CHAPTER 4

Using the Extensible Markup Language in Cultural Analysis and Presentation

NATALIE UNDERBERG AND RUDY MCDANIEL

Insights about using the extensible markup language (XML) can be used to assist cultural heritage research in a variety of ways. In an earlier publication (McDaniel and Underberg 2007), we briefly explained the nature and potential usefulness of XML for humanities and social science research involving narrative:

> Metadata is data about data, or descriptive data that is intended to describe or represent preexisting data from another source. Such data does not need to be visible to the user; in fact, metadata is often invisible and works behind the scenes in much the same fashion as hypertext markup language, or HTML. XML is one such metadata classification system that is derived from SGML (the same parent language of HTML) . . . It is no surprise that the next generation Semantic Web is being created based on the foundational elements of XML. Using XML as a metadata system on the Internet can lead to more relevant searches and substantially improved online experiences for a user . . . Rather than simply performing a keyword search for matching keywords, a search engine would be able to perform the additional task of seeking out content based on semantic anchors. This process involves looking for examples of the usage of the words in the context in which they were originally intended. (58)

XML emerged as part of the overall development of humanities computing, concerned as it was for expanding the potential usefulness of a digital collection beyond that of the original researcher's intentions. One major project in the developing field of humanities computing has been the text encoding initiative (TEI), which began in 1987 and developed a widely used XML version of its own (Hockey 2006). XML works by embedding tags within the document that identify relevant features within that document. It is up to the designer to identify the particular features encoded and the relation between them. XML is particularly useful for humanities and cultural anthropology computing because it enables multiple forms of processing, providing a flexible way to encode aspects of textual structure. Hockey offers the TEI example of using XML to identify structural division like chapters, sections, speeches, and so forth, to enable the researcher to retrieve and analyze all speeches by Ophelia in *Hamlet*. XML is also particularly useful in the humanities because of its ability to imitate the work of humanities scholars themselves. As Hockey explains, XML has the potential to more accurately reflect what researchers desire to do. For example, many researchers may want to directly access a particular part of a document rather than having to wade through the entire piece. XML links can utilize XML-based structures embedded within the document itself to enable more precise linking (to a single chapter, for instance). These links can then be stored external to the document and provide researchers with particular paths through the document (Hockey 2006). In the digital humanities, a well-known example is the Perseus Project at Tufts.

XML is important for digital ethnography because of the nature of digital information itself. XML is flexible and able to be manipulated in multiple ways. It has grown considerably in popularity in the digital humanities particularly for those full text sources encoded according to the TEI. XML allows specific details within documents to be presented, interpreted, and manipulated, as well as making it easier to "chunk" and integrate both data and metadata—resulting in innovative publication

forms that make use of the distinctive features of the medium, such as its capacity to enable multiple paths through information and its ability to provide more robust context for that information (Cameron and Robinson 2007; Hockey 2006). The key is to take advantage of the potential afforded by the digital medium to enact cultural analysis, much like hypertext ethnography (discussed earlier), which exploits the features of digital environments to tell a story.

In the 1990s archives and libraries became interested in the delivery of digital resources rather than simply creating catalogs or finding aids for those resources. Publishers had also begun making journals available digitally, but with an eye more toward delivery rather than manipulation or analysis. Digital media was conceived of more as a communication than a computational medium (Hockey 2006). The use of XML in digital ethnography today allows us to go beyond delivery of resources to enable analysis and manipulation.

In addition, the strong emphasis on linking and connections in the humanities, which is facilitated by the use of XML, also enhances digital ethnography. Sharing these links—the vision of founders of the digital field like Vannevar Bush—enables ethnographic resources to be reusable and allows future potential users to access the thought processes that went into organizing and curating the collection in the first place.

This chapter examines how XML is being used in three such projects at UCF that represent collaborations between computer science and ethnographic experts: an analysis of personal narratives of Catholic nuns; the creation of an online portal to share resources and facilitate dialogue between digital humanities scholars and educational game makers and users; and the creation of an online portal to share the digitized newspaper stories from the pioneering Central Florida Hispanic newspaper *La Prensa*, along with the stories, photographs, and memories of Puerto Ricans throughout the diaspora and the interpretations of scholars and community members who are familiar with the events covered. This work follows Crane et al.'s (2005) call

for increased communication and collaboration between scholars, technical experts, and community members. After briefly introducing each project, we discuss the use of XML in achieving ethnographic and cultural heritage research aims.

Introduction to the Projects

In the first collaboration, which we refer to as "exembellishment" (McDaniel and Underberg 2007), we used sample vocation narratives from Underberg's ethnographic fieldwork with southern US and Peruvian nuns to suggest an XML framework for coding and displaying narratives from a particular storytelling community in a digitized format (McDaniel and Underberg 2007). In this type of work, we advocate using the potential of XML to assign meaningful metadata (descriptive data) to narrative. We argue that by creating metatags and assigning them to narrative documents, researchers are better able to research and solicit new stories from specific storytelling communities. Using XML is particularly useful because it can allow one to pattern story scripts based on what are called "document type definitions" (DTDs).

The second project, the Digital Humanities Exchange (DHE) initiative, builds on work by ourselves and other colleagues at UCF on video games in the humanities. It is designed to share what we have learned with other educators and researchers and facilitate exchange between expert/amateur, artist/scholar, and teacher/student within the domain of games-based learning. The project, in its development phase, consists of an interactive community-driven web portal to manage digital assets that can be used in scholarly game-based learning contexts. This portal functions as a trading post where pioneering scholars can exchange ideas, assets, and knowledge of best practices for using game-based learning in the humanities. The DHE includes the Turkey Maiden Educational Computer Game, a computer game mod (a modification of an existing computer game engine)

project directed by Underberg, designed to teach about Central Florida history and culture (discussed in more detail later in the book).

The third project, entitled Digital Diaspora, is intended to create an online portal to share the digitized collection of *La Prensa*, the pioneering Hispanic newspaper founded in Central Florida in 1981, and enable the creation of a public archive related to the events and topics documented in this historically significant newspaper. This project is in the initial stages of development.

Using XML to Analyze Cultural Information

In "Exembellishment: Using the eXtensible Markup Language as a Tool for Storytelling Research" (McDaniel and Underberg 2007), we point out how ideas from folkloristics can inform the development and use of this digital media tool. We suggest that the term *normalform* can be useful for digital storytelling researchers, distinguishing as it does the basic framework for a particular type of folk narrative (Georges and Jones 1995). While for many years folklorists primarily concerned themselves with traditional narratives such as the European folktale, the legend, or the myth, more recently they have turned their analytical eye toward discerning formulas and patterns in personal narratives (see Stahl 1989).

Underberg was able to identify the sociohistorical context that gave rise to the development and performance of a particular genre of personal narrative, the vocation narrative, among Benedictine nuns who entered religious life beginning in the 1980s in two historically related communities, and to identify the underlying structure of the narrative form that was common to all informants.

Very briefly, this development of a longer, more conflict-oriented vocation narrative shared by those entering religious life in these communities resulted from a vocational crisis brought on by the changes to religious life mandated in the 1960s by the

Street scene in Morropón, Peru, home of one of the communities of Benedictine sisters who shared their vocation stories with Underberg. Courtesy: Natalie M. Underberg.

Second Vatican Council. As in many such communities, nearly half of the nuns left the convent under study during the turbulent late 1960s and early 1970s, which resulted in a growing need for more people to join their order. Unlike the elder sisters in the community, who largely grew up in the vicinity of the convent and demonstrated their likely "calling" to religious life through such externally observable activities as frequent attendance at Mass and an expressed desire to be like the nuns/teachers they admired, the newer entrants to the community tend to be older, come from farther away, and to possess different background experiences from those of the existing convent members. The upshot was that, in the 1980s, an aggressive and media savvy marketing campaign was developed to secure new entrants for a community that was in danger of dying out. Thanks to the business acumen and hard work of the sisters and their advisors, the number of vocations to their community eventually grew, in part because of the ability of aspiring entrants to tell a compelling personal narrative about their call to religious life that would be recognized and legitimated as valid by the nuns of the community they wanted to join.

In our project, we used the unique seven-part structural out-

line of a *normalform* for telling a convincing vocation narrative as the set of metatags in XML. To put it in folkloristic—and narratological—terms, the *normalform* of the story, consisting of seven elements, provides the syntagmatic structure for these stories; the crucial part played by discernment (or determining the will of God through prayer and silence) indicates something about the paradigmatic structure of the narratives (see Lévi-Strauss 1969):

1. the receipt of a call and resistance to that call
2. surrender to the call as an account of the first successful listening to God
3. determination of the validity of the call to religious life in general
4. the narrowing of a general call to a particular subtype of religious life
5. realization of the call to the Ferdinand/Morropón Benedictines
6. the facing and overcoming of obstacles following acceptance of the call
7. the identification of the time of entrance into the order and a statement of contentment with the call

At the core of these stories is an opposition between being willing versus being unwilling to listen; the working out of this conflict serves as a gatekeeping mechanism regulating acceptance into the community (McDaniel and Underberg 2007). Identifying both levels of narrative structure—syntagmatic and paradigmatic—can be used to create meaningful XML coding, which can in turn help researchers better identify, elicit, and display these narratives and their contextual meanings.

XML can also usefully clarify ambiguity within texts. XML can thus help narrative researchers dealing with more complex narratives, enabling them to conduct more precise searches on narratives in a given database. For example, they would be able to search not only for keywords but also for specific aspects of keywords—such as with the use of an <event> tag to relate important events within a particular type of story. Such a system could help create smarter searches, in which tags are stored in grouped units accompanied by meaningful related words that provide needed contextual information (McDaniel and Underberg 2007).

We explain how this potential of XML tags for providing semantic meaning to the narrative would work in the case of the vocation narrative:

> Particular tags such as <obstacle> and <call_realization> enable researchers to compare and contrast new stories based solely on the *normalform* of preexisting vocation stories in the database. After this set of tags has been applied, it is relatively easy to create a simple Internet search engine script to parse and control search results based on keyword searches. A user can now search for very specialized information within a story collection; for example "search all vocation stories where location = X, obstacle = Y, and the call narrowing sequence contains keyword Z" would be a perfectly valid search, and would likely return quite accurate results given a large enough set of stories to search through. (McDaniel and Underberg 2007:66)

By beginning with an XML coding of a particular narrative, we hope to contribute to the creation of a narrative research system that enables improved searching and classification of this type of narrative.

Facilitating Exchange: A Database for Sharing Assets for Educational Computer Game Design

As indicated earlier, the objective of the Digital Humanities Exchange (DHE) project is to facilitate the exchange of information, materials, and assets related to educational computer game design. Although the DHE is open to non-heritage-based games, we will focus here on how the DHE can help cultural heritage experts gain and share information and materials. As mentioned

earlier, one of the games included in the DHE is the Turkey Maiden Educational Computer Game mod, a game based on a Spanish folktale collected in Depression-era Ybor City, Florida, the historic "Cigar Capital of the World" and home to Spanish, Cuban, and Italian immigrant communities. Turkey Maiden is a mod of the popular role-playing game NeverWinter Nights. The DHE makes available assets from the game itself, the curriculum, and the primary historical materials used in the creation of the game, such as historical photographs and documents related to 1930s Ybor City.

Mods are the current focus of the Turkey Maiden project because they offer a way for instructors to create computer games for educational purposes without having to build a game engine from scratch (see Koster 2004 and Pearce, Witten, and Barti 2006 for a discussion of the potential of modding to allow players and nonprofessional game designers to customize the game). David Leonard (2006) notes the predominance of white player characters in commercial games and of racial and ethnic stereotypes when nonwhite player characters are included, suggesting the need for game experiences that more responsibly address issues of cultural diversity. Educational computer game mods offer educators the opportunity to construct game characters and worlds that celebrate cultural diversity in a way that is compatible with educational goals related to history and heritage. For example, Turkey Maiden incorporates Federal Writers' Project Works Progress Administration (WPA) materials (available through the University of South Florida's library), historic photographs, and newspaper stories on Ybor City into a game story that takes students on a folktale-inspired educational tour through Depression-era Ybor City.

The DHE portal is intended to enable teachers, humanities scholars (including cultural anthropologists), and others to apply the ideas of cultural learning embedded in the Turkey Maiden game into their own projects (discussed in more detail later in the book), thus using technology to enable (enact) cultural analysis. Integrating these insights into the DHE demonstrates how scholars and others can combine original source materials with

the newest technologies. The project, to culminate in a "trading post" for scholars, will facilitate scholarly dialogue and share information about how educational computer games can be built—even with scarce resources. The project is intended to be a contribution to interdisciplinary information management, using the same kinds of complex visualization systems and resource management interfaces that are seen in contemporary virtual worlds.

The DHE also relates to a problem that is currently the focus of much research in the digital humanities: how to deal effectively with the proliferation of data that characterizes the digital world today. With so much information, more and more sophisticated methods for sorting through that information in meaningful ways become necessary. It is in this niche that humanities work—and the closely related field of cultural anthropology—figures. Finally, the integration of scholarly dialogue through discussion forums and mechanisms for feedback will enable creative new approaches to combining expert knowledge, scholarly skills, and project assets. Such an approach contributes to the creation of a constructivist model in which new media users are less passive consumers than consumers/producers, and tallies with De Lusenet's (2007) argument that, in today's so-called participatory culture, user contributions are fundamental to the creation of the project itself.

DHE is modeled on resources like TurboSquid (www.turbosquid.com), but geared toward a more scholarly audience. Specifically, the DHE will combine the technological capacities of a resource management system with scholarly discussion and integration of humanities-oriented primary source materials. Similar projects include Valley of the Shadows and Explore Art (Himalayan art). The portal will also use open-source technologies such as Apache and XML to design the system, so that knowledge and assets can be traded among users.

The asset management system used in the DHE will be combined with collaborative tools and threaded discussion forums. The portal itself is intended to contain digital files from games, including 3-D models and digitized source materials (such as the

Depression-era Federal Writers' Project WPA materials used in the Turkey Maiden game), that are described in terms of both scholarly contextual information and potential practical uses in the games. Scholars will code project files and assets, and then public access will be enabled; this in turn will enable additional contextual meanings to be assigned to the assets by users of the portal.

In particular, the storage and access aspect of the project is relevant to the humanities because it will enable user access to humanities assets that may be 3-D (like 3-D models) or otherwise better encountered in a 3-D environment than in a linear archive or database (such as one that only involves texts). This will in turn allow users to find out more about how issues related to culture, identity, and history can be explored within a game (e.g., enabling the player to adopt more than one avatar or digital character's perspective).

Further, a portal is being developed so that it can be searched according to narrative parameters such as the sociocultural context of the creation of the asset and the time of its creation. This unique system will allow users to search for content along content- and context-based lines, that is, according to both asset descriptions and historical or cultural significance of events related to the assets.

In practical terms, the DHE is being built using five technologies: (1) a relational database system and server-side scripting language (such as PHP); (2) technical expert-generated metadata to describe file formats and potential uses; (3) expert-generated subject matter metadata to describe the sociocultural significance of special collections; (4) social classification systems to provide user-generated tags for linking and extending archive materials (e.g., Flickr); and (5) threaded discussion boards to facilitate scholarly and producer discussions related to the contents, meanings, and uses of the database. Thus far, we have set up the high-capacity server to house the project and completed the programming and development of a prototype web-based system that is intended to serve as an initial test phase for the portal.

Digital Diaspora: *La Prensa* and the Recuperation of Collective Memory

The Digital Diaspora project is intended to leverage the community wisdom accrued through the creation of an oral history archive to augment the knowledge encapsulated in the Central Florida newspaper *La Prensa* in a way that brings together the insights and perspectives of humanities scholars with members of the public in a forum that encourages dialogue and discussion. Such an approach follows Cameron and Robinson's (2007) admonition to utilize electronic culture's ability to reframe scholarly authority with multiple meanings. Specifically, the project will involve four components: digitizing the newspapers; making them searchable according to metatags determined in consultation with humanities scholars (significant events, people, and themes); enabling the submission of stories, photos, and commentary from the public that relate to the newspaper stories and the events, people, and themes they document; and interpretations of the historical events on the part of scholars and participants.

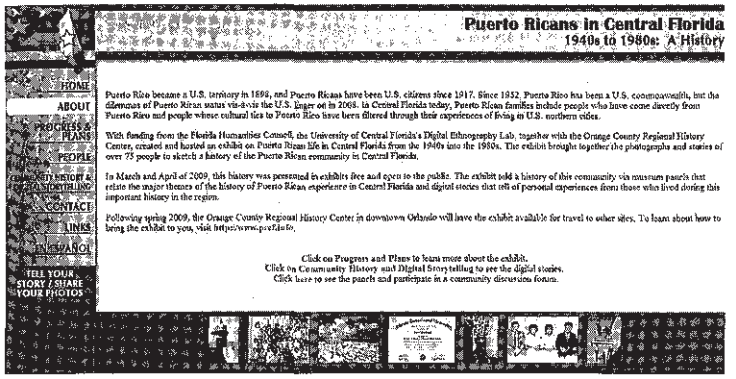
This project builds on fieldwork among Latin Americans in Central Florida that was conducted by Underberg in 2003–2004 and in 2008–2009. In 2008–2009, Underberg and team collected more than seventy-five oral history interviews and hundreds of photographs from Puerto Ricans who had settled in Central Florida in the last fifty years. This project resulted in a 2-D (text panel) and digital story exhibit (**I**). The project also incorporates McDaniel's ideas for using XML (see, for example, McDaniel and Underberg 2007).

Digital Diaspora is important for several reasons. The Puerto Rican population in Central Florida is booming, and Puerto Ricans have played a significant part in building the Central Florida region through the establishment of institutions such as social clubs (e.g., La Asociación Borinqueña), cultural celebrations (e.g., the Puerto Rican parade), and journalism (e.g., *La Prensa*). The growth of the Puerto Rican population in Central Florida has been dramatic, increasing from approximately

100,000 in 1980 to more than 700,000 in 2007. In fact, the number of Puerto Ricans in Florida is second only to that in New York within the fifty US states, and the greatest increase in this Puerto Rican population from 1990 to 2000 occurred in Florida (an increase of 241,354). By far the main destination of Puerto Rican migrants from between 2000 and 2006 was been Orange County, Florida (nearly 35,000). These so-called Disney Ricans have made a significant impact on the social, economic, and political landscape of Central Florida (Duany 2009).

The Puerto Rican experience has been characterized by what sociologist Jorge Duany calls *la nación en vaivén*, or "the nation on the move" (Duany 1996, 2000). Duany uses this phrase to describe the fluid and hybrid identities characteristic of contemporary Puerto Ricans. He encourages scholars to consider the social rather than merely physical spaces in which Puerto Ricans live, and how they create cultural meaning in the diaspora. The project explores the possibility of the Internet to enable the creation and sharing of hybrid identities online (Christensen 2003; Underberg 2006b, 2010; see also García Canclini 1995 for an important discussion of cultural hybridity from the perspective of Latin American studies). With this project we have the opportunity to bring together official and unofficial accounts of important historical events along with the memories and experiences of people who have intersected with these events and may live in diverse communities throughout the diaspora. This project enables us to raise the question: Can the Internet become a space for reconstruction of the Puerto Rican "imaginary" (Flores 2000)?

In order to facilitate a sense of distributed community among Latin Americans residing across the fifty US states, we plan to use Web 2.0 technologies to build an interactive web portal with access to digitized documents from *La Prensa* as well as the oral histories collected from Central Florida Puerto Ricans. The digitized collection will be annotated using searchable metadata that is then further cross-referenced to oral histories collected from Puerto Ricans living in Central Florida from the 1940s to the late 1980s. For the period from 1981 to 1987, official historical records from *La Prensa* can be compared to individual oral histories and accounts from Central Florida residents of that same period. Additional accounts from individuals who have migrated to other areas of the United States will be collected using online bulletin boards and web forms in order to provide a mechanism for bridging different diasporic communities (particularly in regions like New York, New Jersey, Orlando, and Miami).



Puerto Ricans in Central Florida, 1940s–1980s: A History website homepage. Courtesy: Natalie M. Underberg.

Web 2.0 generally relies heavily on social networking and user feedback in order to build a sense of community online. Examples include Facebook, YouTube, and Twitter. In this sense, individuals not geographically co-located can still participate in an online community and participate in and contribute to an online scholarly community based on official historical documents and community stories important to Latin American heritage. Also characteristic of Web 2.0 technologies is the lack of centralized planning (of content) and the emergent and organic growth of the virtual sites based largely on community feedback and participation. In this sense, we will build the foundations of the web portal, and the source scholarly content (*La Prensa* digitized content and the transcribed oral histories and recordings) will act as catalysts for discussion and further contribution of materials from community members.

This process enables a more democratic means of engaging

with source materials and contributing to archival collections. Robert Glenn Howard refers to the category of online discourse made possible through Web 2.0 technologies as the "vernacular web." The vernacular is, by definition, different from the institutional, and serves to call upon an alternate form of authority. He writes that "the concept of a vernacular web provides the theoretical language necessary for speaking about the complex hybridity that new communication technologies make possible" (Howard 2008:192; see also Bolter 2001 and Landow 2006 for a discussion of hypertext's potentially democratizing features). In their introduction to the collection *Democracy and New Media*, editors Henry Jenkins and David Thorburn (2004) note the contrast between older "consensus" forms of media broadcasting and newer technologies for information dissemination that function "according to principles fundamentally different from those of broadcast media: access, participation, reciprocity, and many-to-many rather than one-to-many communication" (Jenkins and Thorburn 2004:2). In this new media economy, the importance of source materials is augmented by the rich contributions of community members, who annotate, extend, challenge, or otherwise shape the new content according to their own insights or perspectives. This is an important feature of social networking technologies and also a central part of Digital Diaspora; this feature seeks to adopt an "epistemic relativist" position as outlined by Cameron and Robinson (2007). This project, in addition to allowing users to comment on source materials and upload new documents for discussion, will feature a mechanism for registering accounts and searching for users based on particular demographic data. The site will also feature a privacy option for those contributors who do not wish to have their profiles accessible to general users.

From a narrative perspective, the implementation of user-generated strategies for sharing personal stories is an interesting research topic because it combines the benefits of flexible hypertext technologies with first-person perspective accounts authored by individuals with ties to historical events. The official reported versions of newspaper articles can be juxtaposed

with the personal accounts of those in attendance during covered events, for example. This will be done using hyperlinks leading from the digitized source materials and custom XML tags that are associated with the personal histories collected from that time period.

Digital Diaspora is intended to draw on a rich lived experience in Florida and the diaspora, and bring together around a unique archive (*La Prensa*) humanities scholars with people who lived the experiences recorded in its pages. We believe this project provides a unique opportunity to use Web 2.0 technologies to address a humanities question of great importance to Latin American studies, specifically, the way that diasporic communities are created and maintained. This project allows us to ask: Can the Internet become a space for creating diasporic communities?

Many researchers working with historical collections recognize the ubiquity, accessibility, and pervasiveness of digital information. In several past projects, content has been organized according to a particular theme, such as the American Civil War theme used in the Valley of the Shadows hypermedia archive at the Virginia Center for Digital History at the University of Virginia, or the Rubin Museum of Art's Explore Art online art exhibit. Others continue work in more ambitious metadata classification systems for the humanities; the Perseus Project at Tufts University is one such example. In addition, we take a cue from such projects as the New York Blackouts at Brown University, which combined archive with oral history research to reconstruct the history of the New York blackouts of 1965 and 1977. Such projects (ours included) raise questions about the verifiability of oral history research when the Internet is involved, and one of the larger research questions to be addressed by our project concerns how to deal with the potential for deception by anonymous participants.

In practical terms, completing a project like this requires several steps of development and collaboration between technical and cultural experts. The first stage, as with all such projects, involves planning and research in order to identify appropriate portal technologies (such as Drupal) and metadata classification

systems; then, a small portion of articles must be selected from the extensive newspaper archives (approximately fifteen) to be digitized; an equal number of oral history excerpts must be transcribed or otherwise prepared for the portal.

The next stage involves digitization and translation, which includes digitization and metadata annotation of documents. Documents will be digitized using optical character recognition (OCR) and corrected by hand. The documents will be translated into English and offered in both Spanish and English for online display. At the same time, the server will be set up with the portal software, and a prototype architecture will be tested to house the document collection.

The third stage is that of encoding, in which the digitized documents will be uploaded to the portal site and annotated using appropriate XML data. At this point corollaries will be identified between *La Prensa* articles of particular thematic interest (e.g., historical, religious, political, cultural, etc.) that may be used as catalysts for inspiring participation from different diasporic communities situated in key geographic areas. Stories added by the community will not be tagged with metadata, but will be linked to the tagged articles using appropriate linked keys in the database tables. This will allow for a realistic rollout of the portal while still connecting these stories and events to the community. This could potentially be a considerable amount of information. These user-added stories will be subsequently annotated and tagged at a later date, if time permits.

When these steps are completed, we will have a prototype launch, which will involve launching a prototype community by mailing publicity materials to press and community contacts identified through research and contacts. As part of the massive publicity campaign necessary to launch in order to develop a vibrant and active online community to create and support the project, the publicity materials will encourage participants to visit the portal and view these Central Florida historical documents as well as to upload their own personal documents and initiate online conversations and make connections in the community. After survey feedback has been gathered from initial users,

the web portal will be modified as necessary before the final step: final release (and continued support).

So far in this book we have investigated new media's ability to tell a story, considered models for collaborative research, and explored the use of digital media to present, manipulate, and analyze cultural information in a way that reconfigures the relation between author and audience, expert and layperson. In the final chapters, we discuss issues surrounding and strategies for teaching about culture using insights from virtual heritage and video game studies.