The Rossetti Archive: Methodologies and Praxis

Andrew M. STAUFFER

Abstract. As the Dante Gabriel Rossetti Hypertextual Research Archive nears the completion of its first phase, we present some of the challenges we have faced in tagging this large body of documents, and offer some potential solutions, as conceived within the realm of Standard Generalized Markup Language (SGML). Issues of textual description, rendering, and linking are discussed.

Résumé. À l'heure où la « Dante Gabriel Rossetti Hypertextual Research Archive » arrive à la fin de sa première phase, nous présentons ici quelques-uns des problèmes rencontrés lors du balisage de ce vaste ensemble de documents. Nous proposons également quelques solutions possibles dans le cadre du *Standard Generalized Markup Language* (SGML). Les problèmes de la description et du rendu des textes sont abordés, ainsi que ceux touchant à la création de liens hypertextuels.

Keywords: Electronic publishing of scholarly works, SGML, Dante Gabriel Rossetti.

Mots-clés : Édition électronique de travaux scientifiques, SGML, Dante Gabriel Rossetti

Currently under development at the Institute for Advanced Technology in the Humanities at the University of Virginia, the Rossetti Hypermedia Research Archive is an electronic database of the complete works of the Victorian poet and painter Dante Gabriel Rossetti. It combines digital images (of manuscripts, printed pages, and artwork) and tagged ASCII texts, all fully cross-referenced via hypertext links within and among files. In development since the early 1990s, the Rossetti Archive has constituted the leading edge of electronic scholarly editions, having first confronted many of the issues and problems involved in such projects, and now approaching the completion of its first installment, to be issued in electronic form by the University of Michigan Press in 1998.

✉ University of Virginia; Department of English; 219, Bryan Hall; Charlottesville, VA 22902 (USA).
Fax: +1 804 924 7105 E-mail: ams4k@virginia.edu

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXXIII, 1 à 4, 1997. C.I.P.L. - Université de Liège - Tous droits réservés.

Jerome McGann, the general editor and creator of the Rossetti Archive, has written several excellent summaries of its background and purposes, and these are readily available both on the World Wide Web, and in print.[1] Further, it seems that the basic editorial implications and radical functionality of electronic archives no longer need to be asserted or explained, given the amount of ink (digital and otherwise) spilled on such topics in recent years.[2] More useful now might be a presentation of some of the general lessons we have learned in developing the Archive, as illustrated by some of the specific documentary instances that allowed us to learn those lessons.

Almost all electronic editions to date, including the Rossetti Archive, use some form of Standard Generalized Markup Language (SGML) to describe and present their materials. Projects which are currently being published on the World Wide Web use Hyper-Text Markup Language (HTML), a subset of SGML that has become the standard for Web browsers. Many scholars and scholarly collectives have found in the Web a fast and inexpensive way to publish editions, and have thus adopted HTML as their tagging language. The problem with this approach is that HTML is a rendering language only: it tells the computer how to display the text on the screen, but it does not allow the editor to describe the logical, structural, or material features of the text. As a result, a great deal of documentary evidence must be edited out of these Web editions, or preserved in ASCII commentary alone. SGML, on the other hand, is a descriptive language, to which rendering protocols can be mapped when desired. From the start, the Rossetti Archive has been conceived as an SGML-based project. Each document in the Archive has thus been

[1] These include the "Introduction" to the Archive itself (available on the World Wide Web at URL http://jefferson.village.virginia.edu/rossetti/introduction.html), MCGANN: 1994, MCGANN: 1996, "The Rossetti Archive and Image-Based Electronic Editing", and MCGANN: 1997.

[2] In addition to the essays collected in the FINNERAN volume, see MCGANN: 1996, "The Rationale of Hypertext", available on-line at http://jefferson.village.virginia.edu/ ✂ jjm2f/public/rationale.html; LAVAGNINO: 1996; LANDOW and DELANEY: 1993, eds.; SHILLINGSBURG: 1986; and BENDER: 1976. For examples of other electronic archives under development, see the "Research Reports" of the projects at the University of Virginia's Institute for Advanced Technology in the Humanities, located on the Web from the homepage at URL http://jefferson.village.virginia.edu. In addition, helpful recent discussions of electronic textuality can be found at the website for Bruce GRAVER's panel, "Electronic Texts and Textuality", at the 1997 NASSR conference: http:// ✂ www.inform.edu/RC/pages/cex/features/graver-panel.html.

marked with a large set of descriptive tags that self-consciously preserve its unique structures and appearances.

Another subset of SGML, based on the language protocols developed by the Text Encoding Initiative (TEI), was also rejected because it made scanty provision for describing the physical features of documents, and because it assumed certain structural and presentational regularities that poets like Rossetti make it a rule to violate. We have developed a unique version (or "flavor") of SGML, designed to account in the most complete and faithful way for the varying features of Rossetti's works. In this article, I will focus on some of the features of the language the Rossetti Archive uses to describe its written and printed documents, leaving aside the pictorial works, which present their own set of issues. In particular, I want to discuss some of the particular struggles between SGML and our documents, and offer some insight into the evolution of our thinking about textual encoding, description, and rendering of the Rossetti Archive documents.

1. Can Logic Contain Art?

As the blacksmith and prophet Los begins his "terrible labors" in William Blake's *Jerusalem*, he announces, "I must Create a System, or be enslavd by another Mans."[3] So too the creators of the Rossetti Archive avoided pre-existing versions of SGML in hopes of escaping servitude to a logical system that was not sensitive to the documentary situations at hand. For example, Rossetti was extremely attentive to the physical appearance of his printed books, so our version of SGML had to include tags for describing material features such as bindings and endpapers. Furthermore, Rossetti revised frequently and produced many intermediate (or alternate) versions and texts—many of which have survived—so our markup language had to be able to account for complex relations among manuscripts, sets of mixed proof pages, one-shot Trial Books, and "double works" which exist both as paintings and poems, among other things. Suffice it to say that the TEI protocols required extensive modification and expansion, and the result has been the evolution of a unique set of logically-related descriptive tags.

[3] ERDMAN: 1970, ed., p. 149, 151. The idiosyncratic punctuation here is Blake's own, rendering nicely the burden of Los' statement.

In SGML, the master file that contains such a set of tags with their relations defined and made explicit is called a Document Type Description, or DTD. We have modified the Rossetti Archive DTD many times as new materials have challenged our previous assumptions about the kinds of logical and physical structures a document might contain. As a result of this often *ad hoc* process of evolution, our DTD may not be as lean as those for which all documentary features and relations can be accurately predicted in advance. Yet it seems that for large and complex collections of materials, particularly those not previously edited by 20th-century scholars (as Rossetti had not been), such complete predictions are impossible. So, for the Rossetti Archive, most often we have had to devise ways to escape not enslavement by another man's system, but the constraining implications of our own previous incarnations of the DTD.

This process of continual adaptation has revealed one underlying limitation of SGML as a means of describing large sets of documents: SGML is by definition a mechanism of logical regularization. No matter how extensively you adapt the DTD to fit a group of texts, you always create a system of rules that defines what is possible. Each time you add a new document to the database, it has a potential to disrupt this system—a potential that increases with the great attention to detail required of scholarly editions. For small-scale archives, or archives in which the material is very regular and predictable, this situation may not be problematic, since the DTD may require no serious modifications, and will become stable quickly. However, for large, complex, and expanding collections of documents (that is, for most scholarly editions), the textual evidence will produce an interminable series of challenges to the assumptions of regularity that SGML requires.

The editor of an electronic archive can respond to such challenges in one of two ways. The first is Procrustean: he may use the existing tagging logic to describe the new documentary situation, overriding or altering the textual features to make them fit. At first, this may sound completely unacceptable, but in fact every creator of a scholarly edition in print or electronic form inevitably makes such compromises with the medium of transmission. Not everything can be preserved, because we have no way of imagining everything a text can be, much less a logical way of describing it. All editing, with the possible exception of strict facsimile reproduction, involves more or less intrusive processes of normalization and regularization. However, in the Rossetti Archive and most other electronic editions, the user may compare the tagged text to digital facsimile images of the page in question, in order to resolve small questions of rendering. For example, our DTD does not allow us to specify precisely where on the page a given bibliographic signature falls; and as we discovered, the locations of these do vary among documents. But since this information can be ascertained by calling up the facsimiles, we have deemed such granularity in our tagging unnecessary. The second response to challenges to one's rules of description is simply to alter the DTD to account for the new situation: in this case, the tail wags the dog. One instance of an unforeseen textual feature causes a change in the basic structuring logic of the archive—a change that can have implications for many of the other documents in the archive. The result is continual updating of files as the DTD changes. For example, in some of the proofs for his 1870 *Poems*, Rossetti had single revise pages struck off, which he then inserted into the sequence of the documents. In order to account for these pages as such, we introduced a new tag (`<revise>`) within the last year, and brought all of the affected files up to date. The change was an important one for those wishing to trace the development of Rossetti's 1870 volume, and was certainly worth making.

Even the simplest of alterations in thinking about the way a text should be described can have daunting implications. In the early stages of the project, we tagged manuscript corrections in the following manner: if Rossetti struck out a word, phrase, or line and wrote in a substitution, we marked the old material with `<del>` tags (indicating a deletion) and let the new reading stand. For example, in one of the proofs for the 1870 *Poems*, Rossetti deleted the word "course" in line 133 of "The Blessed Damozel"; and substituted the word "paths". Under the old dispensation, this was tagged

```
<lg n="23" r="24" type="sexain">
<l n="133" r="139">( I saw her smile.)  But soon their paths
<l n="133a" indent="4">                           <del>course</del>
<l n="134" r="140" indent="1">     Was vague in distant spheres:
```

Within the last year however, we realized that this system was inadequate because it did not make clear which word was added by hand. Put another way, it occludes the distinction between the printed words of the proof page and Rossetti's autograph script. We decided that `<add>` tags needed to be placed around all such additions, so that the text in question is now tagged like this:

```
<lg n="23" r="24" type="sexain">
<l n="133" r="139">(I saw her smile.)  But soon their <add>paths</add>
<l n="133a" indent="4">                                    <del>course</del>
<l n="134" r="140" indent="1">     Was vague in distant spheres:
```

A simple change—but one that meant we had to re-examine and alter every document containing manuscript alterations by Rossetti (an inveterate reviser), comparing page images to the SGML markup and making the required changes. Episodes like this one can consume many hours and seem to be the thousand natural shocks that electronic editions are heir to.

Some of the constant revisiting of files that characterizes such projects can be attributed to the necessarily collaborative environments in which they are executed. Over the course of the edition, tagging assistants (often graduate students) come and go, and as a result, many hands and minds have shaped the myriad files in the archive. Working frequently alone and making decisions constantly, each person executes the minute particulars of tagging slightly differently, even if they all are conforming to the same DTD. Because of this, local variation is commonplace and this can come to have disruptive consequences, particularly if the DTD changes. Developing a comprehensive style sheet will reduce such heterogeneity, but keeping this up-to-date and staying true to it could quickly paralyze the project. The potential immensity of electronic archives makes strict regularity among the files something to be longed for, but perhaps never seen.

In the world of SGML, the best way to limit unexpected compromises or alterations is to invest heavily up front in the research and design of a robust DTD appropriate to the needs of one's project. As we complete the first installment of the Rossetti Archive, we have arrived at a richly-textured and powerful DTD with which to describe and render texts. Yet we came to this by way of some paths that later editors and builders of electronic archives should not have to tread. In this way, the Rossetti Archive has always been a kind of research laboratory, where revisions and refinements occur out of a directed process of trial and error. If we are going to design logical systems of description for works of art, we should accept that art will always find ways to violate those systems.

Several shortcuts can aid the large-scale adaptation of files to a newly-revised system of tagging. Computers are very good at searching for patterns and replacing them with other patterns. Therefore, as long as the changes you need to make are uniform, little problem exists. Several keystrokes in most file-editing programs, such as EMACS or its JOVE subset, will perform such changes in a particular file in seconds. For large groups of files, where such repeated opening and processing would be time-consuming, we have recently developed a series of scripts that will make global alterations, such as changing misconceived work codes to their proper forms, and placing quotation marks around all line numbers. Such scripts make certain that no instances are missed, and can enact the changes instantly. However, you have to be sure that global replacements will not garble or steamroll the text, and frequently the kinds of updating required cannot be performed automatically. Finally, an archive might make use of entity references that call up a separate datafile that can be altered as needed. Let us say, for example, that you have attribution information from a number of libraries and museums that will need to be included in many files. If you begin by simply pasting data—the name of the museum, the director's name, a permission statement—into each file separately, you will have to perform similar cutting and pasting if you decide to include more or very different information later. However, if you set up an entity reference for each museum or library, you would need only to alter the file to which it referred, should any changes need to be made. For example, if you want to present the permissions data from the British Library, you insert the string "&BritLib;" into the marked-up text, which tells the computer to call up the "BritLib" file and insert it into the text at that point. Thus changes made to the "BritLib" file will be replicated automatically in every file.

None of these solutions are particularly helpful when alterations are based on a reconceptualization of textual relations or rendering, which typically require thorough re-examination of all files and involve new decisions, new observations, and new tags. It may be that second or third generation projects of this kind will be able to avoid many of the revisions that the Rossetti Archive has experienced. But ultimately, it may be that we should consider inventing grammars of description more forgiving and flexible than SGML. Alternatively, we might acknowledge that SGML cannot easily preserve all the kinds of information that scholarly editions have valued throughout their bookish history. Such an admission might well free us to create specific electronic toolsets and databases whose radical functionality would exempt them from being 'editions' as such. While it certainly will involve electronic frameworks and digitalization, the future of scholarly editions may come to depend on protocols other than those of SGML.

2. Rendering Conceptual Units: The Need for Double-Tagging

For the textual editor, another major problem with SGML involves the competing claims of formal structures and typographic rendering. As mentioned above, in order for a computer to display an SGML document, it must be told how to render the text marked by a given tag. For example, should it indent at every paragraph tag? Should it place a set of quotations marks at every quote tag? Should all titles of books tagged as such be rendered in italics? SGML is designed to operate most efficiently if the answer to all of these questions is yes. However, anyone who has worked on a scholarly edition will recognize the problem with such assumptions of regularity. Sets of documents rarely conform to such rigid schema, particularly when they include texts drawn from books, periodicals, and manuscripts from across many decades.

At the Rossetti Archive, we tabled this issue during the initial stages of the project, under two working assumptions: one, that the yet-to-be-created browser which would process our SGML files would somehow sort out questions of rendering, and two, that the facsimile images would provide ample witness to typographic specifics. Only recently, as we have approached the conclusion of the first installment, have we recognized the need to detach rendering from some of our conceptual tags, specifically `<quote>`, `<title>`, and `<foreign>`. For example, an italicized appearance of the title of Rossetti's famous sonnet sequence would be tagged as follows:

```
<hi rend="I"><xref doc="22-1869.raw"><title level="wrk">The House of
Life</title></xref></hi>
```

whereas the same title rendered in plain-text inside quotation marks would be tagged

```
&ldquor;<xref doc="22-169.raw"><title level="wrk">The House of
Life</title></xref>&rdquor;
```

The `<xref>` tags mark the text as a link to another file, in this case to the top-level document called a RAW (Rossetti Archive Work), which introduces and itself contains links to all documentary instances of that work. In the first instance, the `<hi rend="i">` tag indicates that the enclosed text is rendered in italics, whereas in the second, the "`&ldquor;`" and "`&rdquor;`" are entity references that represent left and right double quotation marks. Note that in this case, the rendering information is nested outside of the title tags, because the title itself does not include any particular rendering.

In earlier versions of our files, the <quote>, <title>, and <foreign> tags carried with them assumptions of typographic alteration. Currently, we use them as strictly conceptual markers, and include other tags (indicating italics, for example) and/or characters (such as single or double quotation marks) to accomplish the idiosyncratic rendering of the text. Clearly, our desire to describe texts both physically and logically cannot be completely accommodated by a single set of SGML tags to which rendering decisions are mapped. In many cases, double-tagging is necessary to achieve the level of specificity that the scholarly edition requires.

3. The Question of White Space

SGML is good for describing strings of characters (indicating how they function in the document and how they appear on the page), but it can only with difficulty be used to provide information regarding white space. The reasons for this seem to be twofold. First, the editorial enterprise is bound by a bookish history that SGML was designed to transcend. The primary form of displaying an SGML file is on a monitor, rather than a series of pages, so that the page itself becomes a kind of ghost in the machine, a remnant of the previous technology that the current one can't quite grasp. For example, the SGML display will be necessarily affected by various screen sizes and browser window sizes that make the formatting of the electronic archive a more fluid process than it was in the page-bound documents the archive catalogues. Second, as students of literary texts, we have only begun to develop effective ways of talking about the meanings of gaps and white spaces in literary texts. This aporia in our critical vocabulary has left us unable to imagine how such spaces might be usefully and faithfully described.

In a standard SGML document, the computer will not recognize any more than two spaces between characters without the insertion of a tag. As a result, either every space beyond the first two must be tagged, or the rendering can be approximated and reliance for greater specificity must again be placed on the facsimile images. In the Rossetti Archive, we have opted for the latter path in regard to indentation. Thus, a typical stanza, this one from a poem by Rossetti entitled "The Stream's Secret", is marked as follows:

```
<lg n="7" type="sexain">
<l n="37" indent="2">Stream, when this silver thread
<l n="38" indent="1">In flood-time is a torrent brown,
<l n="39">May any bulwark bind thy foaming crown?
<l n="40" indent="1">Shall not the waters surge and spread
<l n="41">And to the crannied boulders of their bed
<l n="42" indent="2">Still shoot the dead leaves down?
```

Here, the lines are numbered consecutively, and the various numbers within the `<indent>` tags are used to indicate relative degrees of indentation, rather than specific quantities of white space. The browser will be told to indent the lines in proportion to the numbers in the tags, with "1" being the smallest indentation and "4" being the largest. Furthermore, these relational indentations are not reset for each poem, so that a line with an indentation of "1" in "The Stream's Secret" will be rendered flush with a line indented "1" in another poem, although they may not be indented an equal number of absolute spaces. The resulting SGML text will contain an approximation of the document's spacing, which can be checked against the page image for greater specificity. This system is admittedly a compromise, arrived at as a way to preserve relational spacing without having to count every space in every document; and later versions of the Archive may well evolve greater levels of fidelity in this regard.

4. Searching Mechanisms

Every one of Rossetti's discrete works, whether written or pictorial, has its own unique work code, so that a user interested in finding all appearances of or references to that work can do so. In a sense, the deployment of these work codes within the SGML files establishes the central nervous system of the hypertextual archive. At the most basic level, a volume of Rossetti's poems exists as a discrete file with a work code, and each poem (and/or division) within that volume also has its own work code, which is provided in the header tags of that poem (and/or division). For example, the poem "Aspecta Medusa" (work code =1-1865) would be tagged as follows for its appearance in Rossetti's 1870 *Poems* (work code =1-1870):

```
<div1 type="lyric" title="Aspecta Medusa." workCode="1-1865"
n="20">

<lg n="1" type="quintain">
<l n="1"><hi rend="sc">ANDROMEDA</hi>, by Perseus saved and wed,
<l n="2">Hankered each day to see the Gorgon's head:
<l n="3">Till o'er a found he held it, bade her lean,
<l n="4">And mirrored in the wave was safely seen
<l n="5" part="I">That death she lived by.
</lg>

<lg n="2" type="quatrain">
<l n="5" indent="1" part="f">Let not thine eyes know
<l n="6">Any forbidden thing itself, although
<l n="7">It once should save as well as kill: but be
<l n="8">Its shadow upon life enough for thee.
</lg>
</div1>
```

Each new work within a document is a division, arranged hierarchically by means of enumerated <div> tags. By embedding the work code in the <div> tag as an attribute, we have allowed the computer to find this poem automatically. Users will be able to gather all available documentary instances of or references to "Aspecta Medusa" (or any work by Rossetti, despite name changes or radical revisions that might disable a string search) by asking the computer to search for its defining code.

However, as we moved beyond the primary documents and began tagging works of biography and criticism, this system revealed a simple but fundamental limitation: how do we identify quotations of Rossetti's work as such? Typically, these are not separate divisions of the document in question, and thus the <div> tag cannot be used. Particularly when they occur in the midst of paragraphs, without the mention of a title, such quotations were in danger of remaining invisible to various kinds of searches. Here again was an example of how, as we expanded the scope of the Archive, we found current practices unsuitable to a new class of documents.

Our solution in this case was an entirely new tag: <workunit>. Its purpose is to identify and describe quotations of Rossetti's work wherever they appear in the Archive. In practice, the <workunit> tag operates as follows, in a passage from William Michael Rossetti's *Family Letters and a Memoir of Dante Gabriel Rossetti*:

```
...the poem which relates his prevision of the death of Beatrice.
In my brother's translation it is rendered thus:--
<quote>
<workunit display="block" wholeness="part" workCode="11d-1861"
type="translation">
<lg n="1">
<l n="1" r="69">&ldquor; And with her was such very humbleness
<l n="2" r="70">That she appeared to say, I am at peace.&rdquor;
</lg>
</workunit>
</quote>
```

Here William Michael quotes two lines from his brother's translation of a poem within the *Vita Nuova*. In addition to the work code and type of quoted material, the attributes of <workunit> allow us to specify whether this is a block or an in-line quotation and whether it quotes the work in full or in part. Thus the <workunit> tag becomes an identifying marker of any quotation of a work by Rossetti, providing for its inclusion in searches for instances of that work, despite its incomplete, untitled, and unpredictable location within another document. Note, too, that the place of the quoted lines within the poem as a whole is preserved through the use of the *r* attribute in the line tag. While the *n* value represents the actual number of the line within the line group on the page, the *r* value refers to the line numbers (here, 69 and 70) of an arbitrarily-selected version, or reference text, of the poem, to which all other instances are compared for purposes of collation. This feature will fully automate the often difficult process of collating and comparing various documentary instances of a given line or set of lines. All versions of lines 69–70 in the Archive can be compared, no matter what their location or numerical value within a given document.

5. Conclusion

Such examples of the adaptive nature of the Archive's methodologies and practices could easily be multiplied. In fact, the Archive has moved forward through a continual series of confrontations between our methodologies and our praxis, between design and implementation—ultimately, between our assumptions of regularity and the intractable variability of documents. Because we have chosen to design our own system of SGML, theoretically independent of any specific hardware or software platform, we have been bound at the highest level only by the rules of set logic and the principles of bibliographical description.[4] However, our work has been continually shaped by the minute particulars of execution and application. Certainly building this kind of scholarly engine requires one to imagine the questions it will be able to process, and the ways that SGML can make those processes produce accurate information. But the real work of developing the Rossetti Archive and projects like it occurs in the repeated reconfiguring of those imaginings as we enact their implications.

Bibliography

BENDER (T. K.): 1976, "Literary Texts in Electronic Storage: The Editorial Potential", *Computers and the Humanities*, 10, p. 193–99.

DEEGAN (M.) and SUTHERLAND (K.): 1998, eds. *Electronic Text: Investigations in Method and Theory* (Oxford: Oxford University Press; forthcoming).

ERDMAN (D. V.): 1970, ed. *The Poetry and Prose of William Blake* (New York: Doubleday).

FINNERAN (R. J.): 1996, ed. *The Literary Text in the Digital Age* (Ann Arbor: Univ. of Michigan Press).

LANDOW (G.) and DELANY (P.): 1993, eds. *The Digital Word: Text-Based Computing in the Humanities* (Cambridge: MIT Press).

LAVAGNINO (J.): 1995: "Reading, Scholarship, and Hypertext Editions", *Text*, 8, p. 109–24.

McGANN (J. J.): 1994: "The Complete Writings and Pictures of Dante Gabriel Rossetti: A Hypermedia Research Archive", *Text* 7, p. 95–105.

McGANN (J. J.): 1996, "The Rationale of Hypertext", *Text*, 9, p. 11–32. Also available at URL: http://jefferson.village.virginia.edu/public/jjm2f/rationale.html.

McGANN (J. J.): 1996, "The Rossetti Archive and Image-Based Electronic Editing", in FINNERAN, p. 145–183.

McGANN (J. J.): 1997, "The Rossetti Hypermedia Archive: An Introduction", *The Journal of Pre-Raphaelite Studies*, N.S. 6, p. 22–32.

SHILLINGSBURG (P.): 1986, *Scholarly Editing in the Computer Age: Theory and Practice*, rev. ed. (Athens: Univ. of Georgia Press).

[4] Of course, certain computer protocols and technical realities have determined some of the contours of the project, just as the availability of Rossetti documents (and publication permissions) have set certain limits to its current contents.