Conceptual Levels of SGML Tags:
a proposed taxonomy based on the tagging in the Orlando Project

Stan Ruecker
Orlando Project
University of Alberta, 3-5 Humanities Centre
Edmonton, Alberta CANADA T6G 2E5
ph: 780-908-9575 • e-mail: sruecker@ualberta.ca

Abstract

Several projects in various disciplines are now using Standardized General Markup Language (SGML) tags at an interpretive level. That is, these projects contain tags which have the potential to provide the reader with additional information that is not already explicit in the text itself. One such interpretive project is the Orlando Project, which is an integrated history of women's writing in the British Isles, currently under development in Canada. Orlando is unlike other projects in that the content is being written and tagged simultaneously. It also contains a wide and rich variety of both descriptive and interpretive tags, which provide the user with a wealth of information on women's writing in the British Isles. But the project does not currently provide an explicit indication of the level of description or interpretation to be expected in any given tag. Without such a taxonomy, projects like Orlando risk introducing potential ambiguities for the scholarly user. This paper therefore proposes a potential conceptual tag taxonomy for literary interpretive SGML projects such as Orlando.

Introduction

Standardized General Markup Language (SGML) is a meta-language, or grammar, for creating text markup systems [19]. It allows its users to define custom sets of Document Type Definitions (DTDs), which are used to structure plain text documents so that a given output device or platform can interpret the structure for formatting (and other) purposes. SGML allows document authors to create and manage their material in a manner relatively free from considerations of where that material will ultimately be displayed, manipulated, or printed.[1]

For projects that include poetry, for example [23], [7], it is useful to be able to specify the poem's title, line breaks, and even line counts, so that any reader will see the poems displayed properly on the screen or page, regardless of whether they are accessing the data on a UNIX, LINUX, Windows, or Macintosh computer. The SGML tags delineate the parts of the poem without regard for platform, but the platform eventually plays a role in determining how each of those parts should be displayed or printed. A Macintosh, for example, might use a default screen font for titles of 12 point Palatino italic; the same title might appear in a Windows machine as 14 point Arial bold, and on a UNIX computer as 10 point Geneva oblique.

SGML, however, has the power to do more than facilitate platform-independent formatting. Because SGML allows the user to define a custom tag set, it has been used in some projects to create tags that help facilitate concept-level searches. In some cases, the tags are used to group common content independent of wording; in others, the tags add information not originally in the content; and finally, tags have been used to provide interpretive material. These three kinds of concept-level tags add complexity to the projects that use them. Interpretation is often not required in deciding what constitutes a poem, or a line from a poem, or a paragraph of prose (although in certain cases even these boundaries are in dispute); but there are several different kinds of interpretation implied by the use of concept-level tags. When does a piece of fictional writing, for example, constitute the use of a literary motif? What is the "normal" name of a geographical location that has had many different names at different historical periods? When is a health problem physical and when is it mental? Each of these questions can be understood as requiring, during its answer, a different kind of interpretive act, and each of the interpretive acts may imply the need for a different kind of interpretive tag.

[1] Other approaches to this problem include eXtensible Markup Language (XML), which can be considered as a subset of the grammar of SGML, and HyperText Markup Language (HTML), which is (at least in its earlier basic forms) a DTD. Because of the similarity between XML and SGML, further references in this paper to SGML can be understood as referring to both grammars.

The Need for an SGML Tag Taxonomy

A review of existing SGML projects in the humanities, law, and medicine suggests that the principle of enhancing archives through custom DTDs is now being applied in a variety of different fields [16], [4], [20]. There are also indications that with the commercial advent of XML there will be an even greater proliferation of similar projects [5], [8]. The advantage of SGML over HTML is that projects using SGML can define their own tagsets and create tags appropriate for their data; in spite of the popularity of e-commerce, for example, HTML has no tag for price. Implementing a <PRICE> tag in SGML, on the other hand, can be a straightforward part of setting up a DTD.

Although there has been some discussion of SGML tag categories in the literature, the primary distinctions seem to have been made between different kinds of descriptive tags, with various levels of formatting information but only one category for interpretive tagging [3]. What the SGML community appears to have not yet formalized are clear distinctions among different conceptual levels of tagging. A conceptual tag taxonomy would consist of a hierarchy or continuum of formal definitions of the kind of data to be expected in each given group of tags.

Such a taxonomy would provide information about the classes of data a user might expect to find within a particular kind of tag. This classification information would be additional information, not currently included in tag definitions. It would help distinguish between tags that are present for reasons of platform-independent formatting (e.g., <P>this is a paragraph</P>), tags that are present for the purposes of document description (e.g., <AUTHORITY>the Orlando Project</AUTHORITY>), and tags that support scholarly reading (e.g., <HEALTH ISSUE="PHYSICAL">Spring 1816, Jane Austen first mentioned feeling ill; that summer, aged forty, she already suspected she might not recover.</HEALTH ISSUE>) [17].

The explicit implementation of such a taxonomy would be valuable both to taggers and to subsequent users, since best-practice issues could then be addressed within the framework of defining the appropriate conceptual level for each tag, which would, in turn, help address the problem of users not knowing implicitly how a given tag was implemented by the document taggers. Although not every user of an interpretive tagging project would necessarily need the kind of information provided in a taxonomy, for scholarly users relying on automatic retrieval for academic research, the extra precision made possible through such a taxonomy would be invaluable.

Assume, for example, that a project DTD includes the tag <NAME>, and that the <NAME> tag is meant to record all personal names. This is a common tag in literary interpretive tagging projects such as [12], [16] and [6]. What both the taggers and scholarly users need to know is whether a name in the text should normally be tagged or not. A text might contain several different kinds of name, such as proper names ("David Jones"), titles ("Sir David Jones"), unambiguous oblique references ("the eldest daughter of Thomas Jones of Cardiff"), and potentially ambiguous oblique references ("the son of Thomas Jones of Cardiff" in the case where more than one son is known). Further decisions need to be made concerning whether the name tag should be used for fictitious characters ("Prince Hamlet"), historical figures of contested authenticity ("St. Christopher"), and, in some contexts, non-human entities ("St. Christopher's dog Job," or "Mother Nature"). Finally, the need for some means of normalization arises as a possible consideration for people whose names change over time (Tibetans, for example, or people who receive titles, or women who take their husband's surname at marriage).

Normally these questions are addressed for the project team members through a combination of training and tagging documentation. However, the issue remains as to the best methods for a) ensuring that various taggers working over time follow the guidelines with reasonable consistency and, more importantly, b) making the user aware of the project practise decisions. Ideally, the interfaces used by both the taggers and the end users should be able to extract some of the project practise decisions from the DTD itself. In cases where a taxonomic decision is the basis for best practises, including the taxonomy in the tag definitions would make this information at least potentially available for such an automated reconstruction. Issues of interface design are clearly involved, but before the interface can be addressed, it may be helpful to provide some more formal analysis of the kinds of information such a taxonomy might provide.

Proposed Tag Taxonomy

The proposed taxonomy for SGML tags contains six levels, which extend between the difference poles of unmarked text on the one hand through to tags intended to provide interpretive glosses on the other (see Table 1).

Levels 4–6 of this taxonomy constitute the proposed addition to the understanding of conceptual levels in interpretive tagging projects. The value of these levels to the scholarly user needs to be evaluated in both theoretical and practical terms.

This taxonomy was created through an informal analysis of the tags used in the Orlando Project [16], [17] at the University of Alberta and University of Guelph. Orlando is a history of women's writing in the British Isles, and one of the most conceptually complex of current literary SGML projects. It contains primary documents on over 1,200 women writers, divided across two DTDs (biography and writing) incorporating a total of 265 unique tags. There is

Table 1: Proposed Tag Taxonomy

Level	Level Name	Level Description	Examples[2] [17]
1.	raw data	unmarked text; or when text tagged in one instance is left untagged in another (c.f. “Pankhurst”)	Although she disagreed with the <NAME STANDARD=”Pankhurst, Christabel”>Pankhurst</NAME>s’ <NAME STANDARD=”Pankhurst, Emmeline”></NAME>militarism, RS acknowledged their contribution to the women’s movement, and in her preface wrote <QUOTE DIRECT=”Y”>I am exceedingly grateful to the late Mrs. Pankhurst for discussing the militant phase of the Movement with me.</QUOTE>
2.	descriptive markup	guidelines for document formatting	<P>A tale of pioneer feminism, it is about a young woman in rural Michigan who first runs away to a pioneer college for women, then participates in the pre-Civil War abolitionist movement.</P>
3.	meta tags	document description for the current document	<DOCTITLE>Jane Austen: writing</DOCTITLE>
4.	internal glosses	flag content that would be transparent to a human reader	<NAME>Susanna Centlivre</NAME>
5.	external glosses	add material not in the content itself, such as references and allusions (c.f. Tom Lefroy’s middle name)	<CHRONPROSE>Jane Austen caused comment by her public flirting with <NAME STANDARD=”Lefroy, Thomas Langlois”>Tom Lefroy</NAME>...</CHRONPROSE>
6.	hermeneutic	interpretive tags	Young Brandon behaved like a <TCHARACTERTYPEROLE CHARACTERTYPE=”HERO”>man of sensibility</TCHARACTERTYPEROLE> in loving a woman who <SOCALLED>fell</SOCALLED>...

also a wide range of corollary data (approximately 14,000 items) on male writers and historical context, much of which has been tagged according to the third project DTD – events.

Each Orlando document is written and tagged by the same member of the project team, although subsequent quality assurance phases ensure that several people review the work. According to the project’s internal writing guidelines, the primary research priority is “to incorporate whatever has been outstanding and interesting about the particular woman writer” [17]. Informally, this instruction translates into a policy of tagging any content that has “retrieval value”; that is, if a piece of information would seem to be of interest to a scholarly user, then it should be tagged.

The Orlando project can therefore be understood as containing material at all levels of tagging. However, because the concept of taxonomic ordering is not inherent in the SGML standard, the Orlando DTDs are not defined in such a way as to coincide precisely with the current proposed taxonomy. Nonetheless, it is possible in many cases to distinguish a given tag’s conceptual level through close reading of the project best-practise guidelines, which reside in the project glossary [17]. Unless otherwise indicated, the following examples are taken from the Orlando project; in some cases, tags not germane to the immediate discussion have been removed to make the passages easier to read.

Levels 1, 2, and 3 – standard tags

The first three levels of the taxonomy include the kinds of tags familiar to most users of tagging systems, whether SGML, XML, or HTML. These tags provide formatting and document catalog information. As a general rule, contemporary SGML projects only define tags to the third level: meta tagging. This generalization holds for the majority of humanities computing projects, including all those implementing versions of the standard set out by the Text Encoding Initiative (TEI). This standard is intended to provide descriptive markup and meta tagging to help make humanities archives platform independent [21]. Examples of projects based on the TEI DTD are Jerome McGann’s Rossetti Archive [13], the Middle English Collection at the University of Michigan [15], and many projects out of the University of Virginia, including archives for Dickinson [7], Whitman [23], and Blake [1].

In the legal domain, additions to the TEI made by the University of Stockholm’s Corpus Legis project [4] serve to incorporate legal apparatus into the DTD, which, in the case of indicating jurisdiction (with the <JURIS> tag), might be considered as extending the tagset into external glossing (assuming there are records where the tag’s

[2] Some tags have been removed in order to simplify the examples.

attribute rather than its contents contain the actual data). The medical community has been developing another system, the HL7 Patient Record Architecture, including DTDs (albeit in XML) which might be seen as incorporating external glosses, with a variety of tags containing patient information [8]. Early adopters of this architecture include the Mayo Clinic [11] and the SynEX system at University College, London [20].

1. raw data

The first level is content only, or untagged text. Since SGML markup can encompass everything from single words to entire documents, there is a sense in which no material within a tagged document can be considered unmarked. However, in practical terms, there are many cases in which the content under question might require tagging in one case but not in another.

If "raw data" indicates content that is potentially conceptually distinct from the material grouped with it, but is nonetheless unmarked, then there are two kinds of raw data in the Orlando project. The first is content that was not considered interesting or important enough to be tagged at a given point, although similar content might be interesting or important someplace else. For example, geographical places in Orlando are tagged only if they might have retrieval value. Deciding whether or not to tag a place name in Orlando is in this respect always a judgment call. For example, "London" is tagged as a <PLACE> in Brigid Brophy's biography because she contributed to an art show there in 1969, but "France" is not tagged, since it was where her mother served in the second World War, before Brigid was born.

The second kind of raw data in Orlando is prose that is sufficiently marked simply as being part of a larger unit, such as a chronology structure, or <CHRONSTRUCT>, but does not require more detailed internal tagging. In the discussion of a visit to Jane Austen by her cousin Eliza Hancock de Feuillide, for example, where the visit may have influenced Austen's epistolary novella *Lady Susan*, the following passage occurs, without any additional tagging: "Eliza was nothing like the scheming and ruthless Lady Susan; but she was a woman of the world who may have made her relations appear somewhat humdrum." Such material may remain untagged in Orlando for any number of reasons, including the judgment of the tagger as to its not having sufficient retrieval value, the lack of appropriate tags in the DTD, and the fact that the visit may be considered as sufficiently marked by the <NAME> tag on Eliza Hancock de Feuillide in the previous sentence.

2. descriptive markup

Descriptive markup provides the necessary information for platform-independent formatting: "this is a paragraph"; or, "this should be given strong emphasis, such as bold type." Early papers on SGML conceptualized descriptive markup in terms of an ordered hierarchy of content objects (OHCO) [18]. An OHCO would indicate such items as book, chapter heading, section heading, paragraph, line – basically, the divisions familiar to anyone who has written technical documents, with some additional levels of granularity included to cover material like plays or poetry (not used by Orlando). One problem with OHCOs is that not all textual material is hierarchical: tables, calendars, revisions, and extra-textual markings all pose difficulties, some of which have been considered grounds for developing entirely new approaches to markup (as in the MECS system designed for encoding Wittgenstein [24]).[3]

The Orlando DTDs include descriptive markup, consisting primarily of two levels of subheading and paragraph tags. A number of OHCO implications are also intended to be drawn from tag nesting. For example, the Orlando <CHRONSTRUCT> tag forms the basic unit of much material, since it occurs in all three of the project's DTDs, and establishes a section of prose that could be isolated as part of a chronology. According to the Orlando tag documentation, a <CHRONSTRUCT> can contain a number of optional items but always requires a <DATE>, <CHRONPROSE> (the sentences or paragraphs), and <BIBCIT> (or reference).

One way of presenting the Orlando data to a user, therefore, involves retrieving only the material tagged with <CHRONSTRUCT>s. The <CHRONSTRUCT>s and their contents are then organized chronologically by the values in the <DATE> tags in order to form a chronological history for a given writer or set of writers. It is also usual (but not required) for a <CHRONSTRUCT> to be followed by a supplementary paragraph or two of related but less central prose, which is contained in a <SHORTPROSE> tag. The <CHRONSTRUCT> and <SHORTPROSE> tags are always both contained in a required <CHRONEVENT>. An expanded chronology could therefore be generated by retrieving all <CHRONEVENT>s, which would show the user a chronology formed, not only of the <CHRONSTRUCT> material, but also of the supplementary <SHORTPROSE> paragraphs.

In one sense, then, the tags <CHRONEVENT>, <CHRONSTRUCT>, and <SHORTPROSE> can be considered a form of descriptive markup, indicating that the prose contained within them belongs together in a display.

3. meta tags

Meta tags include material appropriate to what, in a printed volume, would be title page matter: author, title, language, or "this is a poem." Meta tags and what should go into them is a primary concern of the library community,

[3] Wittgenstein was a philosopher famous for extensive multiple revisions of his work.

in which decades of research and debate inform many current international document description standards [10], [22]. There have, however, been difficulties both in interdisciplinary dialogue and in translating some of these traditional concerns into systems appropriate for electronic material, where the rapidly moving targets of both digital and physical format have rendered issues of container definition less significant than has been the case with printed matter. Cataloguing a particular book, for example, often involves describing the physical printed volume. Cataloguing a digital document, on the other hand, usually does not require mentioning the kind of media on which it has been stored.

The Orlando meta tags are based on the TEI header, and include indications of which sort of document follows (biography, writing, or event), a file description containing title and the authority responsible for the existence of the document (which, in this case, is always the Orlando Project), and an indication of the current state of the document in terms of its revision history, with team members identified by their contribution to the work on each particular document. Given an interface that allowed it, a scholarly user might, therefore, be able to reconstruct the history of a particular document's creation and editing with reference to the actual people who did the work.

Levels 4, 5, and 6 – Scholarly glossing tags

The majority of SGML projects use DTDs that can be considered as stopping at the level of meta tagging (level 3) – only document header information and descriptive tagging is available. There are a few instances, however, of projects that have gone beyond the meta tag into the three remaining levels of interpretive markup. Most TEI projects tag titles, for example, and the Brown Women Writers Project also tags names [2]. More deeply-interpretive projects include the Dante archive at the University of Virginia [6], the medieval Danish History project (where the encoding is intended primarily for index generation) [14], and the Orlando Project [16].

A similarly complex example of textual markup can be found in Willard McCarty's "Analytical Onomasticon of Ovid's *Metamorphoses*" at the University of Toronto [12], which set out to tag the *Metamorphoses* for primary themes such as sex, death, and names. This Ovid material is distinct from other projects mentioned in this paper, in that its markup was done using the TACT system (a combination markup scheme and database) rather than SGML. The project raises similar issues, however, in that the scholarly user of McCarty's "Analytical Onomasticon" has to be made aware of the kinds of decisions implemented by the project taggers. In addition to these literary projects, both legal and medical DTDs exist that may extend as far as external tagging (level 5) [20], [4].

4. internal glosses

Internal glosses provide information that would normally be available to a human reader but needs to be explicitly tagged for automatic processing. There is, therefore, a sense in which nearly every tag in a given project's DTD can be understood as an internal gloss, since two of the standard uses of SGML are to provide the information necessary for automatic document formatting and to group material of a similar kind for subsequent search and display. For the purposes of this paper's proposed taxonomy, however, any tag might be considered an internal gloss if it meets three conditions: it does not primarily provide formatting instructions; it does not add information not already in the text; nor does it comprise an interpretive act in its own right.

The Orlando <BIRTH> and <DEATH> tags are examples of internal glossing, where either a date range or explicit prose discussing these events is already available to readers but not in a form that would necessarily be sufficiently consistent for retrieval through string searching. Dates, for example, are occasionally given in prose form, while the VALUE attribute of the <DATE> tag provides the same designation in a form that can be numerically sorted or searched:

> In <DATE VALUE="1964-04-">April the same year</DATE> Brophy published <TITLE TITLETYPE= ”MONOGRAPHIC”>The Waste-Disposal Unit </TITLE>.

5. external glosses

External glosses add references and allusions that are not already present in the text proper. Such additional information might consist, for example, of details of the kind of school a writer attended – was it co-educational or single-sex, was it primary, secondary, or post-secondary, was it public or private? This information would not normally be available where the prose only gave the name of the institution. The content of an external glossing tag resembles the information often found in the textual apparatus of scholarly editions and, as in scholarly editions, is highly dependent on decisions as to the sophistication and needs of a proposed audience.

Many of Orlando's tags can be understood as external glosses, since the project's tagging goals include reference to a wide range of historical places, institutions, documents, and people. Whenever this material is not present in the text, it is provided – invisibly, as far as the reader of the content is concerned – within the attributes of the tag. For example, one project priority is geographical placement for a given writer's work. One tag that helps mark geographical placement is the Orlando <PLACE> tag, which is based on the TEI guidelines. The <PLACE> tag is the outer shell of a nested set and can contain sub-tags <PLACENAME>,

<ADDRESS>, <SETTLEMENT>, <REGION>, and <GEOG>. The difference between <SETTLEMENT> and <GEOG> is that the former identifies a a town or village, while the latter specifies a country. These tags are used so that prose which only partially identifies a location can be still be retrieved by people searching for the larger geographical context.

A typical set of <PLACE> tags might be assigned as follows:

> Before her multiple sclerosis prohibited travel,
> Brigid Brophy enjoyed becoming familiar with
> <PLACE> <GEOG>Italy</GEOG></PLACE>. She
> also visited <PLACE><SETTLEMENT>Samarkand
> </SETTLEMENT><REGION reg="Uzbekistan">
> <GEOG current="Russia"></GEOG></PLACE>
> (in a blizzard).

The text for this passage (content only, without tags) would appear as follows:

> Before her multiple sclerosis prohibited travel, Brigid Brophy enjoyed becoming familiar with Italy. She also visited Samarkand (in a blizzard).

In this example, the <PLACE> and <GEOG> tags around "Italy" are internal glosses, since they only mark for the benefit of the computer everything a human reader would already know; namely, that Italy is a geographical place. This marking allows researchers to retrieve all the British women writers mentioned in the Orlando project who travelled in Italy. The tags on Samarkand, however, are external glossing tags, since there are several additional pieces of information provided which are not found in the actual text; namely, that Samarkand is in Uzbekistan and that Uzbekistan is in the former Soviet Union.

It is possible that the <REGION> tag could be used hermeneutically, in cases where a somewhat vaguely-defined region is specified. An example might be <REGION>the Thames Valley</REGION> as opposed to the internal gloss <REGION>Berkshire</REGION>.

6. hermeneutic tags

In order to be considered hermeneutic, a tag must introduce an interpretation not already explicit in the content it marks. A tagger using a hermeneutic-level tag makes a choice from a number of possible options. These options might take the form either of the choice of attribute values on the tag, or else of the inclusion of material (as in an external gloss). The difference between an external gloss and a hermeneutic tag is that, for the latter, any additional material would clearly be indicated as an interpretive judgment rather than, as in the former, a statement of widely-accepted and readily-verifiable fact. The Orlando DTDs include hermeneutic tags of the first kind (with the interpretation indicated by the choice of attributes), but not of the second kind (where information not in the content is added to the tag at an interpretive level).

An example of an Orlando tag which indicates an interpretation through its attributes is <HEALTH>, which contains the optional attribute ISSUE. Two of the values permitted by the DTD for the attribute ISSUE are "mental," and "physical." In many cases, however, it is impossible to distinguish, especially at some historical distance, whether a given health issue was one or the other. As with any hermeneutic tag, <HEALTH> is in this respect a judgment call. An example occurs in the biography of Brigid Brophy:

<HEALTH ISSUE="PHYSICAL">In November 1982, BB fell in the street and suffered a blow on the head which knocked her unconscious and had to be treated in hospital. She fell when she ran across a street. Neither she nor her friend <NAME>Elizabeth Jane Howard</NAME>, who was with her, could account for the accident or for its severity.</HEALTH>

In this case the attribute is "PHYSICAL," which is based on information later in the biography which associates the fall with an early stage of Brophy's as-yet-undiagnosed multiple sclerosis.

A third kind of hermeneutic tag performs an interpretive act simply through being applied. In this last class is the Orlando <SOCALLED> tag, which serves to indicate that the tagger is not willing to let a content statement stand without question. An example occurs in the biography of Elizabeth Barrett Browning, where a term used in the past to designate mixed ancestry is now widely interpreted as racist: "During his youth in England, her father came into frequent contact with one set of <SOCALLED>quadroon</SOCALLED> first cousins."

Many of the tags defined in the Orlando project can be regarded as hermeneutic, including both those that are unequivocally interpretive by intention, and tags that may not originally have been conceived as hermeneutic but are which nonetheless arguably so. In the former category are tags such as textual motif (<TMOTIF>), where a literary critical understanding of the material is necessary to make this judgement useful and appropriate. The Orlando <TMOTIF> tag is an example of the kind of hermeneutic tag that provides interpretation simply through its application rather than through a choice of attributes.

For example, Jane Austen's writing document contains the following note on her novel *Sense and Sensibility*:

<TITLE>Sense and Sensibility</TITLE> uses, reflects in its title, and radically alters, the <TMOTIF>paired-heroines</TMOTIF> motif made popular by <NAME>Jane West</NAME> and others.

Although there is no question that *Sense and Sensibility* is a novel with two heroines, the identification of such a configuration as a literary motif constitutes an act of literary critical interpretation which, in this case, is further supported by the content of the sentence in which the tag appears.

Another such category is <TINTERTEXTUALITY> and its subtags <TGENRENAME> and <TTHEMETOPIC>, where the implied or explicit relationship of the writing in question to another piece of written material is marked by the tagger, based on whatever evidence is available. Such evidence might not be included in the tag's attributes or in the document content itself but may exist in one of the various source documents consulted by the tagger. In practise, however, the prose marked with <TINTERTEXTUALITY> on the Orlando project, like the prose marked with <TMOTIF>, is usually sufficient to give the reader at least a basis for the judgment. An example of <TINTERTEXTUALITY> and its subtags <TGENRENAME> and <TTHEMETOPIC> appears in the writing document for Jane Austen:

> <TINTERTEXTUALITY INTERTEXT= "PARODY">Jane Austen's early fictions sizzle with outrageous <TGENRENAME> satire</TGENRENAME> on inter-related topics: the popular <TTHEME-TOPIC>fictional modes of the day</TTHEME-TOPIC>, <TTHEMETOPIC> gender roles</TTHEMETOPIC>, and expectations of <TTHEME-TOPIC>feminine decorum</TTHEME-TOPIC>. Most of the novelists mocked (like <NAME>Eliza Nugent Bromley</NAME>) are individually unknown today. <TVOICENARRATION>These burlesques include stories told in the first person, third person, and in <TGENRENAME GENREREG="EPI-STOLARY"> letters</TGENRENAME> </TINTER-TEXTUALITY>.

What the tags add, in this case, is the experienced reading of an academic. Anyone reading Jane Austen's early fictions may be able to see that she sometimes uses fictional letters to tell a story. That this form of narration characterizes a literary genre called "epistolary," on the other hand, is a piece of information that is not self-evident and that constitutes part of an expert knowledge of literature. It is this expert knowledge which is potentially made availabe to any of Orlando's users through the attributes of the hermeneutic <TGENRENAME> tag. At a more sophisticated level, the <TGENRENAME> tag also serves to allow scholarly users to search and associate genres in a variety of contexts, including the intertextual and chronological.

Some Orlando tags that are arguably hermeneutic include <RACEANDETHNICITY>, <CLASS>, <RELIGION>, and <INTIMATERELATIONSHIPS>. The Orlando <INTIMATERELATIONSHIPS> tag has been widely used on the project in an attempt to counterbalance the tendency of traditional biography to ignore matters relating to the personal involvements of women writers. Although it might be argued that <INTIMATERELATIONSHIPS> is primarily an internal gloss, there are cases where one of the optional attributes has been used to add information which might be considered interpretive, given that information on intimate relationships in the past is not always accurate or even easy to locate. A possible example of <INTIMATERELATIONSHIPS> as a hermeneutic tag occurs in the biography document for Valentine Ackland:

> <INTIMATERELATIONSHIPS EROTIC= "EROTIC-YES"> Valentine Ackland visited Italy with her parents and fell in love for the first time in her life, with a young woman.</INTIMATERELATIONSHIPS>

Like <INTIMATERELATIONSHIPS>, the Orlando <RACEANDETHNICITY> tag is arguably hermeneutic by definition, since issues of what constitutes race and ethnicity are currently under debate. There are cases where either this tag or its sub-tag, <RACECOLOUR>, has been used primarily as an external gloss, such as in the following passage from the biography of Phillis Wheatley:

> Phillis Wheatley was a <RACECOLOUR SELF-DEFINED="SELFYES">black African</RACE-COLOUR>, whose colour dictated all her life-experiences from the moment that she was seized by slave-traders.

However, not all uses of the </RACECOLOUR> tag include the attribute SELF-DEFINED="SELFYES," and in such instances the onus of the interpretive act lies either with the writer of the biography, or with the person assigning the tag, or both. The following example occurs in the biography of Mary Seacole, where the tag is an external gloss, although whether the interpretive decision was made by Mary Seacole or the biographer is not clear: "...she does not fail to credit her <RACECOLOUR FOREBEAR= "MOTHER">Black</RACECOLOUR> mother with having passed on a talent for herbal medicines..." The </RACECOLOUR> tag is intended to allow searchers to retrieve all writers of black heritage mentioned in Orlando; such an identification is, however, not always a straightforward statement of fact.

Another potentially hermeneutic tag is <LOCATION>, which is used to indicate where a woman was physically located at a particular period of her life. The hermeneutic thrust of <LOCATION> lies in its available attributes, which are "travel," "visited," and "lived." Selecting the appropriate attribute is inevitably a judgment call on the part of the tagger, since the distinctions between passing through briefly (travel), spending some time (visited), and staying for a prolonged period (lived) are not clearly defined and, in many cases, the details of duration may be difficult to derive from the sources available.

The complexity of choosing its attributes aside, <LOCATION> might seem straightforward enough but, in theory, the very existence of a tag for geographical location involves a number of critical assumptions concerning the importance of nation, region, and even the establishment and identification of one area as "home." These assumptions inform the entire Orlando project, which is, after all, defined as dealing with "British Women Writers." Post-colonial literary and cultural theorists are, however, currently re-examining the significance of these ideas and, as their reformulations find a wider audience, any interpretation based on geography may find itself enmeshed in controversy. In this case, a tag that may have begun as an external gloss will have migrated to hermeneutic status.

It is important to distinguish between interpretations that arise in the content of the tag (in which case the tag might only be an internal or external gloss) and interpretations that are indicated by the tag itself (in which case the tag is a hermeneutic tag). In the former class would be the following <SHORTPROSE> passage from the biography of Aphra Behn:

> <SHORTPROSE>While in Surinam, Aphra Behn was rumoured to have had either an affair or at least a flirtation with <NAME>William Scot</NAME>, son of a leading anti-monarchist politician.</SHORTPROSE>

Although there is clearly speculation involved, the speculation is on the part of Aphra Behn's contemporaries rather than the biography writer or tagger. This passage receives a hermeneutic level of tagging , however, when it is nested as follows:

> <INTIMATERELATIONSHIPS EROTIC= "EROTICPOSSIBLY"><SHORTPROSE>While in Surinam, Aphra Behn was rumoured to have had either an affair or at least a flirtation with <NAME> William Scot</NAME>, son of a leading anti-monarchist politician.</SHORTPROSE> </INTIMATERELATIONSHIPS>

Here, as in the case of the <INTIMATERELATIONSHIPS> tag on Valentine Ackland, it is the choice of attribute which signals that an interpretive act has been made by the tagger.

Knowledge derivable from a tag taxonomy

A fundamental question in creating and implementing a conceptual taxonomy concerns the degree to which a particular user or tagger might benefit from an explicit inclusion of tag level in an interpretive-level tagging project. The taxonomy's primary potential would lie in making the intended content level of a tag explicit, in order to prevent cases where a scholar using a resource like Orlando might receive a search result that is understood as being a straight internal or external gloss, when in fact it is hermeneutic.

The Orlando <NAME> tag, for instance, might be considered as potentially operating at any of three levels. A <NAME> tag might be an internal gloss insofar as it contains information only concerning the fact that a given set of words constitutes someone's name. On the other hand, <NAME> becomes an external gloss when it provides additional information that associates a person in the text with a person in the real world. An example might be the tagged phrase "Lady Mary's <NAME STANDARD= "Montagu, Edward Wortley">husband<NAME>." Finally, <NAME> could become a hermeneutic tag if its use in some way limits the reader to a particular reading or set of readings. The father of one of Susanna Centlivre's children, for example, is unknown, although Richard Savage is a possible candidate. Attributes on the <NAME> tag might therefore be used to provide the names of various people who have been suggested. A fabricated example (not from Orlando) might look like the following:

> The assassin of President Lincoln was John Wilkes Booth, but was he working alone? A number of possible <NAME CANDIDATE1="Johnson, Vice-President Andrew" CANDIDATE2="Benjamin, Judah" CANDIDATE3="Stanton, Edwin">collaboraters </NAME> have been discussed by conspiracy theorists over the years.

The numbering of the "CANDIDATE" attribute might be used here to indicate probability, in which case the additional information in the <NAME> tag is interpretive. An alternative version would allow multiple instances of the same attribute (simply "CANDIDATE," without a number), in which case the <NAME> tag would be an external gloss rather than hermeneutic.

Orlando <NAME> tags are not currently defined in such a way as to easily permit their use as hermeneutic tags, since the attributes on <NAME> do not provide for more than one name to be included at a time. Any such debate, as far as Orlando is concerned, must therefore be contained in the actual prose.

Orlando's DTD could be redefined to include information as to whether <NAME> were an internal, external, or hermeneutic level tag. The DTD might, alternatively, be redefined so as to include a different <NAME> tag for each level. In either case, taggers and users would have more information available to help disentangle the possibilities inherent in an application of the tag to a given name. A <NAME 4–INTERNAL> would therefore indicate simply that the tagged material is a name. A <NAME 5–EXTERNAL> would indicate that the tag contained the name of a person about which we have some additional information – and the user could be cued as to the availability of that information. A <NAME 6–HERMENEUTIC>, on the other hand, would suggest that the tagged name has some ambiguous quality, and indicate further investigation is required.

One disadvantage of applying a tag taxonomy to interpretive markup projects is that it provides clarification at the cost of increased overhead in terms of conceptual planning, implementation, and technical capability; that is, without an accepted industry standard, individual projects would be required to shoulder the burden of implementation within a given development environment on a case-by-case basis.

Projects like Orlando might thus be required to develop custom software in order to make a taxonomy useful. On the other hand, it is possible that the SGML community might find some means of defining a meta-DTD that would work with existing browsers and allow individual projects to define their own taxonomy. First, however, is the necessity for officially recognizing that there is a need for tag taxonomies.

One advantage of a tag taxonomy is that it could help to prevent unsuspecting academics from misinterpreting results obtained by searching a project such as Orlando, perhaps with the consequence of having that misinterpretation in some way compromise a piece of research. For example, in the biography document of Aemilia Lanyer, the <ETHNICITY> tag is hermeneutic in the sense that the designation it makes is not a definitive one:

> Aemilia Lanyer was English-born but with <NATIONALHERITAGE>Italian </NATIONAL HERITAGE> and possibly <ETHNICITY>Jewish </ETHNICITY> heritage too.

For a reader accessing this document with the question “Was Aemilia Lanyer Jewish?”, the answer currently provided by the tag would be “yes,” in spite of the equivocation provided outside the tag by the phrase “and possibly.” If <ETHNICITY> in this case were redefined as <ETHNICITY 6–HERMENEUTIC>, then the search engine could return the more accurate result, “this is the interpretation of the Orlando project: maybe.” That the Orlando project cannot currently return this result is not a shortcoming of that project; rather, it is an indication of the need for an extended definition of SGML tags and the kind of material they might contain.

Conclusion

There are already a number of SGML and XML projects operating in the humanities, medicine, and law, although only a very limited subset currently make extended use of the potential power of these markup grammars to enhance content with a full range of searchable interpretation. In such projects, however, there is a case to be made for supplementing the existing DTDs with some formalized understanding of conceptual tag level. The tag taxonomy proposed in this paper would add value to interpretive tag projects by explicitly including the intention of the DTD designers in the definition of individual tags. This taxonomic data could subsequently be made available to both the taggers and users as a mechanism for reducing ambiguities in examining specific tag instances.

Acknowledgments

The author wishes to thank the team members of the Orlando project, whose tagging work is being used here. Orlando includes Patricia Clements, Isobel Grundy, Sue Fisher, Terry Butler, and Greg Coulombe (University of Alberta), and Susan Brown (University of Guelph), as well as many others. Thanks are also due Renee Elio (University of Alberta), Susan Hockey (University College, London), and Susan Liepert (University of Alberta).

References

[1] Blake Archive. http://www.iath.virginia.edu/blake

[2] Brown Women Writers. http://www.wwp.brown.edu/

[3] Coombs, James H., Allen H. Renear, Steven J. DeRose Markup Systems and the Future of Scholarly Text Processing. http://www.oasis-open.org/cover/coombs.html

[4] Corpus Legis Project. http://www.juridicum.su.se/iri/corpus, and http://www.ach.org/abstracts/tei10/finke.html

[5] Cover, Robin. The XML Cover Pages – news. http://www.oasis-open.org/cover/sgmlnew.html

[6] Dante, the World of. http://www.iath.virginia.edu/dante

[7] Dickinson Archive. http://www.iath.virginia.edu/dickinson

[8] HL7 XML SIG Mission Statement. http://www.hl7.org/special/committees/sgml/sgml.htm#MISSION STATEMENT:

[9] Liepert, Susan and Stan Ruecker. A Conceptual Taxonomy of SGML Tagging. Unpublished summary.

[10] MARC Standards: Library of Congress: Network Development and MARC Standards Office. http://lcweb.loc.gov/marc/

[11] Mayo Clinic - Clinical Document Management. http://www.hl7.org/special/committees/sgml/PRA/project6.htm

[12] McCarty, Willard. Analytical Onomasticon of Ovid’s Metamorphoses. http://www.kcl.ac.uk/humanities/cch/chwp/mccarty

[13] McGann, Jerome. The Rossetti Archive. http://www.press.umich.edu/bookhome/rossetti/index.html

[14] Medieval Danish History. http://www.ach.org/abstracts/1998/abs31.old.htm

[15] Middle English Collection. University of Michigan Humanities Text Initiative, Ann Arbor, Michigan. http://www.hti.umich.edu/english/ME.html

[16] Orlando Project: An Integrated History of Women’s Writing in the British Isles. http://www.orlando.ualberta.ca

[17] Orlando Project. Unpublished internal project documentation.

[18] Renear, Allen, Elli Mylonas, and David Durand. Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. http://www.stg.brown.edu/resources/stg/monographs/ohco.html

[19] Sperberg-McQueen, C. M. and Lou Burnard. A Gentle Introduction to SGML. http://www-tei.uic.edu/orgs/tei/sgml/teip3sg/index.html

[20] SynEX. http://www.chime.ucl.ac.uk/HealthI/SynEX.

[21] Text Encoding Initiative. http://www.uic.edu/orgs/tei/

[22] Vizine-Goetz, Diane. Using Library Classification Schemes for Internet Resources. http://www.oclc.org/oclc/man/colloq/v-g.htm

[23] Whitman Archive. http://www.iath.virginia.edu/whitman

[24] Wittgenstein’s Nachlass. The Bergen Electronic Edition. http://www.ach.org/abstracts/1998/abs19.old.htm