Report on Creating a Snapshot of the Princeton Charrette Project

RAFAEL C. ALVARADO, PH.D.
Co-Director of the Princeton Charrette Project

Tuesday, July 25, 2006

Contents

1 About this document

This document is a follow-up report to my work for the Department of French and Italian at Princeton University on behalf of the Princeton Charrette Project during the week of 26 June and after. I am very grateful to Janet Temos, Director of the Educational Technologies Center at Princeton, and to Serge Goldstein, Director of Academic Services, for the space and resources they graciously provided to perform this work. Approval for the general plan and outcomes described herein come from Gina L. Greco and K. Sarah-Jane Murray, both project co-directors from 2003 to the present.

The term "Princeton Charrette Project" (PCP), as used throughout this document, refers narrowly to the collaborative efforts of Professor Karl Uitti and his students to produce a digital critical edition of the Old French romance, *Le Chevalier de la Charrette*, up the point of Prof. Uitti's untimely passing in 2003. The term also refers more specifically to the collection of digital materials produced by this collective effort, which have been hosted on the web by Princeton University, and which have been maintained there until the present time. By this definition, the PCP has served as a foundation for at least two other projects, the *Charrette* Project 2 at Baylor University,[1] under the direction of Ms. Murray, and *Le «Projet Charrette» à Poitiers*, a mirror site in France.[2]

2 Aim and Scope of the Work

The aim of the current work is to create a "snapshot" of the extant PCP collection in order to achieve the following goals: (1) to commemorate the contribution of Prof. Uitti and his team of *charretteurs* to both Old French scholarship and digital philology, (2) to create a lasting, library-quality archive of the materials associated with this contribution, and (3) to create a foundation for future and parallel projects that build on this contribution.

To achieve these goals, the work of the PCP editors has proceeded on two levels—the administrative and the technical. At the administrative level, Gina Greco has secured buy-in from Princeton University legal counsel, the Office of Information Technology at Princeton, and the Department of French and Italian to host and sponsor a snapshot of the PCP collection for an indefinite period—the *longue durée* for the Internet being an indefinite concept in itself. At the technical level, I have endeavored to perform the actual work of creating the snapshot. This has meant two things: one the one hand, the consolidation and organization of the PCP collection, and, on the other, the development of a plan for creating a web presence for these materials that will require as little maintenance as possible in the years to come.

The goals of creating a foundation for future projects and leaving an institutionally maintainable installation of materials have required me to go beyond

[1] URL: http://lancelot.baylor.edu

[2] URL: http://www.mshs.univ-poitiers.fr/cescm/lancelot/index.html

the task of simply gathering all extant materials and linking to them from a web page (along with annotations and credits). The problem has been to distinguish between primary content, which should remain intact as a record of work accomplished by the PCP, and secondary content, which merely provides access to the former, and which is need of repair. The following conception of the PCP's information architecture provides a solution to the problem of scoping the project, and a rationale for the work accomplished.

2.1 Information Architecture

I make a broad distinction between three levels of description—a tripartite "stack" of encoding which in practice is often conflated.

2.1.1 Level 1: Core scholarly content of the project

This is the level of scholarship *per se*. It includes the material results of the paleographical and philological work performed by Prof. Uitti and his graduate students. These results typically exist as printed documents, but sometimes may be "born digital" and exist only in the forms described the second level. The contents at this level also form a stack, but of interpretation, beginning with the raw paleographic work of image capture and diplomatic transcription and ending, for now, with the assignment of figures to the critical edition. On top of this stack, it is hoped, will proceed other forms of critical and interpretive scholarly work, including forms of statistical and structural criticism that can take advantage of the work of encoding described next.

The sets of materials at this level are the following:

1. The folio side facsimiles (photographs) taken of from the original manuscripts.
2. The diplomatic transcriptions of these materials, including a typology of what I shall call "glyphic" forms—punctuation marks, alphabetic characters, and majuscules.
3. A critical edition of the poem contained in the manuscript collection as a whole, edited by Profs. Foulet and Uitti.
4. A lexico-grammatic index of the critical edition describing the grammatic features of each word in the critical edition.
5. A catalog of rhetorico-poetic figures (*adnominatio*, *chiasmus*, enjambment, *oratio obliqua*, *oratio recta*, and rich rhyme) that classify segments of the critical edition through a series of overlapping layers.

2.1.2 Level 2: Encoding formats and files

This level refers to the scholarly content of the preceding level as represented in the digital medium for machine processing and networked distribution. These include such items as SQL data tables and their definitions, XML text file files and their definitions (such as the TEI Document Type Definition), JPEG image files, etc.

Work at this level has been performed by both scholars responsible for work at the first level, and digital specialists with no particular expertise in Old French or paleography. The primary contribution at this level is to produce representations of scholarly content that may be stored, processed and distributed effectively through various digital channels, such as but not limited to the web.

The items at this level are:

1. The JPEG images of the Manuscript pages associated with the eight manuscripts that comprise the manuscript tradition at the base of the PCP.
2. The XML-encoded diplomatic transcriptions of the glyphic content of these manuscript pages, including a system for classifying and describing such specific glyphic features as punctuation marks, special characters, and majuscules (large, decorated letters that initiate various lines throughout the manuscripts.) These have been encoded using a variation of the TEI Lite (P4) DTD.
3. The SQL-encoded sets of data associated with the philological work superimposed on the original Foulet-Uitti critical edition, and stored as tables within the *Figura* database. These data sets include: (1) a version of the textual content of the critical edition itself, parsed by word and punctuation mark; (2) the lexical and grammatical data for each word in the poem; (3) the rhetorico-poetic data, comprised of "figures" that mark the presence of variant forms of *adnominatio*, *chiasmus*, enjambment, and rich rhyme throughout the poem, as well as instances of direct and indirect speech along with the name of the character associated with the voice; and (4) a detailed description of the majuscule data associated with the paleographic data in the previous two sets. The *Figura* database also contains tables with transcription data and references to the JPEG images, so that these data can be used in presenting the PCP materials in rich, database driven web applications.

Together, these three sets comprise the encoded data that exists at Level 2, and which effectively represent the snapshot of Level 1.

2.1.3 Level 3: Presentational views

This level provides user access to the contents of Level 2. In our case, these are largely HTML files hosted on Apache-served web sites. (Within this level, one may distinguish further between the "model," "view" and "control" layers associated with web application design, but these may be effectively bracketed off in the current discussion.) Variation at this level occurs to provide different visualizations of the same data for different audiences and contexts. For example, items at Level 2 may be packaged as "learning objects" for use in the undergraduate classroom, or they may be exposed within a more sophisticated application for statistical analysis and 3-D visualization.

As a general design principle, it is wise to separate this level from the previous, in order to allow for the flexible re-deployment of scholarly data for different purposes. Indeed, this is one of the fundamental design principles of XML, which, strictly speaking, is intended to represent *logical* structures, not *rhetorical* views. The latter are produced through the application of "stylesheets," in the form of CSS or XSL, which transform and embellish documents to produced context-specific documents for end user consumption.

Examples of items at this level include the following:

1. The original Princeton web site.[3]
2. The Figura series of web applications, one of which was used to collect figural data, the other to display it.[4]
3. The new web site to present the current snapshot.

2.1.4 Summary

This architecture is summed up in the following diagram:

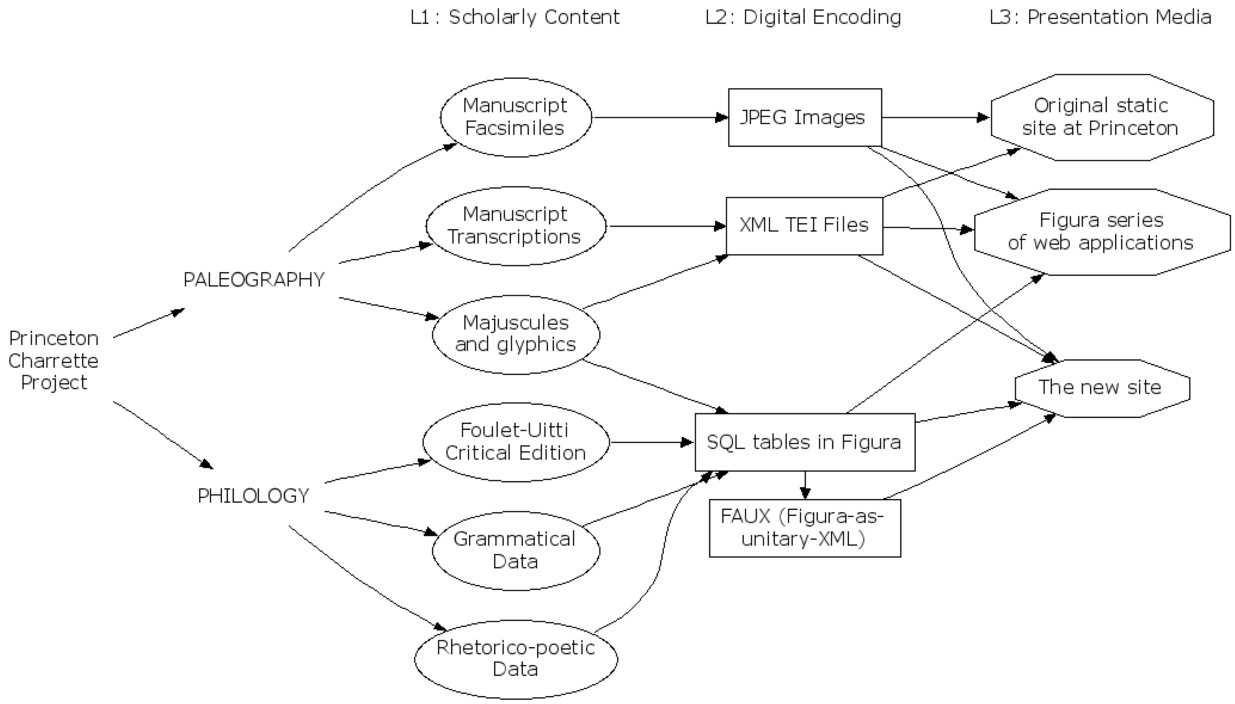

Figure 1: The Information Architecture of the Charrette Project. Note that movement from left-to-right reflects progressive *encoding*, and, roughly, from top-to-bottom progressive *interpretation*.

3 Method and Results

Given this framework, the guiding plan of the current work has been *to preserve the contents of Level 1 by means of small but important changes at Levels 2 and 3.*

[3] URL: http://www.princeton.edu/~lancelot

[4] URL: http://ontoligent.com/figura

3.1 General modifications to level 2

In addition to consolidating all of the digital files associated with this level, it has been necessary (1) to simplify and standardize their formats, (2) to provide essential metadata, including information about rights, provenance, formatting and subject matter, and (3) to provide enough documentation for the materials to allow librarians and researchers to take full advantage of the formats in which the materials are encoded.

In order to achieve the goals of archival longevity and supporting other projects, I have either encoded or recoded the entire collection of textual materials at Level 2 in XML, including all of the data stored in *Figura.*[5] The rationale for using XML is simple: although not the most effective format for data storage and manipulation—relational databases remain superior in this respect—it is by far more "future-proof" than a database application, and has been endorsed by librarians, archivists and digital scholars throughout the world as the *lingua franca* for encoding humanities computing material. Importantly, this view is shared by the Educational Technologies Center, who will be hosting the material on their servers for the long-term.

3.2 Specific modifications to Level 2

3.2.1 Modifications to the Images

The names of all of the images have been changed to match the convention for identifying a manuscript folio side within the transcriptions and *Figura*, namely the following hyphen-delimited sequence: (1) a capital letter standing for the particular manuscript, (2) a number standing for the folio number, and (3) a small "r" or "v" standing for *recto* and *verso* respectively. The filename extension has been changed from "`jpeg`" to the more common "`jpg`." Finally, a set of thumbnail images have been generated to aid developers of web sites and applications that use these materials.

3.2.2 Modifications to the Transcriptions

Transcriptions have retained the formatting accomplished by Alexei Lavrentiev, with the following major change: each manuscript line element now contains an attribute to define its physical position on the manuscript page—defined as column ("a," "b," or "c") and a "row" number indicating the line number in the column. This eliminates the need to include empty line numbers in transcriptions to represent critical edition lines not found in the source manuscript. References to critical edition lines remain as "key" attributes in the line elements, e.g. as "`FU-32`". In addition, naming conventions for manuscript pages and other items have been normalized. Note that the effect of viewing synoptically the relationship between

[5] XML stands for "eXtensible Mark-up Language," and refers to a standard for encoding textual documents that has been in existence for over ten years—much longer if once considers that it is a simplified subset of SGML (Standard Generalized Mark-up Language), which was invented in the early 1970s and has become a standard in the publishing industry.

critical edition lines and their associated manuscript pages is now accomplished with the "FAUX Charrette," which is described below. Transcriptions are organized into files by whole manuscript, and files are named as follows: "`MS-A.tei.xml`", where the capital letter following "`MS-`" refers to the manuscript code.

3.2.3 Modifications to *Figura*

All of the data stored in *Figura* have been exported to a single, comprehensive XML document known as "FAUX" (Figura-as-unitary-XML). The file uses a custom set of tags, as opposed to TEI, for sake of simplicity. In addition, the identifiers for each of the significant elements encoded in the document collection have been normalized for linking among documents, and for reference in scholarly publications. These include specific lines and words in the critical edition, manuscript pages, and figure instances.

In addition to FAUX, the entire SQL data definition file for *Figura* has been exported from the MySQL database server it is currently hosted on, and may be downloaded for importing.

3.3 Modifications to Level 3

To provide maximum user access, the entire inventory of properly encoded materials will be presented on a single, simple web page that will be hosted by Princeton University in perpetuity and made available off-line as a CD or DVD. This page will contain links to the files themselves, in compressed format, along with metadata, documentation, and credit information about the participants of the PCP. In addition, it will contain links to extant items at Level 3—the current Princeton web site, and the *Figura* interface—along with links to parallel projects. This page is currently under construction, but will be available by August 2006.

3.4 Summary

In sum, the final concrete deliverable of the project is a web page dedicated to the PCP, to be hosted in the Princeton University web system, sponsored by the Department of French and Italian, and maintained by the Educational Technologies Center. When completed, the web site will be located at:

http://www.princeton.edu/~lancelot

The web page will contain the following sections and links:

1. A brief description of the web page and its purpose.
2. A brief description of the PCP.
3. An inventory of each of the items and Levels 2 and 3 above, including metadata. These items are:
    a. Level 1:
        i. The Manuscript Images

ii. The Manuscript Transcriptions

iii. The *Figura* SQL data, including:

1. The Majuscule and Character Data
2. The Foulet-Uitti Critical Edition
3. The Lexical-grammatical Data
4. The Poetico-rhetorical Data

iv. FAUX, or "*Figura*-as-unitary-XML," including the same Figura SQL data

b. Level 2:

i. The Original Charrette Web Site

ii. The *Figura* Series of Web Applications

* * *