

# Administering small- and medium-sized humanities database projects in an age of shrinking budgets

## PANEL

*ORGANIZER: David Gants*

---

*Electronic Text Center, Alderman Library, University of Virginia, Charlottesville, VA 22903*

### AUTHORS:

Tom Corns  
Roy Flannagan  
David Gants  
Thornton Staples

AFFILIATION: University of Virginia

E-MAIL: dgants@virginia.edu  
FAX NUMBER: (804) 924-1431  
PHONE NUMBER: (804) 924-3230

## I. Introduction

A humanities scholar contemplating computer-aided research on texts in digital format must pay careful attention to the source of those texts, lest inaccuracies and corruptions mar the analysis. Likewise a teacher wishing to employ electronic texts in the classroom needs to check that the text of the work in question meets certain standards of reliability. In response to these needs a number of well-funded projects have emerged, aimed at creating high-profile literary or linguistic databases, of which the electronic *Oxford English Dictionary* is probably the most well-known. A number of equally ambitious projects are currently underway, for example Jerome McGann's Rossetti Archive at Virginia and Peter Robinson's *Canterbury Tales* Archive at Cambridge. In these cases the undertakings have had the scholarly leadership and institutional funding to support the creation of databases that will withstand close scrutiny.

Most of the etexts now residing on Internet archives, however, have emerged from the efforts of modest, low-budget operations, and to date, little concerted action has gone into addressing the problems and issues faced by these smaller projects. Some of the currently-archived texts were originally put together as part of a specific study, and thus have been subject to various emendations called for by the requirements of the project itself. Still others have been scanned or typed into digital form by etext enthusiasts and lack the rigorous (and tedious) checking necessary to eliminate er-

rors introduced in the act of entry. In this second group the print analogue used as text source may itself be an inexpensive reprint or abridged "popular" edition, lacking the editorial rigor now applied to the creation of most scholarly editions. Furthermore, a significant number of texts have been systematically stripped of all marks of editorial intervention in preparation for mass circulation. Through the well-known work of the Text Encoding Initiative, and more recently the MLA committee to establish guidelines for the preparation of electronic texts, the humanities computing community has begun to produce a practical framework of common standards around which one might design a humanities database project. The TEI P3 in particular offers a large number of quite useful solutions to encoding problems, and the MLA guidelines will establish techniques for ensuring textual rigor. With work on these aspects of electronic text development well underway, we must next face the task of incorporating the emerging standards into small-scale projects, ones that frequently lack the recommended resources necessary to fully implement these standards.

This panel seeks to explore ways in which humanities database projects can maintain the high level of textual and intellectual integrity achieved by well-funded organizations while working within the limitations of time and financial resources experienced by most small-scale undertakings. The speakers will discuss their current projects, the problems they faced in maintaining high standards on low budgets, and the administrative and technological innovations they developed in response to those problems. Drawing upon practical examples stressing process they will illustrate database-specific implementations of the TEI P3, conversion of data compiled in smaller, proprietary applications to the broad SGML platform, synthetic integration of existing utilities and applications for text processing, relative feasibilities of product-delivery systems, and the challenge to database designers when dealing with widely-varied classes of data. They will additionally offer suggestions on strategies for organizing administrative structures that enhance modest human resources, as well as techniques for fund-raising.

## **Building a Mom and Pop Database: the 2000 books in the Milton Quarterly library become a relational database, through community effort**

*Roy Flannagan*

---

*Roy Flannagan, Department of English, Ohio University, Athens, OH 45701-2979*

AFFILIATION: Ohio University

E-MAIL: flannagan@ouvaxa.cats.ohiou.edu

Over the course of five years, every book in the *Milton Quarterly* library, a collection of about 2000 books, most of which dealt with Milton, Renaissance literature, theology, culture, or history; Puritan or English Revolutionary War history or theology; or classical Greek or Roman classical civilization or literature, was entered into a relational database, using the powerful library software ProCite. The database is a dedicated product, designed to be a unique aid for a Renaissance scholar who might want, say, to connect Donne with Calvin, or feminism with Milton. It now finds its ideal distribution through the *Milton Quarterly* home page on the World Wide Web. Open Text Corporation's Pat software will be used to query the database at the Electronic Text Center of the University of Virginia. The database will be copyrighted so that it cannot be lifted and sold as a unit; access will be limited to academic readers. The proprietary software Pat, developed at the University of Waterloo, will be used through a machine-license at the University of Virginia. ProCite generously supports output in RTF files, so that the *Milton Quarterly* can afford also to be generous in distributing its property.

Development work on the database gave valuable training to undergraduate apprentices, through the State of Ohio's Program Excellence and through the Honors Tutorial College of Ohio University. Graduates of the program are now in graduate school in English at Ohio State, at Stanford Law School, in teaching on various levels and, in the case of the most recent graduate, in a private business that employs skills learned in database construction. In each case, work on the database has given the student valuable training in scholarly methods and even in Aristotelian methods of collecting and labeling data empirically. Students have been treated as honored apprentices and allowed to be creative within the system. The im-

pression left on them was that of a happy factory, where workers set their own schedules and had a stake in both job description and product-creation. As administrator of the *Milton Quarterly* database, the speaker has built a useful and informative set of guidelines:

- Go to the strengths of the apprentice compiler (Greek, Latin, Japanese, critical theory, computer programming), and let each apprentice work within his or her circadian rhythms and on a unique schedule.
- Build the database using a flexible program that can talk across computer platforms and electronic mail addresses.
- When planning a project, bear in mind that a good database never has to end, and is perfectible. Mistakes can in input be forgiven and corrected. Mistakes in procedures can be amended.

## **The Rossetti Archive: From Conception to Publisher**

*Thornton Staple*

---

*Institute for Advanced Technology in the Humanities, Alderman Library, University of Virginia, Charlottesville, VA 22903*

AFFILIATION: Institute for Advanced technology in the Humanities

E-MAIL: tls@virginia.edu

The Rossetti Archive is being created by Jerome McGann at the Institute for Advanced Technology in the Humanities at the University of Virginia. From the beginning it has been designed as a complete archive of the written and graphical works of Dante Gabriel Rossetti and intended as a model for the development of electronic critical archives. I will trace the development of the project from its inception in 1992 through the present, focusing on the challenges presented by developing such a large project with relatively limited resources. Included will be a discussion of the creation of a set of SGML DTDs, data development and management, and project design and implementation. I will also discuss the evolution of McGann's publishing agreement with the University of Michigan Press.

## **Four computer-aided approaches to determining the provenance of Christian Doctrine, attributed to John Milton**

*Tom Corns*

---

*Department of English, University of Wales, Bangor, Wales, United Kingdom*

AFFILIATION: University of Wales

E-MAIL: [corns@bangor.uk.ac](mailto:corns@bangor.uk.ac)

Over the last two years an international and multidisciplinary team of scholars has attempted to approach with scientific detachment the currently much disputed issue of the provenance of Christian Doctrine. This is an account of the methodologies adopted, which combine digital imaging and stylometrics with traditional approaches augmented by computer assistance.

## ***The Studies in Bibliography* Full-Text Database**

*David L. Gants*

---

*Electronic Text Center, Alderman Library, University of Virginia, Charlottesville, VA 22903*

AFFILIATION: University of Virginia

E-MAIL: [dgants@virginia.edu](mailto:dgants@virginia.edu)

FAX NUMBER: (804) 924-1431

PHONE NUMBER: (804) 924-3230

In 1997, the Bibliographical Society of the University of Virginia will issue its fiftieth volume of *Studies in Bibliography*, a landmark in its history of continued scholarly excellence and diversity. The Society plans to mark this anniversary by publishing the first release of its Full-Text Database, an Internet-accessible, fully SGML marked-up, machinereadable database that will, when completed, contain the entire back catalogue of *SB*.

The *SB* Full-Text Database is essentially an in-house operation that receives very little outside funding, relying instead upon a collaboration of

resources among the Society, the University's Library, Electronic Text Center and Department of English, and the Institute for Advanced Technology in the Humanities. Because of the modest budget afforded the Database its participants have developed a number of procedures, practices and strategies for maximizing resources. Labor and equipment costs, two major elements with which all projects must deal, have been shared among the parties to take advantage of the rhythms inherent in the academic calendar. The project has developed techniques for text conversion that minimize the error-checking labors posed by the use of low-cost scanning equipment. It has also combined the strengths of various text processing tools such as UNIX utilities and the Oxford Concordance Program to increase efficiency of mark-up chores. When coordinated by a long-term administrative plan, these individual elements allow the project to create a high-quality database with a minimum budget.

The speaker will outline the basic operating system of the *SB* Database Project, provide examples of the TEI SGML mark-up scheme developed for the texts, discuss text conversion, markup and delivery procedures, and demonstrate the project's operation using Open Text's Pat search and display software, Internet Web browsers, and product access limitation components.