3. METHODOLOGY

Working differently

In 1949 Father Roberto Busa decided that building a complete concordance of Thomas Aquinas' works would be useful. At that time large-scale computing had become stable enough that he was able to persuade the then-head of IBM, Thomas J. Watson, to provide computational and human resources to the project (Pegues 105-6). What would have possibly taken a professional lifetime was done in a matter of a few years instead. Busa's research inaugurated a slowly building wave of humanities projects in which computing has been a core component. Although the *Index Thomisticus* was initially done on mainframe computers using punch cards, today it is freely available on the web as part of the Corpus Thomisticum project.[24]

Projects such as Busa's work on Aquinas have followed commercial and academic technological developments. One thing that has unified them over the past decades is the excitement over new possibilities of thinking differently presented by a computational approach to the study of cultural artifacts. This excitement continues to drive research that, in addition to its specific subject matter, often considers philosophical and neurological aspects of human perception and memory, and their implications for computer processing of cultural objects.

[24] 28 January 2007 <http://www.corpusthomisticum.org/>.

Philosopher Henri Bergson and scientist Vannevar Bush attempted to get at new modes of conscious thought by appealing to how our minds already work in the background. In his 1896 *Matter and Memory*, Bergson argued that there exists information which natural language cannot communicate. For him that information includes the associative processes of thought. "Essentially discontinuous," he wrote, "since it proceeds by juxtaposing words, speech can only indicate by a few guideposts placed here and there the chief stages in the movement of thought" (Bergson 116). In my continuing conversation with the Roland corpus both as a reader and as an analyst,[25] recurrent themes and imagery constitute the "words" which make up the "language" of the meta-Roland. If this analogy is accepted, then the encoding is a tangible representation of my thought processes, an articulation of complexity that needs to be understood, but that linear speech is unable to express.

Bush went so far as to propose a practical manifestation of our thought processes in the memex. He called it "a sort of mechanized private file and library"; a library of "books, records, and communications" stored on "improved microfilm," (As We May Think) in which one may create manual links that indicate semantic connections between pieces of information.

[25] This dual function has prompted the construction of *Roland^{HT}* as a Barthesian writerly text. In its multiplicity the Roland corpus facilitates, and implicitly suggests in the first place, the electronic project aim "to make the reader no longer a consumer, but a producer of the text" (Barthes 4). In the case of the electronic project, the text in question is the ephemeral sequence of excerpts seen, heard and read by a user over the course of a session. But the semantically encoded excerpt repository is also a writerly text. By making frequent explicit and implicit reference to other corpus objects, the primary sources of the Roland corpus indicate a persisting assumption of prior contextual knowledge on the part of the audience, making it difficult for the recipient to experience them as anything *other* than writerly texts.

While the memex may not have materialized, Bergson's and Bush's writings about the processes involved in thought and memory mingled with developments in computer science to produce such fundamental aspects of personal computing as the operating system based on the metaphors of folders, windows and [hyper]links.[26] The above intellectual and technological developments in turn contributed to the emergence of textual semantic encoding as a method of classifying and grouping information that can then be output in different ways.

A digitized text can be made richer by the addition of semantic code. The nature of this richness depends on the nature of the text. Addresses and telephone numbers may be defined and associated with names in a telephone book; pharmacological substances may be categorized and subcategorized; a word spelled in five different ways in a text written before spelling conventions set in may be indicated to be the same word. This last has been particularly useful to scholars studying materials created before the regularization of spelling, and in the case of *Roland^{HT}* — to tracing character names through different languages.

Knowing your instrument[27]

In order to use any resource critically it is important to know certain aspects of how the resource was created. Books, particularly ones whose authors are not well known, are valued more highly if they are published by reputable presses – and less if the editing is faulty. Scholarly journals and conference pro-

[26] In personal computing hyperlinks first appeared in the form of aliases pointing to real document locations.

[27] The following three sections were written in collaboration with Ethan Fremen.

ceedings that employ a peer-review system are trusted more than journals that do not. But evaluation of the trustworthiness of an above-named resource is only possible if the researcher understands the basic principles behind editing, error-correction possibilities and limitations of the print culture, as well as the mechanics of peer review.

The computing age is here to stay. Personal computers, the internet, search engines, online informational resources, subscription-based electronic distribution of journals, and social computing (photo and video hosting sites, tagging sites, massive multiplayer online role-playing games) have all become available and entered widespread use within the span of four decades ("personal computer"). Computers are ubiquitous in industrialized nations, and efforts are being made to provide access to computing and the internet to poorer parts of the world.[28] In the context of humanistic studies, computers and the internet are not merely tools for recording cultural heritage; they are part of it. Just as it was essential to understand the consequences of movable type for literature, in order to continue work in the humanities it is vital to understand the implications of computing in society, culture and pedagogy.

Valuable online scholarly resources are in continued development. The largest directory of such resources, Intute: Arts and Humanities (mentioned in Chapter 2), lists over 18,000 of "the best Web resources for education and research, selected and evaluated by a network of subject specialists." Among

[28] For more on the unequal worldwide distribution of computing resources, see the end of this section.

these resources is Wikipedia,[29] an open online encyclopedia created and edited by any interested party. The credibility of Wikipedia content has sparked significant controversy in academic circles. Nevertheless it is mentioned in twelve Intute-evaluated records, and over 75 records found by Harvester, a specialized search engine "created by taking each of the resources listed in the Intute: Arts and Humanities database and 'harvesting' up to 50 pages from those sites," whose search results "will not necessarily have been evaluated by Intute staff" (Harvester). Clearly Wikipedia is a presence that cannot be ignored in an educational setting. It has already compelled secondary-education teachers as well as college and university faculty to familiarize themselves with it well enough to advise students on how to best use it, and on its limitations.

Semantically encoded texts and corpora are in a different category from Wikipedia, and as a category have a set of technical features in common that are crucial to awareness of them as critical and pedagogical tools. These features in large part functionally overlap with those of older tools for scholarly research, such as indices and editorial practices. A basic understanding of the principles behind semantic encoding, search engines and organization of information in turn facilitates the appreciation of features and exigencies that are unique to the electronic medium. Examples of the latter include fragmentary and/or multilinear reading modes, visualizations of texts, interoperability with related projects, and – in the case of *Roland^{HT}* – presentation of semantic hypertextual corpora that would be simply impractical to outline on paper.

[29] 2 Apr. 2007 <http://wikipedia.org/>.

Extensive usage of resources built using these emerging technologies highlights the necessity for humanists to collaborate with technical experts at an unprecedented level. Collaborative practices enacted and successful so far have prompted humanities scholars to reconceive the concept of primary and secondary sources as data, and to radically restructure presentation of data results. Understanding results presented using new technologies requires that the recipient know something of these technologies, their benefits and limitations in a humanistic context.

For reasons outlined above, I will enumerate and explain the tools I have used to put the theory into practice, and the reasoning behind using them.

Tools

1. XML (Extensible Markup Language). This remarkably simple markup system is misleadingly termed a language. It might be more accurately described as a "metalanguage", or "metagrammar", as it is a tool for defining languages. The languages so defined are artificial languages designed for indentifying data elements, especially in text. In reality XML is a small set of rules, a syntactic framework with only a few predefined "words." Its simple syntax, outlined in Appendix E, makes it the most flexible currently available tool for semantic markup.

I chose to design and follow my own idiosyncratic encoding scheme rather than adhere to the Text Encoding Initiative's Guidelines for Encoding.[30] Working with a large primary-source corpus and a time constraint, I concentrated on inventing sub-categories for themes, imagery and characters, which together constitute all of the semantic (as opposed to structural)[31] encoding.

The TEI Consortium aims to provide a modular semantic encoding framework that would satisfy the needs of any encoding project in the humanities. The TEI Guidelines are "an international and interdisciplinary standard that enables libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation" (TEI Consortium). A considerable benefit of following this encoding standard is that by using the same formal semantic syntax disparate electronic projects are easy to cross-reference if the need should arise. Community-driven development of the Guidelines also centralizes theoretical work on semantic encoding, reducing the likelihood of duplication of effort.

So why develop a tagset specifically for *Roland^{HT}*? Two principal reasons led to this decision. First, the TEI Guidelines is a hefty document, spanning well over a thousand pages in print. For a researcher previously unfamiliar with them, the prospect of using the Guidelines for semantic encoding is a serious deterrent

[30] Several versions of the Guidelines exist; although there are paper publications, the most recent information about the Guidelines may be found at <http://www.tei-c.org/Guidelines2/>. 28 Jan. 2007.

[31] The encoding I call "structural" here is really also semantic; paragraphs, lines of poetry, segments of text etc. are meaningless as structural categories without existing in a specific context. The dichotomy is simply a convenient one for separating the substance of my argument from its structure.

to beginning the process in the first place. One of the hypotheses posited in the *Roland*^{HT} experiment is the usefulness of semantic encoding in the process of humanistic research and thinking. I have thus attempted to remove as many obstacles as possible to using semantic encoding, with the aim of making this mode of work available and attractive to scholars with little or no experience in the electronic medium.

The other reason for developing my own semantic tagset is that the TEI Consortium has been working with a predefined (though evolving) set of assumptions about electronic objects for over two decades. *Roland*^{HT}'s status as a semantic hypertextual corpus makes it at present a unique electronic object. Separating the encoding process from the TEI has allowed my primary sources to dictate their own semantic structure, without the necessity to fit into a pre-existing framework developed without these materials in mind. In theory — this is the TEI's goal — any concept a humanities researcher can imagine should be encodable using the modular tagsets it makes available. All of my code should therefore be translatable into TEI compliance in a straightforward manner; if this proves impossible, the incongruencies may be useful data for the developers of this now-established standard for encoding humanities texts.

In addition, the present dissertation is a proof of concept: the Roland corpus is far too large to process in the time allotted for dissertation writing. I plan to work on expanding it after graduation; the cataloging and semantic encoding of the corpus will be at least one lifetime's work. Hence, the conversion of *Roland*^{HT}

into TEI-conformant code will occur once the project — and the Guidelines — are relatively stable.

For a detailed description of the encoding process, please see "Process Description" below.

2. The *Roland^{HT}* interface was created in collaboration with Ethan Fremen. For display purposes we use XSL (the eXtensible Stylesheet Language) and JavaScript, which transform encoded information into something one can view in a browser. XHTML (eXtensible HyperText Markup Language) is utilized for basic web syntax.[32] We jointly edited the XHTML, CSS and XSL; Fremen contributed most of the JavaScript functions in use. The written-down XML structure is the product of solitary work on my part; however, it has been greatly influenced by conversations with Fremen, which have honed a common vocabulary and conceptual space usable by a programmer (Fremen) and a humanist (myself). In this aspect, *Roland^{HT}* is a practical example of the benefits of interdisciplinary collaboration.[33]

3. For text editing and encoding, I used several tools. OmniOutliner[34] is an inexpensive and simple to use software for hierarchical organization of ideas, with space to make notes. Scribe[35] is a free FilemakerPro[36] database created at

[32] XHTML is a derivation of HTML. Its principal differences from its parent language are stricter syntax requirements and the elimination of certain elements and attributes, whose functions were folded into stylesheets (formatting) or other elements/attributes.

[33] For related discussion see "Collaborative Possibilities" ff, below.

[34] 28 Jan. 2007 <http://www.omnigroup.com/applications/omnioutliner/>.

[35] 28 Jan. 2007 <http://chnm.gmu.edu/tools/scribe/>.

[36] 5 Apr. 2007 <http://www.filemaker.com/>.

the Center for History and New Media at George Mason University. It allows for the creation of source, author and note "objects" which can be interlinked arbitrarily and tagged with keywords. Finally, the XML editor oXygen[37] (limited-time renewable license for updates, less expensive for academics) was useful in encoding the corpus.

4. I used a versioning system, the open-source software subversion,[38] as a way of preserving the work in different stages of development. This has helped on a number of occasions when I found myself having gone in an erroneous direction. Almost all of these occurred during encoding;[39] any earlier version could be taken up again and developed as another branch of the database. Although it is possible to host both the subversion server (where the versioned archive is kept) and client (the working copy) on the same computer, in *Roland^{HT}*'s case the archive was hosted on a remote server.

The choice of software in this case proved to be helpful in an unexpected additional way. Some older versioning systems, particularly CVS (Concurrent Versioning System) and RCS (Revision Control System), have a flaw that becomes quite serious when backing up to a remote server. During the backup process, both CVS and RCS copy the entire document regardless of how much

[37] 28 Jan. 2007 <http://oxygenxml.com/>.

[38] 28 Jan. 2007 <http://subversion.tigris.org/>.

[39] Most notably, midway through the encoding process I found myself wondering whether to encode imagery of buildings (castles, fortresses, churches etc.). I created a version (snapshot) of the already complex XML file just before encoding the buildings. They proved to be insignificant to the corpus formation: they occurred too infrequently to be considered a consistently recurring element, and their geographic locations were not reliably noted. Since other imagery was present, it would have been time-consuming to remove the buildings encoding by hand. Instead, I restored the above-mentioned snapshot and proceeded with my work; the restoration took less than five minutes.

has actually changed. If this document is large and the network connection is unstable or something causes the version control system to crash, the incomplete repository becomes unusable and must be manually fixed. In addition, such a backup archive balloons quickly in size.

Subversion does something smarter: it only backs up the parts of a document which have changed. This has multiple benefits: the backup process is much faster; archive size is manageable. Atomic commits, as they are called, have no effect until they are completed; so if the connection is interrupted, the versioned archive does not break. Finally, subversion's approach allows multiple people to edit different parts of the same document. With older versioning software, any document must be "checked out," as from a library, and locked from other users until it is checked back in. After receiving the initial documents, subversion only checks in changes. If a user's changes happen to be in conflict with another user's simultaneous changes, the data is not allowed to be checked in until the conflict is manually resolved — but in this case the user gets the benefit of her colleague's recent work, and edits the document itself, which does not require the additional technical expertise required to fix RCS/CVS.

Placing importance on "atomic" changes in electronic files prompted my consideration of the importance of atomic academic input. Further discussion of the term can be found in "Virtual Humanities Lab" below.

Except for some of the text editing tools, all of the software I've used is free and open-source. This has been a conscious decision. As with many academic projects, the most exciting technological developments that have come out of the digital humanities community so far seem to be the result of funded work combined with a large amount of unpaid and largely unrecognized labor. Developing open-source software is a collaborative venture that the humanities need in order to flourish in the networked present. Supporting open-source projects has been a political act in a market still very much monopolized by political censorship, as well as products built by large organizations and financially inaccessible to most of the world.[40]

Strengths and weaknesses of the above-described approach

Our approach is standards- and client-based. All currently available features can be used with static files accessed locally. This permits *Roland^{HT}* to be archived on a CD, DVD or other storage device; access to a networked server is not necessary to view the work.

Because it is based on current standards, standards-compliant browsers now and in the arbitrarily far-off future can correctly display the work. XHTML 1.0, which is a stricter subset of HTML 4.01 (all XHTML 1.0 documents are also valid HTML 4.01 documents), allows us to exploit HTML 4.01's status as an International Standards Organization archival standard.[41]

[40] Internet usage worldwide varies more or less in direct proportion to national per capita income. According to the Google Gapminder World project (<http://tools.google.com/gapminder/>, 28 Jan. 2007, in beta), internet users per 1000 people are as low as 0.78 (Tajikistan). India and China, the two most populous countries, hover near the middle of the GNP/GNI range but count only 32 and 73 internet users per 1000 inhabitants, respectively.

[41] More information on the ISO is found at <http://www.iso.org/>. A description of the ISO HTML standard is at <https://www.cs.tcd.ie/15445/15445.html>. 29 Jan. 2007.

Our use of JavaScript complies with the Ecma-262 Language Specification,[42] and the XSL transformations conform to the World Wide Web Consortium's (W3C's) XSL 1.0 standard.[43] Our use of cascading stylesheets follows the CSS 2.0 standard, also by the W3C.[44]

These choices mean that, once the dissertation is submitted, it will be readable by any conformant future web browser. The thesis must be represented as a fixed work, which I will be unable to update as technologies evolve; and so we have made the current version as future-proof as possible.

Because *Roland^{HT}* is entirely client-side, our method would not scale well: a substantially larger corpus will require an XML database in order to remain efficient. As it stands, the roughly 850KB XML file containing all text data (but not images or multimedia files; those are stored separately) must be loaded entirely before the project can be viewed, rendering it unwieldy for slow connections if *Roland^{HT}* is accessed over the web. However, once it is loaded, subsequent operations are performed client-side and connection speed no longer matters, except when fetching the aforementioned images and multimedia files.

The interface is rudimentary. Excerpts are provided as a simple list, with no suggestion of their relative significance, or even immediate indication of when and where they originate. This is a strength insofar as *Roland^{HT}* is intended for exploration; however, during informal testing sessions users were briefly disori-

[42] 29 Jan. 2007 <http://www.ecma-international.org/publications/standards/Ecma-262.htm>. Ecma International's homepage describes it as "an industry association founded in 1961, dedicated to the standardization of information and communication systems."

[43] 29 Jan. 2007 <http://www.w3.org/TR/xslt>.

[44] 29 Jan. 2007 <http://www.w3.org/TR/REC-CSS2/>.

ented by the seeming lack of direction. This has been partially remedied by the inclusion of a help file into the website.

There is no visualization of the interrelationships among excerpts. As of this writing there is no standards-based way to effectively visualize networks. Scalable Vector Graphics (SVG) is a W3C recommendation whose implementation allows dynamic graphical manipulation. Of all the currently available browsers, only Firefox 2.0 and above implements this standard; but it does so incompletely. Increased dissemination of the SVG standard will afford possibilities for representing relationships within the Roland corpus in a visually sophisticated way.

Arbitrary string-based and/or semantic queries are not supported. The choice not to implement a search engine was made in order for the project to remain entirely client-side. Clicking on encoded themes, characters or imagery performs a "canned query," which reduces the list of excerpts to those in which the given element occurs. The greatest benefit of moving to a server-side XML database in the future will be the opportunity to create a robust semantic search interface.[45]

Process description

XML is an appealing humanistic tool in large part because it serves two complementary purposes. It is a syntactic framework for the expression of an

[45] Semantic search engines of the kind *Roland^{HT}* would need are already available, and intended to be customized for the needs of specific projects. The most notable such tool is PhiloLogic <http://www.lib.uchicago.edu/efts/ARTFL/philologic/>. 3 Apr. 2007.

overarching argument; on the other hand, it is also a tool for recording minute, quotidian results of research. XML's structural formalism has allowed me to combine two steps in the research process: gathering data, and putting it in an internally consistent format. It may seem that this is a time-saving move, but in reality the incremental nature of encoding a large data set has required at least the same investment of time as would work performed entirely in English. The advantage becomes apparent during data analysis and preparation of the web interface. Well-formed XML is easily queried using the XML companion syntax XPath, a feature built into oXygen; XSLT and CSS are powerful enough transformation tools to accommodate a web interface of arbitrary complexity.

Because I am using XML to build the project at the same time as I study my data set, there was no opportunity to get to know the corpus before encoding – in fact, the code *was* the way to get to know my data. It therefore seemed at the time that to start out with a pre-built Document Type Definition (DTD, see Appendix E) would be to impose an incomplete and/or incorrect structure upon the corpus. Instead, I closely re-read the HTML version of *Roland*^{HT} submitted as my Master's Thesis in Italian Studies at Brown (Zafrin). Already then the unidirectional, one-to-one hyperlinks[46] had been (manually) invested with semantic meaning. Each link within the text of a primary source led to another specific excerpt, and it did so indirectly, by way of a "blurb" that explicitly stated the connec-

[46] Encoding an anchor (`<a>`) hyperlink between two nodes in HTML enables the user to move from each one of them to the other but not back; anchors are thus unidirectional unless manually encoded otherwise. Unidirectionality is not a given in XML, which encourages categorizing rather than pointing, and then styling the code to suit particular needs. Thus, for example, `<theme>` elements in the XML version of *Roland*^{HT} cannot be said to have a single "direction."

tion expressed in that link. (See Appendix B.) The initial semantic encoding consisted of transferring the semantic information implicit and explicit in these hyperlinks and contextual blurbs into XML.[47]

The approach of beginning to build the DTD only after a significant amount of encoding had already taken place bore out. Being free to invent tags spontaneously left me to concentrate on the semantic patterns that revealed themselves as I became more familiar with the corpus. I also made the conscious decision not to use any pre-built standard DTD, such as that offered by the Text Encoding Initiative, or the Dublin Core Metadata Initiative (DCMI)[48]. Because I do not know either of them well, trying to make a literary/artistic critical argument in TEI-conformant code would be similar to learning a foreign language by writing a journal article in it. As mentioned above, *Roland^{HT}*'s idiosyncratically created code should easily translate into TEI.

Letting the code evolve from the primary-source analysis was key to the use of XML encoding as a tool for thinking. Throughout the encoding process, my understanding of the corpus has evolved in directions I could not have anticipated from the beginning. Changes and encoding inconsistencies became even more stark after the two-year hiatus that followed the encoding of the initial seven-work set, after which over thirty-five additional works were added. So, af-

[47] It should be noted that the semantic encoding found in *Roland^{HT}* is of two types: descriptive and interpretive. Descriptive elements are in turn of two types: structural (for example, paragraph and verse boundaries) and transmissive of information found in the original documents (author's name, work title). Interpretive elements include `<theme>`, `<imagery>`, `<note>` and `<textnote>`. In addition to the present theoretical discussion, he combined content of interpretive elements – element names, attribute names and attribute values – constitutes *Roland^{HT}*'s contribution to Roland scholarship.

[48] 28 Jan. 2007 <http://dublincore.org/>. See in particular <http://dublincore.org/about/>.

ter the first-pass encoding, it was necessary to normalize tag usage, particularly that of themes and imgery which had first occurred to me somewhere in the middle. I used oXygen's automatic document structure learning function to create a DTD for *Roland*^{HT}. I then combined XPath queries within oXygen with XSL transformations using custom stylesheets to extract statistics of theme and imagery elements, and later many of the attributes. Taking advantage of the perspective afforded by having recently worked with the entire corpus at once, I selected several that had occurred too rarely to be significant, and either deleted them or folded them into other elements. (See Appendix C.) This helped clarify the main semantic threads recurrent in the corpus.

The process of cleaning up the DTD was performed in consecutive iterations, and yielded an approximately 30% reduction in DTD size, bringing it down to about 80 lines. A bare-bones, valid TEI Lite DTD for the same XML document would be at least three times as long. The current proof of concept and corpus proposal deals with a small enough data set that this minimalist and idiosyncratic DTD suffices. Future expansion of *Roland*^{HT} – both in the amount of material with which it deals and in the need to interface with related online projects – will require a more elaborate and TEI-compliant DTD.

The unbearable incompleteness of code

The code is partially incomplete. This is the case in instances where an element or attribute was first implemented somewhere in the middle, or it turned out that an attribute enccoded interesting data that was not necessary for the

Roland-corpus argument. For example, some – but not all – of the characters are encoded as belonging to a certain religion (Christian, Muslim, Jewish, Pagan).

In case of family relationships, all of their instances were deleted from the encoded data set. The decision to delete them was made because, although in the future it will be interesting to explore kinship and loyalties to blood and sovereign, the encoding of these relationships was inadequate, and must be approached in a different way. The complexity of blood relations is high and information-rich enough to merit its own large project in the future.

Further work on the above-described semantic issues is being put aside until after this dissertation's completion. Why not just take them out altogether, as was done with families? The reason is simple: family relationships were expressed with bad code that would not have scaled up gracefully; otherwise they would have been left in as well. Taking out the partially-done code work would not significantly reduce file size. More importantly, compared with the laconic elegance valued in computer science, humanistic dialects of artificial languages are rooted in complex semantic disarray. They are permitted incomplete threads of thought, with the understanding that such threads are intended for further development. Properly used and well-documented metadata may suggest future research directions by its very structure. Partially encoded information can also be hidden from the end user until its encoding is complete, as was done with the rich encoding of proper names in the online text of Giovanni Boccaccio's *Decameron* (part of the *Decameron Web* project[49]).

Collaborative possibilities

Being messy in this semantically suggestive way is an advantage of a digital, encoded argument over one made in, for example, a journal article or a book. Writers of prose (specifically, of publications other than progress reports for ongoing projects) may explicitly suggest further research directions arising from their main arguments. But they have no incentive to do so: it is both time-consuming and implicitly discouraged by the academic community, which prefers a certain finality to the author's argument. Such finality is an illusion at best; many humanistic arguments and data are debatable and debated. Yet the misperception remains widespread, in large part because the physical properties of books and journals (when they are not electronic objects) encourages the feeling of well-defined intellectual edges to the work. Properly documented code can emphasize the nature of hypertext as an "open work" (Eco) – and, by its networked, open-source nature, encourage further work on *Roland^{HT}*.

This in turn should make it easier to build a community of interested scholars who would perform collaborative work and make its results publicly available. Reducing the barriers to entry by ensuring good code documentation and publishing an ongoing list of waiting tasks should make it easier for scholars

[49] In particular, social roles were encoded as one of the attributes in the element describing people. However, categories for those social roles are still under debate; so the stylesheets and search engine do not instruct software to read that attribute – it is skipped altogether. The *Decameron Web* is online at <http://www.brown.edu/decameron/>. 19 Apr. 2007.

to *decide* to participate, even with small contributions (a work mode that currently is not intuitive to most humanists outside of face-to-face conversations with colleagues). The success of an open-ended collaborative project greatly depends on keeping a certain momentum, and increased contribution volume would help provide this momentum.

Collaborating this way is a continuous, informal peer review process, and its result is a better-honed argument. It is as essential to an electronic project as peer review is to an article and editing to a book. But for the most part humanists are disinterested in collaborating electronically. There is no formal reward system set up for recognizing their input, so at the moment it would seem that any small contributions one makes to an online project (that is, not as a Principal Investigator on one) are a labor of love.

In a 2006 working paper, Daniel O'Donnell, James Cummings and Roberto Rosselli Del Turco consider why busy scholars would want to contribute to online projects. What are the rewards – what is the motivation for an already over-exerted, over-committed researcher to contribute to what amounts to a scholarly Wikipedia (with some important differences in format, methods and content)? In *Roland*^{HT} as in the Virtual Humanities Lab (see project description below), the rewards will be recognition of input, and a specific value placed on small bits of it – what at VHL we termed atomic input.

Virtual Humanities Lab

Virtual Humanities Lab (Prof. Massimo Riva, Principal Investigator) was a two-year project co-sponsored by the National Endowment for the Humanities and Brown University. From July of 2004 through August of 2006 I was VHL's Project Director. Our activities amounted to two major sub-projects. First, we semantically encoded three important but relatively little-known 14th-century Italian texts, created a web interface with a semantic search engine for them, and put them online for free perusal. Secondly, we built an annotation engine which allows scholars with (free, but moderated) accounts on our site to annotate and cross-reference parts of these texts, optionally anchoring their annotations to a specific phrase within the excerpt in question. As of the time of this writing, over fifty scholars from Europe and the Americas have annotators' accounts. Because the semantic encoding of these texts is partly interpretive as opposed to descriptive, contingent on resource availability, the encoding itself will be made visible to the users and disputable by the annotators.

Collaboration has been central both to the website and to the initial semantic encoding of the texts themselves. This was particularly true in the case of Giovanni Villani's historical account of Florence up to 1348, when its author died of the plague. *Cronica Fiorentina* was encoded by an Italianist working at Fitchburg State College in Massachusetts and a historian from University of Massachusetts at Dartmouth. The two – Rala Diakité and Matthew Sneider, respectively – had worked together before, when both of them had been pursuing graduate studies at Brown. But the physical distance between them necessitated online collaboration. Neither had much experience working remotely before; so they not only received basic XML training but were also introduced to subversion (see "Tools" above). In addition to occasional face-to-face meetings, we used telephone, email and a real-time chat system for resolving technical difficulties. The result of this collaboration is the first online publication of the *Cronica*, richly encoded by a team of scholars who are now translating it for the first time into English.

The networked project's potential to become collaborative both server-side and [web-] client-side is obvious, given the right built-in tools. Modelling itself on VHL, *Roland^{HT}* will be a venue in which people may work out their own arguments through dialogue, or else publish minor thoughts and findings that may be useful to themselves or their colleagues. All of an individual's contributions will be viewable at once – this should be useful to job search and tenure review committees. Individual contributions will be publicly tested and validated by interested researchers working in the relevant subject area(s).

VHL's example illustrates the value of atomic scholarly input – subject-relevant contributions that are the result of solid research but are too small to constitute a paper, article or book.[50] Combining an approach that values such contributions to a knowledge base on one hand, with a networked presentation on the other, implies a great deal of flexibility for participating scholars. Similarly to already-successful electronic means of communication (email, weblogs, discussion lists), VHL allows small information packets to be published and dis-

[50] This discussion is a modified version of the discussion of this topic in Riva and Zafrin (3).

cussed. Being unsuitable for a more extensive discursive format because of their seemingly incomplete, fragmentary nature, these bits of information may not otherwise be expressed at all (the processing of preparatory materials, including alternative philological or interpretive solutions, is often left implicit or made entirely invisible in printed editions). Reducing the minimum size of a contribution to the knowledge base from an article to a paragraph or sentence (or a modification of the encoding structure), provided a peer review process is still employed, increases the net amount of useful knowledge available for discussion. It creates the possibility for researchers to branch out and participate in more conversations, perhaps creating a distributed version of the encoding and editing process. While the eventual size and richness of the VHL (and, separately, *Roland*^{HT}) knowledge base will hopefully be significant, the time commitment required of any individual researcher is minimal. From this atomistic form of collaboration new (hyper)textual scholarly models can also be born.

Networks in the noosphere

By putting their arguments on the network and thus making them available for hyperlinking, searching and other manipulation, researchers help, in Eric Raymond's words, "homestead the noosphere." Raymond's homesteading paradigm comes from John Locke's theory of property: if a piece of land is not claimed, you claim it by working the land, investing of yourself in it. Similarly, Raymond writes, the open-source hacker community works by its members' claiming ownership of technological gaps by making (or else improving) software to fill those gaps.

The hacker community Raymond (himself a programmer of open-source software) describes is characterized by two main aspects: pragmatism, and gift culture.   This does not mean that there is no ownership: "[t]he owner of a software project is the person who has the exclusive right, recognized by the community at large, to distribute modified versions."   There are also rewards, the main among them being reputation, which in turn leads to a meritocracy. The community's pragmatic approach to knowledge work (write only software that people actively need or that has obvious widespread appeal), combined with the importance of reputation based on an individual's useful contributions, is effective in discouraging redundant knowledge production and increasing the usefulness of products to the community at large.

An approach to humanistic research similar to the above-described principles has proven useful at Virtual Humanities Lab. Our own feelings on this were corroborated by university faculty and administrators at "Transforming the Culture: Undergraduate Education and the Multiple Functions of the Research University," a Reinvention Center[51] conference held in Washington, DC in November 2006. There, Riva led a breakout session titled "Applying Principles of Learning and Technology In the Humanities and Humanistic Social Sciences: Creating New Modes of Scholarly Activity."[52] Riva's work on applied pedagogy of Italian

[51] 28 Jan. 2007
<http://www.sunysb.edu/Reinventioncenter/conference2006/urconfschedule.htm>.

[52] 28 Jan. 2007
<http://www.sunysb.edu/Reinventioncenter/conference2006/massimoriva/summary.htm>.

and humanities computing received enthusiastic interest by an audience of administrators and faculty from across the country.

Collaboration in digital humanities at large

Humanities' semantic encoding is messy in part because, like any other language, it attempts to constrain knowledge into discrete units. In reality these "units" are interconnected in such complex ways that they lack clear borders. As computer technology evolved humanists have addressed this by working closely with scholars outside of their general area, computer scientists, librarians, new media artists, and others. Whereas only a few decades ago dialogue among these disparate groups was rare to nonexistent, recent projects have taken progressively more interdisciplinary approaches to research.

Why collaborate at all? Humanities have over time evolved a system of largely individual production, occasionally supplemented by small-group (two- to three-member) collaborations. Why change this system?

The obvious answer is access. As has been pointed out numerous times, electronic dissemination of data is vastly cheaper than handling, transportation and storage of physical artifacts over the same geographic area. Digital objects are also easier to protect from destruction: one of the internet's biggest advantages is decentralized redundancy of data storage.

Until our ability to manipulate our own senses greatly increases, there will be no substitute for handling a physical book, vase or pyramid created long ago.

But, however imperfectly, their digital representations do convey a lot of information to interested parties removed from these objects by thousands of miles.

Perhaps more importantly, humanistic research methods that have been relatively stable for hundreds of years do not take into account the profound impact that networked communications and computation have had on humanity at large – on politics and warfare, on artistic production, on economic systems. If the aim of the humanities is to study all aspects of being human, humanists must incorporate the ways of thinking afforded by electronic media that have begun to emerge.

The future of collaboration

Now and for the foreseeable future, humanities scholars serve triple duty as researchers, teachers and scribes for future generations. It has always been this way; but with the advent of printing and reduced materials costs came an explosion of books and articles often redundant and unaware of each other, and an expectation of a certain peer-reviewed publication volume for tenure and promotion purposes. With the much increased volume of academic writing, it has become difficult to maintain a reputation-based reward system.

In his *Down and Out in the Magic Kingdom* (2003) writer Cory Doctorow posits a reward system that gives credit for contributions of all sizes – not just ideas incorporated into larger arguments by their originators, but also thoughts thrown out into the noospehere and taken up by others. In Doctorow's sciencefiction novel emphasis shifts from volume produced to *useful substance* produced.

Current technology is not quite ready to handle automatically updated social-approval scores and project them onto citizens' retinas upon request. But we have already begun leaving a digital trail; and it is no coincidence that social-networking software has gained popularity so quickly. It is the perfect compromise between, on one hand, the imposition of ideas upon a minority by a consensus majority, and on the other – enabling individuals to judge popularity for themselves based on sources they trust. For example, the social-linking site del.icio.us[53] allows every linked digital object to be viewed in different contexts – a registered user may see how many del.icio.us users have linked to it in general, or else how many members of a trusted subset have done so.

Computing has afforded some previously impractical or impossible forms of research itself. The most conspicuous example of this is our ability to quickly process large amounts of data. We have discovered new ways of looking at information – specifically, considering primary humanities sources as *data* – and presenting it for pedagogical purposes. Projects such as the Ivanhoe game,[54] the Rossetti Archive,[55] Valley of the Shadow[56] and Great Unsolved Mysteries in

[53] 28 Jan. 2007 <http://del.icio.us/>.

[54] 28 Jan. 2007 <http://speculativecomputing.org/ivanhoe/>. The summary at the referenced URL claims that Ivanhoe "promotes self-conscious awareness about interpretation and seeks to encourage collaborative activity in fields such as literature, religious studies, history, and other humanities disciplines."

[55] 28 Jan. 2007 <http://www.rossettiarchive.org/>. Presents and analyzes all of Rossetti's art, originally created in different media. For a list of media in which DG Rossetti worked, please see http://www.rossettiarchive.org/exhibits/index.html.

[56] 28 Jan. 2007 <http://valley.vcdh.virginia.edu/>. Presents information on Civil War Virginia.

Canadian History[57] would have been impossible before electronic networking, and have proven both useful and appealing to students.

In the humanists' task list, the creation of new, previously unavailable kinds of projects is joined by digitization and preservation of extant artifacts, textual and otherwise. Collaborative efforts allow for greater longevity of such digitized materials – individuals may leave but the project goes on, new maintainers protecting it from obsolescence in the face of the fast pace of technological developments.

Too much expertise is needed to accomplish all of this for anyone to do it alone. Infrastructures are difficult to maintain and often functionally redundant, so the number of separate infrastructures will be minimized by the desire for efficiency. Besides this, even specialized humanities topics are often large enough for many different researchers to work on them. At times these researchers live so far apart that the only effective way for them to follow each other's work – indeed, to know who is working on their topic – is the internet. Thus collaborative, cross-disciplinary efforts make more sense to pursue than individually run digital projects.

At the 2006 Pauley Symposium titled "History in the Digital Age," held at the University of Nebraska-Lincoln, Alan Liu pointed out another very practical reason for which collaboration among humanists, and between humanists and computer scientists, is beneficial to a project. "Spare change" that might have

[57] 28 Jan. 2007 <http://www.canadianmysteries.ca/indexen.html>. Is used, among other things, to teach Canadian history by way of assignments in which students "play detective" by examining archival materials of unsolved crimes from the 19th and 20th centuries.

otherwise fallen through the cracks of departmental administration, Liu said, tends to find its way to such projects. Humanities departments in particular benefit from increased equipment access in science departments. Scientists, on the other hand, work with data sets they may find interesting and unique.

The model of the conference is reasonably well developed in the humanities. However, humanists (especially those who do not practice workshopped forms of art) are still working out the kinks of workshopping their intellectual production. Formal and informal infrastructure necessary to conduct digital humanities research is also still in development. It is heartening that digital humanities efforts have generally been met with support and enthusiasm on the part of academic and "lay" users, as well as private and public funding agencies.

We are currently in the middle of a unique confluence. On one hand, the digital humanities knowledge base is growing at a healthy speed. On the other, recent developments in electronic communications make it relatively easy to *practice* digital humanities with only a moderate learning curve. Combining overlapping fields of expertise is the most efficient way to take advantage of this moment.