itself to analysis on research done by students of different national origin.

4.3 Structure of the Database

The database currently consists of 2 tables: **<etd revised>** and **<Research Fields>.** <etd revised> is the primary table that contains the stored ETD data, <Research Fields> contains a listing of the various research categories used to classify documents in this project. There is a one-to-many relationship between <etd revised> and <Research Fields>, where an entry in <etd revised> can have one of many <Research Field> categories applied to it.

The structure of this database was kept purposefully simple to avoid imposing any classification restrictions. For example, there are currently many duplicate entries under the <institution> attribute of the database. To fix this, the name of a university should be standardized or a mapping table made.

The current database platform, Microsoft Access 2013, includes features for creating simple front-end user interfaces for editing data. This is not an essential portion of the database, and will likely not transfer to another platform. This project's interface, a portion of which is depicted below in Figure 4, contains simple text and drop down elements for editing data:

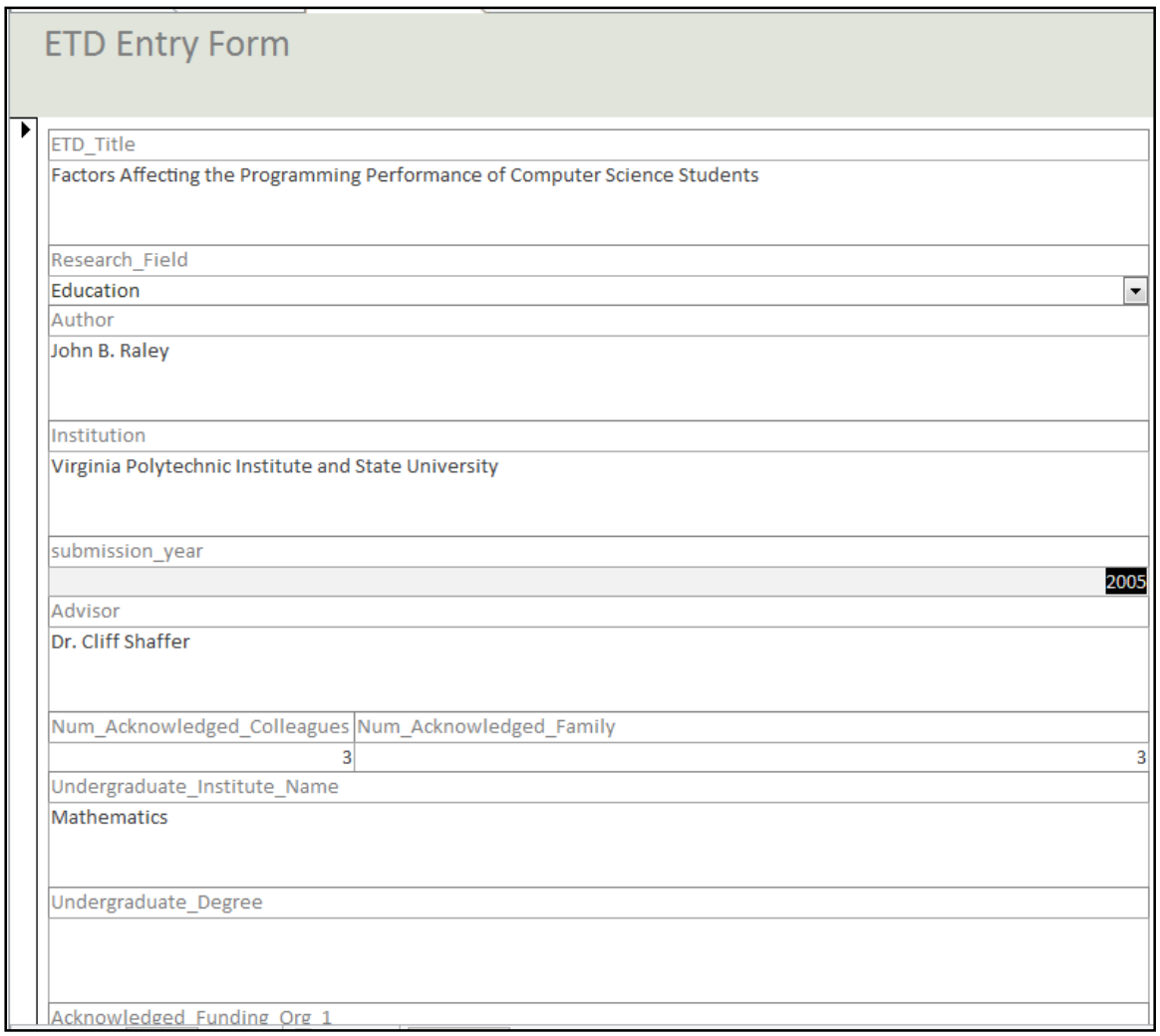


Figure 4. Microsoft Access Record Entry Form

5 LESSONS LEARNED

5.1 TIMELINE

- 2-22-13 – Began reading over resources, started to analyze and document ETDs (looked for information about the author and where that information was in the paper).
- 3-8-13 – Started to draft techniques to use for analysis and outlined database with XML tags.
- 3-22-13 – Developed preliminary database and began to develop Amazon Mechanical Turk procedures.
- 3-25-13 – Presented midterm presentation.
- 4-19-13 – Finished Amazon Mechanical Turk procedures, regarding how to find useful information in an ETD.
- 5-3-13 – Finished database and created final presentation.
- 5-6-13 – Presented final presentation.

5.2 PROBLEMS

Most of the problems experienced with this project had to do with missing or incomplete information. Most ETDs did not have all of the information predicted in our XML tags, and it was decided to leave these blank rather than try to extrapolate this information from the rest of the paper (except as where mentioned). To prevent confusion and differences in data entry, only explicit mentions of a particular attribute were entered into the database. When faced with multiple locations for an attribute (author's name in multiple places, for example), the most explicit location was used (this was normally the first mention).

Another small issue that was encountered was page numbering. The system created provides an attribute for number of pages (num_pages), but most ETDs have pages numbered both with Roman numerals first (title page through abstract) and then with standard Arabic numerals for the rest of the paper. To solve this initially, the total number of pages was used, regardless of numbering. Any page numbers used in the system were then indexed off of the title page (or first page if different) being page 1 and ignoring the nominal system. However, page numbers were removed from the XML schema later on in the project, and this ended up not affecting the final project.

Although not critical problem, another issue was deciding which universities and which papers from those universities to use. In the end, convenience and access to web sites restricted the universities to

North Carolina State University (NCSU), Florida State University (FSU), Auburn University (AU), Wake Forest University (WFU), and Virginia Tech (VT). Papers were then selected randomly from those available. However, this is an aspect of the project that can be improved in the future.

5.3 FUTURE IMPROVEMENTS

This project could be improved in many ways, both in its construction and its usage. The most obvious improvement is simply to add more entries. The more data that is contained in the database, the better any analysis of that data will be. Secondly, the database structure could be expanded to better use many of Access's features (cross-relations). This would allow for more use cases (querying for various combinations of attributes), but would restrict the database to Access. Similarly, the XML tags could be expanded to allow specific or just larger quantities of information from ETDs. Lastly, one piece of the project that could not be completed given the time requirements was to analyze the data. Analysis with a program such as WEKA could provide many relationships and clusters that are not immediately obvious. This would also allow testing predictions and applying machine learning algorithms to the data in order to predict attributes of future waves of computing researchers.

6 ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Edward Fox for helping us throughout this project, and filling in as client when ours was unable to perform this function this semester. We would also like to thank our original client, Dr. Venkat Srinivasan for the idea and impetus for the project.

7 REFERENCES

Aniket, P., & Fox, E. (2002). *XML for ETDs.* Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA.