Journal of the Text Encoding Initiative

Issue 1 | June 2011
Selected Papers from the 2008 and 2009 TEI Conferences

Computational Work with Very Large Text Collections

Interoperability, Sustainability, and the TEI

John Unsworth

Electronic version
URL: http://journals.openedition.org/jtei/215
DOI: 10.4000/jtei.215
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
John Unsworth, « Computational Work with Very Large Text Collections », *Journal of the Text Encoding Initiative* [Online], Issue 1 | June 2011, Online since 08 June 2011, connection on 01 May 2019. URL : http://journals.openedition.org/jtei/215 ; DOI : 10.4000/jtei.215

Computational Work with Very Large Text Collections

Interoperability, Sustainability, and the TEI

John Unsworth

1 The "I" in TEI sometimes stands for interchange, but it never stands for interoperability. Interchange is the activity of reciprocating or exchanging, especially with respect to information (according to Wordnet), or if you prefer the Oxford English Dictionary, it is "the act of exchanging reciprocally; giving and receiving with reciprocity." It's an old word, its existence attested as early as 1548. Interoperability is a much newer word with what appears to be military provenance, dating back only to 1969, meaning "able to operate in conjunction." The difference is worth dwelling on for a moment since it's important to the discussion here: for the interchange of encoded text you need an agreed-upon interchange format to which and from which various encoding schemes are capable of translating their normal output. Interoperability, on the other hand, implies that you can take the normal output from one system and run it, as is, in a different system—or to put it another way, the difference between an interchange format and an interoperable format would be that various systems actually operate directly on the interoperable format, while an interchange format is just a way-station between two other formats, each of which is required by different target systems. Even if there's a single interoperable format, then, it has to be a common or baseline representation that is technically valid and intellectually acceptable in multiple systems. The conditions for interoperability would be some combination of flexibility and shared purpose in the systems, strictness in encoding, and consistency in practice. The TEI has a role to perform at each position in this combination, but it hasn't always embraced these roles, with respect to interoperability.

2 In the P4 Guidelines, the word "interoperability" only appears twice, once in Volume 1 of the print edition in connection with Unicode, and once in Volume 2, in connection with Z39.50 (Bath Profile). On the other hand, interchange has been a core goal of the TEI from

the earliest meetings at Vassar College in 1988 where the effort to produce the TEI Guidelines began. The first principle emerging from those meetings is that

1. The guidelines are intended to provide a standard format for data interchange in humanities research. (TEI 1988)

3 In fact, TEI is an acronym with two possible expansions: it can stand for the "Text Encoding Initiative," when it refers to the activity of producing and maintaining the Guidelines, but in the title of those Guidelines, it stands for "Text Encoding and Interchange." Interchange is the subject of an entire chapter in the TEI Guidelines, as well —Chapter 30 (P4), "Rules for Interchange," the headnote to which says:

> This chapter discusses issues related almost exclusively to the use of SGML-encoded TEI documents in interchange. XML-encoded TEI documents may be safely interchanged without formality over current networks, largely without concern for any of the issues discussed here. This chapter has not therefore been revised, and will probably be withdrawn or substantially modified at the next release. (p. 647)

4 This would seem to indicate that, at least in the universe of TEI, XML has solved the problem of interchange. One significant way in which it has done so is to require Unicode for character representation. In the pre-Unicode era in which Chapter 30 was first written, character encoding was the major concern in the area of interchange especially when the interchange might take place over a network:

> Current network standards allow—indeed, require—gateway nodes to translate material passing through the gateway from one coded character set into another, when the networks joined by the gateway use different coded character sets. Since there is no universally satisfactory translation among all coded character sets in common use, the transmission character set will normally be the subset which is satisfactorily translated by the gateways encountered in transit between the sender and the receiver of the data. (p. 647)

5 TEI tackled this level of the problem by developing writing system declarations and entity references—strategies later adopted by HTML.

6 Beyond the character-encoding level of the problem, interchange advice in TEI P4 and earlier consisted mostly of recommendations to expand minimized tags and supply omitted tags. Since tag minimization and tag omission are not allowed in XML, and since Unicode is required, this chapter's advice on encoding and formatting of marked-up documents is now unnecessary. And by the same token, these features of XML take us (in theory) a step closer to being able to achieve some functional level of interoperability across text collections, at least for particular well-defined purposes. If this is true, this will be important when one wants to work at library scale with documents produced by different projects, publishers, or libraries. However, those who have tried to move from interchange to interoperability have quickly discovered that it's an extremely difficult step to take successfully.

7 In a part of the MONK project (http://www.monkproject.org) called Abbot, we did take this step successfully, and we learned some things in the process. First and foremost, we learned that even within a single project, there may be significant deviations from the norms of tagging and transcription established for that project: this ranges from apparently unmotivated variations in the application of attribute values to apparently random behavior in transcribing and encoding documentary features like line-end hyphens. For the fullest discussion of the challenges met and overcome by Abbot, see Brian L. Pytlik Zillig's essay "TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text Analysis" in *Literary and Linguistic Computing* (2009). TEI-A (for "TEI

Analytics") is a TEI customization developed for the MONK project,[1] and it is deliberately strict and stripped down. TEI-A is related to TEI Tite (Trolard 2009), a customization developed for use with keyboarding vendors. Both are intended to allow minimal variation and require minimal interpretation. As Brian notes in his LLC essay:

> If one were setting out to create a new literary text corpus for the purpose of undertaking text analysis work, the most sensible approach might be to begin with one of TEI's pre-fabricated tagsets (TEI Corpus, perhaps). In the case of the MONK project, however, we are beginning with collections that have already been tagged using diverse versions of TEI with local extensions. TEI-A is therefore designed to exploit common denominators in these texts while at the same time adding new markup for data structures useful in common analytical tasks (e.g. part-of-speech tags, lemmatizations, word tokens, and sentence markers). The goal is to create a P5-compliant format that is designed not for rendering but for analytical operations such as data mining, principal component analysis, word frequency study, and n-gram -analysis. (188-189)

8 Brian goes on to talk about the "schema harvesting" technique that is embodied in Abbot, consisting of a meta-stylesheet which is used to analyze the input text and identify TEI-A elements that are either similar or identical to the elements in the input text; the result of this analysis is a second stylesheet, automatically generated by the first, that contains XSL templates for converting the input documents into TEI-A format. Files that fail validation after running through this second stylesheet are set aside for further (human) analysis, after which stylesheet logic might be extended and the process re-run or (in rare cases) files might be edited by hand. Brian writes:

> All processes are initiated by the Abbot program in the following sequence:
> 1. Use the MonkMetaStylesheet.xsl stylesheet to read the TEI-A schema
> 2. Generate the XMLtoMonkXML.xsl stylesheet, as a result of the prior task
> 3. Convert the input collection to TEI-A
> 4. Parse the converted files against the MONK schema and log any errors
> 5. Move invalid files to a quarantine folder
>
> These steps are expressed in a sequence of Unix shell scripts, and all source files are retained in the processing sequence so that the process can be tuned, adjusted, and re-run as needed without data loss. (191)

9 Getting the world to adopt TEI-A probably isn't the answer to interoperability problems, though. As general as it is, TEI-A has a purpose in mind other than interoperability, namely analysis. A better choice might be TEI Tite, which has its purpose comfortably behind it, as soon as its texts come into existence. But it would be easy to get from one to the other. TEI Tite was developed (by Perry Trolard) as a sort of union-set of encoding practices in large libraries (Michigan, Virginia, Indiana) that contract out for substantial amounts of text-encoding. It focuses on non-controversial structural aspects of the text, and on establishing a high-quality transcription of that text.

10 Abbot, for its part, seeks to deduce similarities in the encoding practices of those entities that contributed text to the MONK project, namely ProQuest's *Early English Books Online* and *Eighteenth-Century Collections Online*, the University of North Carolina at Chapel Hill Libraries' *Documenting the American South*, the Indiana University Digital Library Program's *Wright American Fiction*, ProQuest's *Nineteenth-Century Fiction*, the University of Virgina Library's *Early American Fiction*, and Martin Mueller's Shakespeare texts. The input formats here varied quite a bit, but they included both SGML and XML with both entity references and Unicode for character encoding. As Brian notes:

> Local text collections vary not because archive maintainers are unaware of the importance of standards or interoperability but because particular local

> circumstances sometimes demand customization. The nature of the texts themselves may necessitate a custom solution, or something about the storage, delivery, or requirements for display may favor particular tags or particular structures. Local environments also require particular metadata conventions (even within the TEI header). (188)

11 Or as I put it, in a talk at the NEH back in 2007:

> Once you start to aggregate these resources and combine them in a new context and for a new purpose, you find out, in practical terms, what it means to say that that their creators really only envisioned them being processed in their original context—for example, the texts don't carry within themselves a public URL, or any form of public identifier that would allow me to return a user to the public version of that text. They often don't have a proper Doctype declaration that would identify the DTD or schema according to which they are marked up, and if they do, it usually doesn't point to a publicly accessible version of that DTD or schema. Things like entity references may be unresolvable, given only the text and not the system in which it is usually processed. The list goes on: in short, it's as though the data has suddenly found itself in Union Station in its pajamas: it is not properly dressed for its new environment. So, there's some benefit to the library, and to the long-term survivability and usefulness of their collections, or publishers' collections, to have them used in new ways, in research. (Unsworth 2007)

12 In interchange scenarios, as long as you can get from schema A to schema B by some agreed-upon intermediate step, it doesn't matter that the source texts from the two environments are incompatible in their markup. In an interoperability scenario like MONK, you are trying to bring texts from a number of different sources into a kind of lowest-common-denominator format that can then actually be used in processing.

13 In fact, though, in the MONK project the TEI-A format isn't the last stop: it's a stage in a process with more specific goals than interoperability. The TEI-A produced by Abbot is subsequently processed through Morphadorner,[2] which tokenizes, marks sentence boundaries, extracts named entities, and provides trainable part-of-speech tagging. The result of that process is fed to another program, called Prior,[3] which feeds the texts into a MySQL database—the final representation and the one that is queried for statistical information about the texts. However, we keep the TEI-A and TEI-A "morphadorned" states of the text as well, and in MONK we call on the former to provide a reading text for the user of the system at various points in the analysis process.[4]

14 I think, actually, that this is what interoperability looks like, or will look like in the future: it's a state or a stage in the processing of data, and not necessarily (perhaps not often) the final state or stage. To attain it, you have to supervise the process, mindful of the need to produce an opportunity for interoperability. If libraries and scholarly projects that require the keyboarding or OCR of texts could use a common format (like TEI Tite) as the target of that stage of the process, and if that could be saved and made available for other purposes, it would allow other projects and processes to pick up those texts and either process them in that state or process them from a predictable source format into some more heavily tagged format that supports a more specific purpose. Interoperability, I'm suggesting, is a plateau and a publication, and it's a matter of influencing the workflow for what you and others do so that it passes through that plateau and undertakes that publication. I'm not suggesting that TEI-A is necessarily the spec to use here—more likely, it would be something like TEI Tite, meant as spec for vendors and now stipulated as the output format for TEI members who wish to take advantage of the AccessTEI member benefit (a discount on keyboarding services offered by Apex

CoVantage).[5] No doubt, in most cases this output will receive further processing for particular purposes and for the local environment, but if TEI members, libraries, and publishers using specifications similar to TEI Tite could learn to think about the Tite output as having a purpose of its own, namely interoperability, that would go a long way toward solving the kinds of problems that we encountered in MONK and that are certain to be encountered by anyone else who tries to make texts from different sources work (and play) together.

15 Interoperability is not just a matter of text format, though: it's also very much a matter of license conditions. In the MONK project our final act was to present MONK to the public in two instances. The first instance[6] is available to all users: it includes about 50 million words of American literary text from North Carolina, Indiana, and Virginia, plus the Shakespeare texts. The second instance[7] is available only to users with login privileges at a CIC Institution:[8] it provides access to a corpus of 150 million words that includes licensed material from ProQuest and Cengage. Login is negotiated through InCommon, which is an Internet2 implementation of the Shibboleth authentication protocol that has been set up at each CIC institution. All of those universities license the ProQuest materials, so permission for this re-presentation of their materials was not hard to get; however, only about half of them licensed the Cengage materials, so special permission was required from Cengage to allow them all uniform access to a single instance of MONK. Thankfully, that permission was provided; otherwise, it would have been a good deal more complicated to sort out who was allowed access to what.

16 This solution to the problem of heterogeneous access to licensed material is not scalable, obviously: there isn't time for each new research project to negotiate access in the way that we did, and there's no guarantee that other publishers would agree, as these did. In this connection, "scale" is represented by the Google Books project, which aims to digitize all printed books. As of October 2009, Google would admit to having scanned 10,000,000 books (Brin 2009), but Google estimates that there are about thirteen times that many books out there (Taycher 2010), so they're far from done. The scalable solution might come out of the Google Books Settlement agreement, if a settlement is ever finalized.

17 The proposed agreement (Case No. 05 CV 8136-DC 2009), which has preliminary approval from the courts, calls for Google to set up two research centers in which public domain and copyrighted works would be available for computational research, on the condition that the use of copyrighted material is "non-consumptive" (Case No. 05 CV 8136-DC 2009, section 7.2.d). Non-consumptive research is defined in the settlement as:

> ...research in which computational analysis is performed on one or more Books, but not research in which a researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book. Categories of Non-Consumptive Research include:
> (a) Image Analysis and Text Extraction—Computational analysis of the Digitized image artifact to either improve the image (*e.g.*, de-skewing) or extracting textual or structural information from the image (*e.g.*, OCR).
> (b) Textual Analysis and Information Extraction—Automated techniques designed to extract information to understand or develop relationships among or within Books or, more generally, in the body of literature contained within the Research Corpus. This category includes tasks such as concordance development, collocation extraction, citation extraction, automated classification, entity extraction, and natural language processing.

> (c) Linguistic Analysis—Research that performs linguistic analysis over the Research Corpus to understand language, linguistic use, semantics and syntax as they evolve over time and across different genres or other classifications of Books.
> (d) Automated Translation—Research on techniques for translating works from one language to another.
> (e) Indexing and Search—Research on different techniques for indexing and search of textual content. (Case No. 05 CV 8136-DC 2009, section 1.93)

18 The uses defined in (b) and (c) would cover all of what we did in MONK, and everything I can envision as falling under the general heading of text-mining. However, the notion that you can, for example, do supervised learning in text-mining without reading or displaying substantial portions of the book or understanding its intellectual content is more than a little implausible, and the whole idea of non-consumptive research, should it survive, will need to be refined in light of actual research and research use-cases. In any case, the settlement has not been finalized and the judge under whom it was negotiated has been promoted to a higher bench, so the whole thing may start over, or the suit may be withdrawn.

19 Even if that happens, though, HathiTrust is considering proposals for a research center that would leverage their shared digital repository which was set up by many of the libraries that participate in the Google Books project (Hagedorn, York, and Levine 2009). I am involved in a HathiTrust proposal submitted jointly by Scott Poole at the University of Illinois and Beth Plale at Indiana University under consideration by the HathiTrust Executive Committee as of this writing. At this time, the HathiTrust includes 7.1 million books, about 24% (or about 1.7 million) of which are in the public domain (HathiTrust 2010). By comparison, MONK included about 1500 titles, so even the public-domain content of the HathiTrust component of the Google Books collection is over 1,000 times the size of MONK. That counts as scale.

20 Working with only that portion of the potential research corpus, you could still seriously pursue the research goals spelled out in the HathiTrust RFP:

- aggregation/distillation – "raw texts or abstracts covering particular topics or types of materials are reduced to subsets or databases of interest that can be used by one or multiple researchers"
- development of tools for research – for "textual analysis, entity extraction, aggregation of data, and the representation and analysis of results"
- collaboration – the Center must offer the ability to share processes, results, and communication with individuals and groups in a secure manner.
- Miscellaneous additional needs and concerns of researchers, e.g.
  - "The ability to include additional data."
  - "The ability to have access to both raw and pre-processed texts" (HathiTrust 2010, 7–10)

21 and complexity envisioned here will raise challenges in that area. One possible strategy for sustainability in this case would be to connect the maintenance of a research corpus, institutionally, to the maintenance of rights information. Another proposal in the Google Books Settlement that may survive even if the settlement agreement does not is the establishment of a non-profit clearinghouse for settling claims against money earned by the use of orphan works—those works that are in copyright, but for which a copyright holder cannot be located. A conservative estimate puts the number of orphan works in the Google Books collection at about 580,000 (Cairns 2009),[9] but some estimate the number in the millions. If the rights clearinghouse and the research host site were connected, the activity of the first might contribute to the sustainability of the second.

Even if that subsidy were prohibited or constrained (as it would be, under the proposed settlement), the two activities obviously need to be conducted with awareness of one another, so that it's clear what rights conditions apply to what works. And even if there's no cross-subsidy, a research center could support itself with a combination of budgeted funds in research proposals that use the resource, plus institutional support.

22 These are the bits of an emerging cyberinfrastructure for disciplines that work with text. Characteristically, they include standards, strategies, organizations (like scholarly societies), institutional structures (like libraries and perhaps publishers, as well as research and its funders), and commercial players (including at least software developers and publishers, in this case). These characteristic bits also include moments of production, transmission, storage, representation, and analysis. And because cyberinfrastructure is also a social structure, it is a process. The TEI has a leading role to play at several points in that process, including of course as a standard, but also as a standards organization that interacts with institutional structures and commercial players. TEI competes—whether it wants to or not—in intellectual and institutional ways with various other disciplines and institutional commitments.

23 In general, one area of competition is in the academic recognition of computational research into ontologies. As more and more material has been digitized, people have begun to work toward what Tim Berners-Lee and others call the "semantic web" (2001). The Semantic Web Conference is a high profile academic event, but it is also a very large and fairly commercial event, and semantic web topics are discussed not only in AI and other CS contexts, but also as the foundation of business activities. Semantics, in this case, depends on ontology, and ontology is therefore "one of the pillars of the semantic web."[10] The Text Encoding Initiative has been doing the ontology of literary and linguistic texts since 1987. TEI has an Ontology SIG, in fact, that it should probably fund to represent TEI in semantic web contexts. TEI may have been here first, but it is coming from behind in terms of institutional recognition or functional centrality in semantic web contexts, possibly for the same reason that we seemed late to arrive at the Hypertext Ball when it was first thrown, by the World Wide Web. Neither the semantic web nor the web itself is a pure and well thought-out system, and they're both over-commercialized already. But the TEI has a lot to offer both—and in fact, has offered it to the Web, the point of continuity being Michael Sperberg-McQueen, former North American Editor of the TEI, and his work for the World Wide Web Consortium on the XML standard.

24 We need to make a similarly important contribution, perhaps with more recognition, in the development of the semantic web, or at least in developing what is understood by that term. Doing this may help the TEI to track and participate in proposals for the research use of our expanding corpus of digital cultural heritage material in the form of text. By participating, we can assert the needs and the ontological views of a diverse humanities user community, and we can do that with more historical perspective and more authority than any other organization I can think of. If the TEI were to participate in such proposals, we could help to ensure that the emergent research environment is TEI-friendly, something that will serve the interests of the humanities research community. Through this participation in research proposals and in the research center, we can also contribute to the sustainability and the interoperability of a research corpus. And if TEI is part of doing that, the TEI will also be sustainable, and participation in the TEI will be increased. Simple things like reminding users of the potential interoperability of texts produced through AccessTEI, and perhaps maintaining a record of whose

institutions produced what texts with which access rights, would allow us to begin to carve out a role in the rights discovery and maintenance part of this ecology as well.

25 Finally, although we will certainly need research efforts like Abbot in order to move toward interoperability in the very large corpora of the near future, we need organizations like the TEI itself even more, and we need the TEI to have a vision and a strategy for asserting its role in the semantic web—by engaging early and often with emerging text-research centers and collections, and by promoting the potential interoperability of the materials produced through its AccessTEI service.

BIBLIOGRAPHY

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. "The Semantic Web." *Scientific American*, May 2001. http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

Brin, Sergey. 2009. "A Tale of 10,000 Books." *The Official Google Blog*, October 9. http://googleblog.blogspot.com/2009/10/tale-of-10000000-books.html.

Cairns, Michael. 2009. "580,388 Orphan Works – Give or Take." *Personanondata*, September 9. http://personanondata.blogspot.com/2009/09/580388-orphan-works-give-or-take.html.

Case No. 05 CV 8136-DC: Amended Settlement Agreement. 2009. http://www.googlebooksettlement.com/agreement.html.

Hagedorn, Kat, Jeremy York, and Melissa Levine. 2009. "Call for Proposal to Develop a HathiTrust Research Center." *HathiTrust*. http://www.hathitrust.org/documents/hathitrust-research-center-rfp.pdf.

HathiTrust. 2010. "Welcome to the Shared Digital Future." *HathiTrust*. http://www.hathitrust.org/about.

Ontology. 2010. Semantic Web wiki. http://semanticweb.org/wiki/Ontology.

Taycher, Leonid. 2010. "Books of the World: Stand Up and Be Counted! All 129,864,880 of You." *Inside Google Books*, August 5. http://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html.

Text Encoding Initiative. 1988. "Design Principles for Text Encoding Guidelines." http://www.tei-c.org/Vault/ED/edp01.htm.

Trolard, Perry. 2009. "TEI Tite—A Recommendation for Off-Site Text Encoding." Version 1.0. Text Encoding Initiative Consortium. http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

Unsworth, John. 2007. "Digital Humanities Centers as Cyberinfrastructure." http://www3.isrl.illinois.edu/~unsworth/dhcs.html.

Zillig, Pytlik, and Brian L. 2009. "TEI Analytics: Converting Documents into a TEI Format for Cross-Collection Text Analysis." *Literary and Linguistic Computing* 24, no. 2: 187–192. doi:10.1093/llc/fqp005.

NOTES

1. The TEI-A schema can be retrieved at http://www.monkproject.org/downloads/texts/schemata.gz and documentation is available online at http://segonku.unl.edu/teianalytics/TEIAnalytics.html.
2. See http://morphadorner.northwestern.edu/.
3. See http://monkproject.org/docs/monk-datastore-doc/doc-files/prior.html.
4. With respect to the need to read, see the discussion below, on the subject of non-consumptive research.
5. For more information, see http://www.apexcovantage.com/content-solutions/accessTEI-digitization.asp.
6. See http://monkpublic.library.illinois.edu/monkmiddleware/public/index.html.
7. See https://monk.library.illinois.edu/secure/mainMenu.html.
8. For a list of CIC institutions, see http://www.cic.net/home/AboutCIC/CICUniversities.aspx.
9. See http://personanondata.blogspot.com/2009/09/580388-orphan-works-give-or-take.html.
10. In the Semantic Web wiki entry on Ontology (Ontology 2010), we learn that there is no universally accepted definition of ontology, raising the specter of recursion.

ABSTRACTS

This essay will address the challenges and possibilities presented to the Text Encoding Initiative, particularly in the area of interoperability, by the very large text collections (on the order of millions of volumes) being made available for computational work in environments where the texts can be reprocessed into new representations, in order to be manipulated with analytical tools. It will also consider TEI's potential role in the design of these environments, these representations, and these tools. The argument of the piece is that interoperability is a process as well as a state, that it requires mechanisms that would sustain it, and that TEI is one of those mechanisms.

INDEX

Keywords: interchange, interoperability, text-mining

AUTHOR

JOHN UNSWORTH

john.m.unsworth@gmail.com
Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

John Unsworth is Dean and Professor at University of Illinois' Graduate School of Library and Information Science (GSLIS) and Director of the Illinois Informatics Institute.
He organized, incorporated, and chaired the Text Encoding Initiative Consortium, co-chaired the Modern Language Association's Committee on Scholarly Editions, and served as President of the Association for Computers and the Humanities and later as chair of the steering committee for the Alliance of Digital Humanities Organizations.
During the previous ten years, from 1993-2003, he served as the first Director of the Institute for Advanced Technology in the Humanities (IATH), and a faculty member in the English Department, at the University of Virginia. For his work at IATH, he received the 2005 Richard W. Lyman Award from the National Humanities Center. He chaired the national commission that produced Our Cultural Commonwealth, the 2006 report on Cyberinfrastructure for Humanities and Social Science, on behalf of the American Council of Learned Societies.
He has also published widely on the topic of electronic scholarship, as well as co-directing one of nine national partnerships in the Library of Congress's National Digital Information Infrastructure Preservation Program, and securing grants from the National Endowment for the Humanities, the National Science Foundation, the Getty Grant Program, IBM, Sun, the Andrew W. Mellon Foundation, and others.