Journal of the Text Encoding Initiative

Issue 5 | June 2013
TEI Infrastructures

TAPAS: Building a TEI Publishing and Repository Service

Julia Flanders and Scott Hamlin

Electronic version
URL: http://journals.openedition.org/jtei/788
DOI: 10.4000/jtei.788
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Julia Flanders and Scott Hamlin, « TAPAS: Building a TEI Publishing and Repository Service », *Journal of the Text Encoding Initiative* [Online], Issue 5 | June 2013, Online since 24 June 2013, connection on 01 May 2019. URL : http://journals.openedition.org/jtei/788 ; DOI : 10.4000/jtei.788

This text was automatically generated on 1 May 2019.

TEI Consortium 2013 (Creative Commons Attribution-NoDerivs 3.0 Unported License)

TAPAS: Building a TEI Publishing and Repository Service

Julia Flanders and Scott Hamlin

1. Introduction

1 The question every TEI workshop instructor dreads: "So, how do I publish my TEI data?" This question and the setting that triggers it are central to the genesis of TAPAS, the TEI Archiving, Publishing, and Access Service (http://www.tapasproject.org). Following a TEI workshop at Wheaton College, participants reflected on the fact that learning and creating TEI data seemed to lie well within their grasp, but publishing that data seemed to depend on institutional resources (server space, programmer time, XML publishing expertise) that they could not count on. The central question arising from that meeting came to be the mission statement for TAPAS: how can creators of TEI data at small institutions publish and store their data, and get access to technical support for their projects? It is primarily to serve this set of needs, and this community of users, that TAPAS exists. TAPAS is currently under development by a team that includes a full-time web applications programmer, several repository experts, a TEI schema developer, a user interface design and development group, and a growing team of scholarly collaborators whose TEI projects and data form the testing ground for the service. More information is available at http://www.tapasproject.org. We plan an initial launch date early in 2014, and our goal in this article is to describe the goals and philosophy of the service and the general approach being taken to its design. It should be noted at the outset that all details given here are provisional and subject to change as the project develops.

2 Several key factors have shaped the way TAPAS is developing in response to the needs articulated at its inception. One of the most important is the question of scale: in the age of "big data," digital publication increasingly calls for interfaces that support visualization, text analysis, and other forms of reading that situate the individual text within a much larger ecology of meaning. However, such systems typically lie beyond the

capacity of individual data producers, for reasons of technical ability, funding, and also the quantity of data they have to work with. TAPAS seeks to respond to this issue by offering its users access to a shared publication infrastructure. This infrastructure can provide large-scale services that operate at the TAPAS level, thereby situating the individual scholar's data within a larger ecology of TEI data and also providing a set of tools for publication, analysis, and visualization that would otherwise be out of scope for an individual to develop and maintain. However, TAPAS and its users will face challenges in designing a system that will work gracefully both at this level of scale and also at the level of the individual text or project.

3 A second crucial question for TAPAS to address is that of data vulnerability. Ironically, despite the claim that the digital medium puts authors in a position of greater self-reliance and independence (by removing dependence on commercial publishing infrastructure), in fact, for humanities scholars, the digital medium requires greater and longer-term dependence on institutional infrastructure (or a commercial equivalent) to ensure the ongoing viability of scholarly data and electronic publications. Unlike books, digital data and publications cease to exist very quickly if not actively maintained, so scholars cannot simply publish their work and move on to another project. TAPAS seeks to respond to this problem by offering a place where data can be curated and published without ongoing involvement by the contributor. However, there are significant challenges here as well: most importantly, how to frame TAPAS's responsibilities so that they are feasible, funded, and clear. An open-ended commitment to unlimited curation of complex TEI data is neither feasible nor fundable, but in order to meet the needs of TAPAS contributors we do need to establish a bounded time period during which TAPAS will take responsibility for contributed data, and a business model that will provide support for TAPAS during that period.

4 The third substantial challenge that TAPAS is grappling with has to do with the customizability of the TEI Guidelines, and with the variability of data produced even under "unmodified" TEI. Although in discussions within the TEI community, this variability is often treated as a disadvantage to be overcome and if possible eliminated, for TAPAS it poses a very interesting set of research questions about the nature and purpose of the actual variability exhibited by TEI data from small scholarly projects. Because these projects often use markup to engage in a detailed analysis of a single text (in many cases chosen precisely because of its idiosyncrasies or structural difficulty), the ability to use the TEI Guidelines in flexible and perhaps unprecedented ways is an important aspect of these projects' scholarship. Customizations—particularly those that establish new controlled vocabularies, create new elements, or modify the structure of existing elements—may constitute an important contribution to our understanding of textual structure or of a particular text's ways of making meaning. TAPAS thus has an interest in supporting these forms of scholarship. At the same time, because of its data curation goals, TAPAS also has an interest in reducing casual or unintentional variation. The project responds to this complex set of desiderata by studying the customizations and encoding of contributors' projects, by providing a target schema that reflects good practice, and by encouraging convergence in areas where data variance does not reflect principled disagreement. At this early stage in the project's development, studying the problem is the first priority.

5 This project takes place within a landscape already well populated with large-scale infrastructural projects (Hedges 2009), such as TextGrid (http://www.textgrid.de/;

TextGrid 2010), DARIAH (http://www.dariah.eu/), CLARIN (http://www.clarin.eu/external/), and the Canadian Writers Research Collaboratory (CWRC, http://www.cwrc.ca/). Projects of this kind must all confront a central set of strategic concerns and design challenges, including questions about how much uniformity to impose upon the data, how to accommodate variation, how to create interoperability layers and tools that can operate meaningfully across multiple data sets (Blanke et al. 2011), and how to manage issues of sustainability (of both the data and the service itself). TAPAS is distinctive within this landscape because of its focus on a single form of data (TEI-encoded research materials) and also because of its initial emphasis on serving an underserved constituency (scholars at smaller or underresourced institutions) rather than on providing an infrastructure that can operate comprehensively. However, there is clearly a great deal that TAPAS can learn from these projects, and there are potentially significant connections to be made between projects as well.

2. Background and Genesis

6 The development of TAPAS has proceeded in several stages. Following the original conversation described above, representatives from several small liberal-arts colleges including Wheaton College, Willamette University, Hamilton College, Vassar College, Mount Holyoke College, and the University of Puget Sound, and later joined by Brown University and the University of Virginia, sought and received a one-year planning grant from the Institute of Museum and Library Services (IMLS). This grant, which began in September 2009 and ended in December 2010, funded a series of four planning meetings at which the group developed the basic outlines of the TAPAS service and a rough specification for its activities. At this stage, the key features of the service included a repository in which TEI data could be stored on behalf of contributors, and which would also expose TAPAS data to third parties via an API; a public-facing interface that would permit readers to search, browse, and read texts from TAPAS collections; and an access-restricted interface for the upload and management of contributed data (including the configuration of publication and retention options). At this stage the planning group also articulated the principle that TAPAS data would be published on an open-access basis, without charges to the reader, and that the cost to contributors should be as low as possible (consistent with sustainability) with some form of free upload options available to support short-term experimentation. This work served as the basis for two further grant proposals: one, also funded by the IMLS, for a two-year National Leadership Grant (2011–13) to implement the essential infrastructure of the service; and another, funded by the National Endowment for the Humanities (NEH), for an 18-month Digital Humanities Start-Up Grant (2011–12) to fund the design and development of the user interface. Both of these development activities are now underway as of the time of writing.

7 In the development process the project's understanding of its scope and audience has undergone some shifts which are worth examining. First, although researchers at small liberal-arts colleges are still a crucial constituency, we are increasingly aware that the challenges they face (lack of access to XML expertise, programmer time, and repository services) are not limited to these institutions. Repository services at large institutions may still not provide the kinds of TEI-sensitive publishing opportunities that TAPAS seeks to offer, and researchers may still have difficulty getting support for a complex TEI

project if they are the only TEI user at their institution, however well-resourced. Furthermore, the value of TAPAS may not be simply remedial: even for projects with publication sites of their own, contributing their data to TAPAS may offer other kinds of benefits, such as the value of putting their data into a shared ecology where it can be examined in relation to that of other projects or accessed via the TAPAS API. Another set of considerations has to do with TAPAS's ambitions towards sustainability and longevity, and the nature of the commitments the service is able to make to its contributors. It is clear that at this stage of its development TAPAS cannot make firm long-term commitments, let alone open-ended ones, but the question of how long TAPAS can preserve contributed data also entails further research concerning the needs of TAPAS contributors. It may be that long-term preservation (however that is to be defined) is less important than short- or medium-term publication and project support.

8 The initial launch of the TAPAS service is tentatively planned for the end of the IMLS implementation grant, in 2014; additional funding is already being sought for further development work, since the service at that point will be quite basic. The most important aspect of the service's future development, however, lies in the development of a TAPAS community, both individual and institutional. TAPAS was conceived from the start as a community-supported endeavor and as a way of focusing existing forms of mutual assistance. In this respect TAPAS anticipates working closely with the TEI Consortium itself. By providing ways for digital humanists (who may not immediately identify themselves as part of the TEI community) to engage with others at the same level of expertise and with the same kinds of support needs, TAPAS seeks to strengthen the existing connections between the TEI Consortium and other parts of the digital humanities community. At the institutional level as well, TAPAS has from the outset drawn on a collaborative network of colleges and universities and will seek to broaden that base of participation. In particular, while the TAPAS repository is currently hosted at Brown University, other aspects of the service can be hosted at other institutions, and even the repository hosting is envisioned as mobile over time.

3. Services

9 When completed, TAPAS will offer several different kinds of services and points of access to several different kinds of users. TAPAS *contributors*—creators of TEI data—will be able to upload TEI files (and perhaps other kinds of related data such as page images) to TAPAS, using either a "sandbox" interface which will require very little metadata and minimal authentication, or a fuller ingestion interface aimed at those who want to create durable projects within TAPAS. The sandbox interface is currently envisioned as a space for experimentation and short-term usage, through which users could quickly and easily see their TEI files displayed on the Web. In a teaching context such as a workshop or humanities course, the sandbox could offer an easy, immediate way for novice TEI users to see the results of their encoding work. Access to the fuller set of TAPAS publication and data curation services described below will probably require some form of membership, probably with a fee, but these details have not yet been fixed.

10 Contributors will also be able to publish their TEI data through TAPAS, creating *projects* which may in turn include one or more *collections* of TEI documents. TAPAS will offer a standard publication interface at the outset, which as the service develops will become increasingly configurable so that projects can adapt the interface to the specific needs of

their data. Expert users will be able to develop and use their own XSLT and CSS, while novice users will be able to take advantage of the standard appearance and its simple configuration options to create simple publications quickly.

11 As part of the data upload and management process, it will be important for contributors to be able to inspect, validate, and troubleshoot their data. Because of the likely diversity of data contributed to TAPAS, there may be some TAPAS tools—for instance, visualization options that require the presence of specific elements—that will only be relevant to a small subset of TAPAS data. In addition, some novice users may not be aware of inconsistencies or invalidities in their data, and will need a user-friendly way of discovering and remedying these problems before their data can function within the TAPAS publication system. The larger question framing these features is that of the constraints TAPAS would impose on the data contributed. While TAPAS does not have a strong interest in imposing such constraints for their own sake, it does have a functional interest in constraint that operates at two levels. The first level has to do with the feasibility of the most essential TAPAS functions, such as the basic TAPAS publication interface and the corpus-wide search and discovery features. At this level, TAPAS has an interest in imposing some minimum requirements for metadata, and also in ensuring that the data contributed is some form of TEI. Data not meeting these basic requirements might still be accepted (and this is a question TAPAS still needs to answer), but the contributors might be advised that certain basic functions of the service may not work as expected. These first-level requirements will probably be expressed as a kind of "gateway schema" that imposes only minimal constraint on the data. The second level has to do with the more ambitious and specialized tools that TAPAS can offer to projects whose data meets a higher standard of consistency, explicitness, and detailed encoding. At this second level, contributors have an interest in meeting some optimum requirements for the encoding of specific features, if they want to take advantage of the more powerful handling of those features. For example, projects wanting to have dated entry–based materials (such as diaries or letters) navigable via a timeline would need to encode the dates of the entries in a specific way. These higher-level requirements would be expressed in a "target schema" that contributors could use in creating their data, or to which they could convert existing data. Such requirements (like the basic requirements described above) might be optional for users who were prepared to create their own publication stylesheets from scratch, or who were only interested in TAPAS as a repository service.

12 TAPAS will also offer a set of supporting services to contributors, supplying the technical expertise that is often difficult for small projects to procure for themselves, particularly in small increments. The services currently under consideration that seem likely to offer the greatest value to TAPAS contributors are assistance with data conversion (for instance, to assist in migrating data from non-XML formats or from older versions of TEI, or to help fix problems of inconsistency), technical consulting, TEI workshops, and assistance with project management. For instance, TAPAS could assist contributors in developing encoder training materials and local documentation, or in designing schema customizations for use in data capture. Revenue from these services would help fund TAPAS's core activities.

13 TAPAS *readers*—members of the public with an interest in TEI data or in the content of specific TAPAS projects—will be able to access and work with TAPAS data in many different ways. TAPAS as a corpus of projects will have a corpus-level interface through

which those new to TAPAS can explore, find projects and texts of interest, analyse the TEI encoding used, and perform other corpus-level activities. Individual projects will also offer their own views of their data, through their published collections, and a reader interested in a specific project might visit it directly and use its collection interface to search, browse, and read the materials it contains. TAPAS also plans to offer an API to the TAPAS data, to support third-party use of TAPAS data.

14 The final dimension of the TAPAS service is the preservation and curation of TEI data, for which the TAPAS development team is still considering various different models. It has been clear since the early planning process that long-term preservation of data is of great concern to many of those who constitute TAPAS's potential body of users, so the demand for such a service is clear. However, several questions need to be addressed before the terms of such a service could take a firm shape. First, how sustainable is TAPAS? The services it plans to offer are clearly of value, but can its potential users afford to pay the real costs of providing such services? Or are there other ways of funding these services? Second, what kind of long-term preservation do contributors really need? Do they need a trustworthy place to store their data for a clearly defined interim period, or do they need a more open-ended guarantee that their data will remain usable and accessible? Can TAPAS count on users to remain active participants in the management of their data, or should we assume that some data will become the sole responsibility of TAPAS? And if the latter, is that a responsibility we can realistically take on? Third, what provision does TAPAS need to make for migrating data into future versions of TEI, and how might such migration be funded?

15 Clearly the TEI community, and the TEI Consortium itself, have a significant role to play in the shaping of TAPAS. Although the service has a primary responsibility towards users in its target constituency—individuals and small projects who lack access to supporting resources—there may be advantages to all in planning for much broader use. Some TAPAS services, such as training and technical support, have been clearly needed for some time and TAPAS may be a useful way of organizing and supporting them. For others, such as long-term data curation, it may require the input of the entire community and a further development phase to make an appropriate plan. TAPAS welcomes your input and support.

16 TAPAS is made possible in part by a grant from the U.S. Institute of Museum and Library Services, and in part by a grant from the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed in this article do not necessarily represent those of the National Endowment for the Humanities.

BIBLIOGRAPHY

Blanke, Tobias, Michael Bryant, Mark Hedges, Andreas Aschenbrenner, and Michael Priddy. 2011. "Preparing DARIAH." In *Proceedings: 2011 Seventh IEEE International Conference on eScience*, 158–65. Los Alamitos, CA; Washington, DC: IEEE. doi:10.1109/eScience.2011.30.

Hedges, Mark. 2009. "Grid-enabling Humanities Datasets." *Digital Humanities Quarterly* 3.4. http://www.digitalhumanities.org/dhq/vol/3/4/000078/000078.html.

TextGrid 2010. "Roadmap Integration Grid/Repository." TextGrid, December 2010. http://www.textgrid.de/fileadmin/berichte-2/report-1-2-1.pdf.

ABSTRACTS

This project report details the goals and development of the TEI Archiving, Publishing, and Access Service (TAPAS). The project's mission centers on providing repository and publication services to individuals and small projects working with TEI data: a constituency that typically lacks access to institutional resources such as server space, XML expertise, and programming time. In developing this service, TAPAS addresses several important challenges: the creation of a publication ecology that operates gracefully at the level of the individual project and the TAPAS corpus; the problem of the vulnerability of TEI data in cases where projects cease their activity; and the variability and complexity of TEI data. When completed, the TAPAS service will enable its contributors to upload, manage, and publish their TEI data, and it will offer a publication interface through which both individual project collections and the TAPAS collection as a whole can be read, searched, and explored. It will also provide a range of related services such as technical consulting, data curation and conversion work, TEI workshops, tutorials, and community forums. This article discusses the philosophy and rationale behind the project's current development efforts, and examines some of the challenges the project faces.

INDEX

Keywords: infrastructure, publishing tools, repositories, customization

AUTHORS

JULIA FLANDERS

Julia Flanders is the Director of the Brown University Women Writers Project (http://www.wwp.brown.edu), and a member of the Center for Digital Scholarship in the Brown University Library. She has served as president and vice president of the Association for Computers and the Humanities, and vice-chair and chair of the Text Encoding Initiative Consortium. She is the editor-in-chief of *Digital Humanities Quarterly* and also co-directs the TAPAS project. Her research focuses on text encoding, scholarly communication, and digital editing.

SCOTT HAMLIN

Scott Hamlin is the Director of Research and Instruction in Wheaton College's Library and Information Services (Norton, MA). At Wheaton, he leads a group of academic technologists and research librarians responsible for supporting information fluency and effective teaching and learning experiences through the use of information resources and technology. He has worked for many years with multiple institutions to promote the use of TEI at small undergraduate liberal arts colleges and is currently a project director for the TAPAS Project (http://tapasproject.org).