A Companion to Digital Literary Studies
Edited by Ray Siemens and Susan Schriebman
Presented by ADHO with the permission of Blackwell Publishing

30.

Practice and Preservation — Format Issues

Marc Bragdon, **Alan Burk**, **Lisa Charlong** and **Jason Nugent**,

Introduction

Text analyses, project collaborations, and the myriad other research activities of humanities computing center largely on the use of artifacts in various digital formats. As containers through which content is engaged and transformed, these formats are currencies for scholarly trade. They must inspire confidence, evidenced by broad use, in their ability to meet present and future research needs. To each format its purpose, though, regardless of respective uses, they must meet common criteria to be considered appropriate to a given task. Not surprisingly, these criteria are very much identified with standard application support. Community and cross-community use of artifacts, and the long-term preservation of digital content, demand of formats a high level of interoperability across platforms and applications as well as the sort of version compatibility identified with industry-based standards.

The formats discussed herein — eXtensible Markup Language (XML), Portable Document Format (PDF), Tagged Image File Format (TIFF), Joint Photographic Experts Group (JPEG), and JPEG 2000 — are all either de facto or emerging standards for formatting many of the digital objects employed in humanities computing research. Each is discussed in light of their facility in promoting associated discipline goals.

XML is a file format used by many humanist researchers and scholars for the creation, description, and exchange of textual materials. While use of XML is often in the areas of text description, text analysis, and processing, XML is also used for methods of storing or preserving texts. It is this point, XML as a text storage or archival format, which is the focus of XML in this chapter.

Unlike XML, PDF and PDF/A are not formats that digital humanists normally employ for capturing and distributing literary texts or other genres. Still, examining the strengths and weaknesses of PDF is important because of its prevalence on the web and for assessing its place relative to the use of other formats for digital humanities.

TIFF and JPEG are the de facto standards for digitally representing visual matter —the former for preservation and the latter for access enablement. JPEG 2000 is an emerging standard that strives to combine the richness of TIFF with the portability of JPEG to create a single standard that accommodates multiple purposes on multiple platforms.

XML

Discussion about XML as a file format for preservation generally refers to two distinct yet overlapping contexts: XML as a file format for describing text and XML as a format and mechanism for describing information *about* that text or *metadata*. Both contexts are addressed in this section.

Before discussing preservation and XML as a preservation file format, it may be useful to look at XML's predecessor, SGML, as well as the syntax and general principles within each language. SGML (SGML, ISO 8879:1986) is a descendant of Generalized Markup Language (GML), developed in the 1960s at

2008

Practice and preservation: Format issues

Bragdon, Marc