Journal of the Text Encoding Initiative

Issue 5 (2013)
TEI Infrastructures

Lou Burnard

The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure

Electronic reference
Lou Burnard, « The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure », *Journal of the Text Encoding Initiative* [Online], Issue 5 | 2013, Online since 21 June 2013, connection on 25 July 2013. URL : http://jtei.revues.org/811 ; DOI : 10.4000/jtei.811

Publisher: Text Encoding Initiative Consortium
http://jtei.revues.org
http://www.revues.org

Document available online on:
http://jtei.revues.org/811
Document automatically generated on 25 July 2013.

Lou Burnard

The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure

1. In the Beginning

1 The Text Encoding Initiative was born into quite a different world from that of today. In 1987, there was no such thing as the World Wide Web, and construction of the tunnel beneath the English Channel had only just begun. A major political power called the Union of Soviet Socialist Republics still existed, while in the UK, Margaret Thatcher's government had just been reelected for a third time, and in the US the Senate rejected for the first (and so far only) time a presidential nomination to the Supreme Court. In academic life, it was still (just about) possible to finance an undergraduate degree on the basis of government grants. A typical "home computer" cost about 1,500 pounds in the UK, had an Intel 80286 processor and up to 640 Kb of memory, with maybe up to 50 Mb of storage on its internal hard disk, and probably ran some version of Microsoft's ubiquitous MS-DOS, unless of course it was a Macintosh. New machines were beginning to appear on the market, some of them with nearly enough memory and processing power to run Microsoft's new Windows operating system, or IBM's optimistically named "OS/2," also launched in this year. And meanwhile in another part of the forest Steve Jobs was busy imagining the Next computer, which would run something like Unix, but with a Windowing interface. However, any serious computing would still be done on your departmental minicomputer (perhaps a VAX or a PDP) or your institutional "mainframe," as the massive energy-hungry arrays of transistors and magnetic storage systems sold by such companies as IBM, Univac, Burroughs, ICL, or Control Data were known.

2 At the same time, much of the work done on those massive machines looks quite familiar today. The process of digitization of the office environment had already begun in some scientific disciplines with software such as TeX, Scribe, or tRoff becoming dominant in the production and dissemination of research articles and documentation. The tide of personal computers able (allegedly) to close the gap between the writer and the publisher would soon engulf us as surely as Microsoft Word would replace the seemingly unstoppable Word Perfect 4.2 (released in 1986). Such neologisms as "desktop publishing," "expert systems," and "digital resources" began to appear in serious academic journals, as well as dominating the discourse of fledgling online communities such as the Humanist discussion list launched in 1987. The Internet already existed, as did many theories about how it might be used "hypertextually," though the World Wide Web was still barely an idea. Both in the research community and, increasingly, beyond it, the goals of corpus linguistics and artificial intelligence alike had established a need to work on large-scale digitized textual resources just as the technologies to support such work were beginning to appear. Launching a major revision of the Oxford English Dictionary, Oxford University Press for the first time proposed to do so using computational methods. Text-based disciplines of all kinds were beginning to imagine the possibilities offered by computationally tractable corpora of source texts created by such projects as the Thesaurus Linguae Graecae, the Trésor de la Langue Française, or the Dictionary of Old English. And, given the rapidity with which one storage technology was already replacing another, new discourses concerning the best methods of guaranteeing long-term access to newly created digital resources, and about the necessities for open access and platform-independent formats alike, were beginning to be heard.

3 Symptomatic of that discourse were two key meetings held in 1987: one, in April, organized by a group of practicing historians in Europe, debated the possibility of establishing some kind of consensual standardization for the encoding of primary historical source data in computer-readable form. A paper prepared for that workshop by Manfred Thaller, and subsequently published in the influential journal *Historical Social Science*, together with a collected

proceedings volume edited by the French historian Jean-Philippe Genet (see, among others, Thaller 1986; Genet 1988), may have strengthened the case put forward to the US National Endowment for the Humanities for the funding of a larger international workshop on the issue; or perhaps they just spurred on the organizers of that workshop. But in either case, the majority of the European institutions represented at one workshop reappeared at the other, along with many other scholars from North America and elsewhere in the world. The TEI's foundational event[1] was held at Vassar College, Poughkeepsie, New York, immediately after the joint annual conference in Toronto of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, twin venerable precursors of today's Alliance of Digital Humanities Organizations.

4 However, the purpose of this paper is not simply to chronicle the history or foundational myth of the TEI, but rather to try to answer a more interesting question. If the TEI is so very old—preceding as it does the Web, the DVD, and widespread use of technologies such as the portable phone, cable television, and Microsoft Word—and given that computer related technologies are hardly renowned for their longevity, how is it that the TEI is still with us, and still occupying a significant position in the world of research, after nearly thirty years?

2. Transformations of The Text Encoding Initiative

5 As first conceived, the Text Encoding Initiative was an international research project intended to define a kind of digital demotic, as indicated by its alternative expansion *Text Encoding for Interchange*. In a world dominated by mutually incompatible formats, each computer manufacturer could impose its own conventions for the structuring and representation of textual data. This was a world in which some computers worked in EBCDIC and others in ASCII, where even the number of bits in a byte could vary between 6, 8, and 16, and where, as a consequence, there were serious technical obstacles even to the simple transfer of data files from one machine to another, to say nothing of the difficulties posed by mutually incompatible and proprietary file formats. Nevertheless the TEI announced that it would facilitate the creation, exchange, and integration of textual data in machine-readable form, for all kinds of texts, in every human language, from every historical or social context. In the world before the Web, these were ambitious goals. But this quixotic ambition was further compounded by the TEI's declared goal of making its proposals accessible both to the novice, looking for guidance on well-established best practice, and to the expert, seeking to establish new practice in response to new research goals. It was this latter objective which, in retrospect, gives the TEI its distinctive nature, and which largely distinguishes it from other standardization efforts, both those now forgotten, and those currently entrenched or in formation.

6 In its earliest stages, the TEI was a creature of a time of transition, when the notion of "humanities computing" was just beginning to invent itself as a form of interdiscipline, a space within which information specialists, computer scientists, and traditional humanists might meet, if only to trade secret handshakes and suspicious glances; it was a period in which the notion of the disciplinarity (or otherwise) of that interchange was not yet more important than its simple existence. The founding parents of the TEI were, from the standpoint of traditional academic life, a very mixed bunch of people, who had for the most part foregone the safety of traditional career paths for the excitement of transgressing disciplinary boundaries. Many of those who met that snowy weekend in Poughkeepsie came from research teams or institutions on the fringes of traditional scholarship, owing allegiances to computer science, linguistics, philology, lexicography, or literary studies in many languages, but not centrally placed within any of those disciplines. What united them was an expertise in the creation and management of digital text, a vision of its future importance to the traditional disciplines, and a concern that lack of standardization and commercial pressure might prevent the realization of that vision.

7 The outcome of the conference was a set of eminently practical recommendations as to how an extensible set of guidelines consistent with the goal of a universal text-encoding scheme might be achieved. These "Poughkeepsie Principles," with commentary focusing on implementation, have long been available from the TEI's website[2] and we do not discuss them further here, though they repay reading for anyone curious about the theory underlying the shape and

content of the TEI scheme. Instead we propose to focus more on the organizational and managerial issues which have determined its evolution.

8 It is instructive to compare the organizational structures of the TEI during its first decade of existence with that which has come into being during its second decade. Initially conceived as a research project answerable to a self-selecting group of experts, the TEI of the 1990s manifested a typically centralized structure, in which all the work is done by a small number of people under the control of a small executive authority (see fig. 1). The TEI of the present decade has a more distributed structure in which the task of maintaining and developing the Guidelines is the responsibility of many different people drawn from a loosely defined community (see fig. 2).

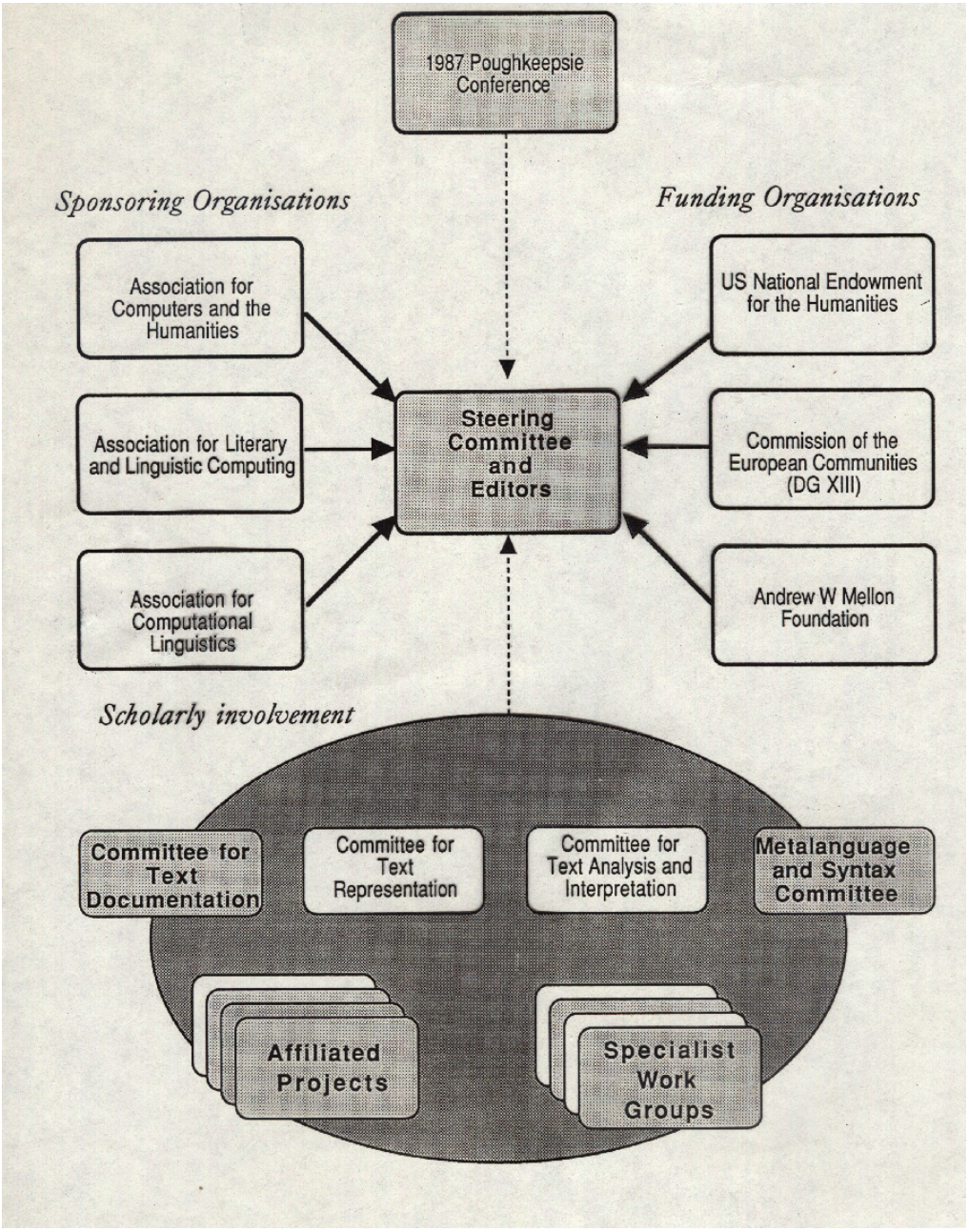

Figure 1. TEI organizational structure, 1991

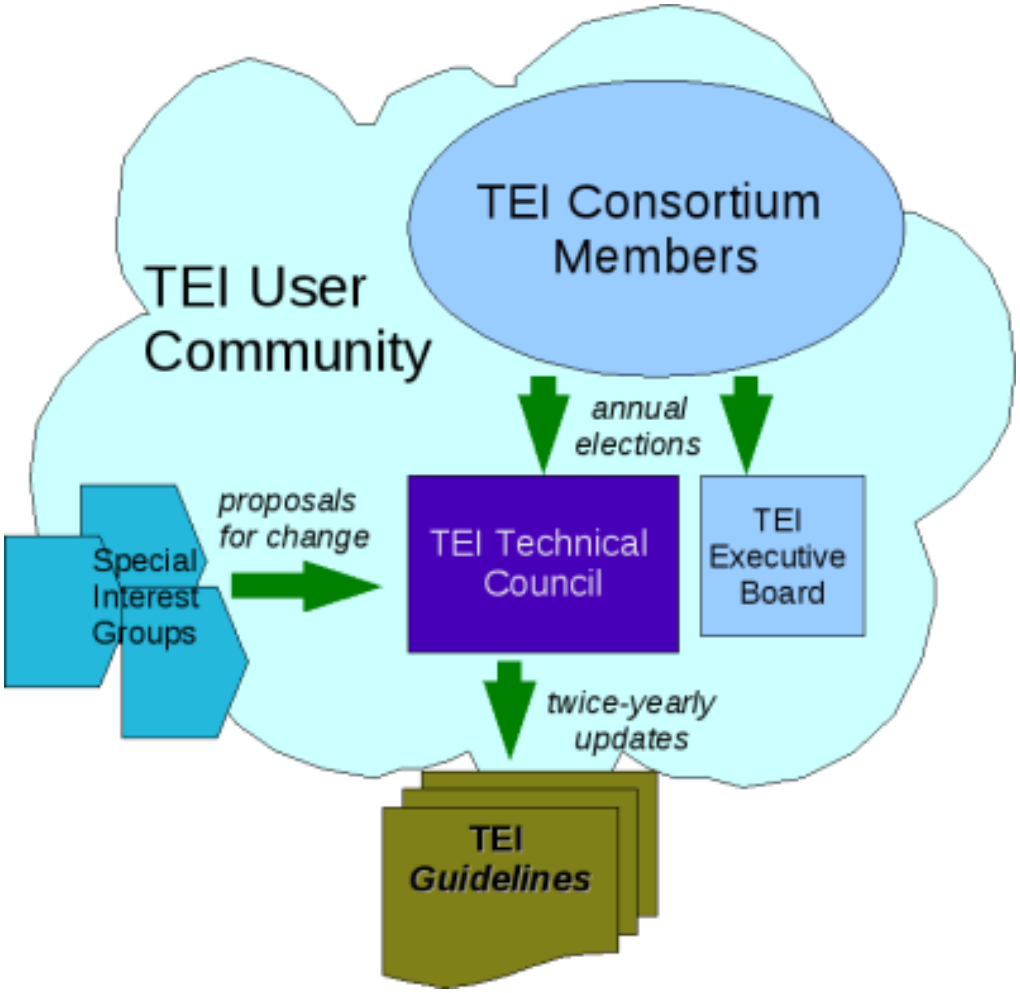

Figure 2. TEI organizational structure, 2012

9 That is not, of course, to imply that the original TEI editors worked in isolation from the many scholarly communities that their work was intended to benefit. On the contrary, the organization of work required the editors to spend much of their time extracting proposals and requirements from a very wide range of experts, which could then be subsequently organized into a coherent whole for appraisal and ratification by further groups representing those communities. In all, nearly 200 experts from both sides of the Atlantic gave their time and energy to codifying their own practice in such a way as to facilitate its integration with that of others.

10 During the first funding cycle (1992–93), the TEI's work was carried out by four appointed committees. The Text Documentation Committee, populated by documentalists and librarians, produced recommendations that eventually became the TEI Header. The Metalanguage and Syntax Committee, populated by computer scientists, made recommendations on the formal metalanguage in which the TEI outcomes should be formulated. The other two committees divided the universe of possible objects for text markup between them, along rather loosely defined conceptual lines: one committee, Text Representation, was supposed to identify the "significant particularities" of written texts, specifically those which needed to be made explicit in a marked-up document; the other, Text Analysis and Interpretation, was supposed to focus on the encoding of "added value" brought to a text by an analytic tool, such as a parser or (indeed) a careful reader. This opposition between **analysis** and **representation** typifies much contemporary debate about the proper function of markup, and indeed the process of reading itself.

11 The Text Documentation committee may be credited with the idea, still useful today, that the TEI Header is intended to act as a **primary source of information** for a digital resource, in the rather specialist sense that the librarian community uses that phrase. The title page of a printed book may not correspond exactly with a catalogue entry for that book, but cannot be ignored

during the production of one. Given that most digital resources were going to be produced by non-cataloguers, the committee argued that the TEI Header should make it possible for the non-expert to record whatever information they deemed useful in such as way as to simplify, but not to replace, the task of the subsequent cataloguer. This idea perhaps underlies much that professional cataloguers find frustrating in the TEI header today.

12 The choice of SGML[3] as a vehicle for expression of the TEI's recommendations was a fairly obvious one for the Metalanguage Committee. This committee also established the principle that the use of this ISO standard was conditional, and that the TEI itself should as far as possible be independent of any particular metalanguage syntax; the wisdom of this decision and the consequent system architecture became apparent when, several years later, the task of re-expressing the TEI in a new metalanguage called XML was taken on.

13 The Text Representation committee began by consolidating current encoding practice across a number of related disciplines to provide the essentials of the TEI scheme: the basic components of textual objects, viewed as hierarchically organized containers, in which identifiable, possibly internally structured, components such as bibliographic references, highlighted phrases, or proper names float in a kind of "soup" of plain text. To these, the committee added recommendations for markup of textual variance and hypertextual links, amongst other topics reflecting its diverse membership. A recurrent theme in its discussions was the desire to encode an idealized version of the text itself, independently of its realization in a particular source, while at the same time wanting to preserve what was significant in that source. This debate found expression in a number of published scientific articles[4]; it also probably underlies the TEI's notorious tendency to provide both cake and the eating of it (or, as they say in France, both the butter and the money for the butter). Nevertheless, as critics were quick to point out, whenever it became hard or impossible to give equal time to both the presentational or visual and the analytic or structural properties of a text, this committee, and therefore the TEI, consistently came down on the side of the latter. Many of those at Poughkeepsie felt that contemporary digital publishing systems, with their emphasis on "camera-ready" copy, engendered a distraction from the true business of scholarship, as had been cogently argued in a highly influential article published in the *Communications of the ACM* shortly before the Poughkeepsie Conference.[5]

14 The Analysis and Interpretation Committee began with the rather ambitious goal of providing ways of encoding in a normative way the full range of linguistic analysis. In an early working paper, Langendoen (1990) notes that amongst many other topics, it would need to deal with:

- Underspecification, uncertainty, multiple hierarchies
- Phonology and prosody
- Morphology and word-level tagging
- Higher-level syntactic analysis
- Structural ambiguity
- Anaphora and Deixis
- Idioms
- Figures of speech (not for this cycle)
- etc. etc.

15 If this seems a touch ambitious, it should not be forgotten that this was the period during which Europe's Language Engineering industries were born, and that part of the motive behind the EU's funding of the TEI was precisely to formulate recommendations about specific linguistic categories for the use of that industry.

16 The linguists represented on the A&I Committee did not, however, belong to a school in which such concepts as "noun" or "verb" were unproblematic: instead they formulated a more abstract system, a kind of metamodel of linguistic annotation based on feature structure theory,[6] which could be used to represent any kind of linguistic (or indeed non-linguistic) annotation. This model eventually achieved wide acceptance in the language engineering field and is, so far, the only part of the TEI to have been formally adopted as an ISO standard,[7] but its power and generality of application were not immediately appreciated by the computational linguist who

just wanted to build a simple parser. To meet such needs, the committee also proposed a range of generic segmentation and synchronization mechanisms of various kinds.

17 Inevitably, the encoding needs and mechanisms identified by these two committees frequently coincided in all but name. Despite the vigorous application of Ockham's razor, the TEI often found itself proposing a range of different ways of encoding linguistically motivated segmentation and of recording analytic judgments, with more or less internal structure, reflecting the diversity of the user community. This process continued further with the appointment, during the second funding cycle, of a number of specialist working groups, charged with testing the TEI proposals within specific research areas and identifying any gaps. As with the larger committees of the first phase, the members of these specialist working groups were nominated by experts in the field, and their meetings funded on the understanding that they would produce detailed proposals for integration into the Guidelines, the first version of which—a modest 275-page volume known as P1—had been distributed in print form in 1990. In the event, some of these working groups were content to assess the applicability of that draft to their own area of expertise, or to make general comments about their usability, but in many cases this work resulted in significant expansion of what was already proposed, and in a few cases substantial entirely new material was developed, extending the scope and coverage of the Guidelines considerably. One working group, for example, produced proposals for the encoding of transcribed speech (which had not featured at all in P1), while another produced a complex tagset for the markup of existing print dictionaries. Meanwhile, the TEI editors revised chapter by chapter on the basis of the feedback received. Rather than delay publication until the whole work was complete, the new version (P2) was distributed in fascicle form during 1992, as each new chapter took shape.

18 The chairs of all the working groups and other significant contributors were invited to sign off on the editorial process at a four-day Technical Review Meeting, held at Eynsham Hall near Oxford in May of 1993. After further revisions consequent on that review, the first complete version of the TEI Guidelines, known as P3, appeared at the international SGML conference held in Montreux in May 1994. TEI P3 took the form of two substantial green volumes, making up a 1300-page reference manual documenting and defining some 600 SGML elements which could be combined and modified in a variety of ways to create specific SGML document type definitions (DTDs) for particular purposes. In 1994 the first edition of TEI Lite also appeared. This was a simplified subset of the whole scheme, which was originally written to accompany the first of many training workshops held in Chicago in December of the same year. This Chicago "metaworkshop" was conceived as a training session for trainers, and did much to establish how the TEI is generally introduced to a novice audience.

19 In the five years that followed, the proposals of the TEI established themselves, without benefit of grant funding or centralized management, as a necessary part of the intellectual infrastructure. In the US, several influential university libraries began to deliver online versions of set texts and to host the textual databases needed for research. Training in using the TEI for basic encoding appeared as a module on specialist library courses; research projects in their grant submissions would routinely add the phrase "we will follow the TEI recommendations" (often followed by "but . . .") ; a new generation of research assistants found it necessary to understand the difference between <div1> and , or why there was a <pb> element but no <page> element. A detailed history of the way in which this happened is hard to write, so rapidly did the TEI become a part of the humanities computing ecosystem at this period. To some extent we may speculate that there was no serious alternative for people who found inadequate the purely presentational focus and lack of formality in the HTML of the time, or that in a world where the technical infrastructure was rapidly changing, something so clearly grounded in traditional scholarship, and so clearly underpinned by theoretical rigor, offered at least the promise of a long-term return on the serious costs of investing in the digital future.

20 In his preface to a collection of working papers about the TEI published in 1995, Charles Goldfarb, inventor of the SGML standard, remarked with typical prescience, "The vaunted 'information superhighway' would hardly be worth travelling if the landscape were dominated

by industrial parks, office buildings, and shopping malls. Thanks to the Text Encoding Initiative, there will be museums, libraries, theaters, and universities as well."[8] The TEI Recommendations were endorsed by the US National Endowment for the Humanities, the UK's Arts and Humanities Research Board, the Modern Language Association, the European Union's Expert Advisory Group for Language Engineering Standards, and many other agencies that funded or promoted digital library and electronic text projects. They had become part of the intellectual infrastructure of the day.

21 For the purposes of this article, however, the key point to note is that by the end of the twentieth century, although everywhere cited as if it had some kind of formal existence, the TEI was no longer under the active care or management of anyone. Many of those most responsible for its original development had moved on to other challenges, most notably the principal editor Michael Sperberg-McQueen, who in 1996 had been appointed co-editor of the World Wide Web's emerging standard XML. At a conference organized to celebrate the TEI's tenth anniversary at Brown University in 1997, there was cake and an a cappella choir, but no consensus as to the organizational way forward. Did the TEI need any kind of formal management? Was anyone interested in maintaining it into the future, or was it a great idea to be fixed in stone, until consensus emerged that its time had now passed?

22 Ideas about an appropriate structure for continued maintenance and development of the TEI had been discussed within the original Steering Committee of the project over two or more years without practical result; the need for some kind of scientific council able to make decisions about the future technical development of the Guidelines had also been foreseen, and indeed a body not unlike the TEI's current Technical Council had held two meetings. As a result of their work, a new and final version of P3 appeared online, with a small number of egregious errors corrected, and the addition of one new element. It was already clear, however, that a very much more extensive program of work was needed to bring P3 up to date. The difficulty was in setting up a financial and organizational infrastructure within which that extensive program of work might be undertaken.

23 The task was eventually accomplished after what might be caricatured as a kind of "management buyout." At a meeting held at King's College London in January 1999, representatives of the three learned societies in whose name the TEI Guidelines had originally been published, and representatives of four key academic institutions at which the TEI was widely used and promoted, met and agreed to transfer ownership and management of the Guidelines to a new entity: a non-profit international membership organization called the TEI Consortium, the constitution and structure of which were laid out in a detailed proposal submitted to the meeting by representatives from the University of Virginia and the University of Bergen (see TEI 1999).

24 The goal of the TEI Consortium would be to establish a permanent home for the TEI as a democratically constituted, academically and economically independent, self-sustaining, nonprofit organization. It was clear that the TEI Guidelines had a major role to play in the application of new XML-based standards that were driving the development of text-processing software, search engines, web browsers, and indeed the Web in general. To re-express the TEI's SGML schema fragments as XML was not particularly difficult, but the TEI also needed to adapt to a technical environment completely transformed by the rise of the Web. The first act of the Consortium was to publish a new version of the Guidelines, known as TEI P4. This was a largely unmodified version of TEI P3, except for its re-expression using XML syntax rather than SGML. At the same time, the new Consortium declared its intention of embarking on a major revision of the entire Guidelines which would bring them up to date, extending and improving their coverage, and also radically transforming the way in which they would be maintained and developed in the future.

25 In a presentation given at the first annual meeting of the members of the new TEI Consortium, held in Pisa in November 2001, Michael Sperberg-McQueen (2001) itemized the things that in his view the "old" TEI had done right. These included some interesting technical aspects: for example, the TEI interchange format, a subset of SGML that corresponded very closely with what subsequently became XML; the TEI extended pointer syntax, which approximated

to what became XPointer and XPath; the fact that DTDs are not written, but generated using an additional layer of abstraction which integrates the formal and informal or documentary aspects of the architecture—the ODD system. But Sperberg-McQueen also underlined the importance of the work invested "to earn community buy-in," in his phrase, thus re-asserting the essentially community-driven aspects of the project.

26 At the same meeting, Lou Burnard and Syd Bauman listed a dozen or more possible areas in which the Guidelines needed enhancement. These included topics evidently in need of update because the world had moved on, such as character encoding following the definition of Unicode, and also areas where other research communities had already been active, independently of (but perhaps inspired by) the TEI, such as the detailed proposals for the encoding of manuscript descriptions arising out of the European MASTER project. But for the most part they presented an ambitious shopping list of items in which they felt work was needed, with no promise that the work would actually be done without support from the user community.

27 Beyond simply adding yet more elements and more recommendations, it was clear that the TEI's internal architecture needed substantial revision. The move to XML meant that many new technologies were now available, and many assumptions about document processing which were entirely valid in 1994 needed to be rethought for the new digital world, of which the TEI wished to become a good citizen. In his report to the members in 2005, Christian Wittern, as chair of the Council, likened the process to the restoration of a venerable piece of architecture: "When I grew up in the small southwest German town of Tübingen, conservation and re-creation of the medieval town centre was a major project there. This resulted in many a half-timbered house being completely redone from within without actually taking down the structure, but resulting in a totally new interior . . . . Something similar is happening at the moment with P5 . . ."[9]

28 During the TEI's first two development phases, working groups had communicated extensively by e-mail, circulating and discussing drafts at a leisurely pace punctuated by expensive if productive face-to-face meetings funded by the TEI. The working environment for this new phase was rather different: for most working groups, and most notably the TEI Council itself, telephone conferencing became the norm, with only very occasional face- to-face meetings funded by the TEI. Nevertheless, when the first version of TEI P5 was finally published in 2007, much of its new intellectual content was the result of working groups[10] independent of the TEI Consortium. The new chapters on manuscript description, on the description of named entities, on the documentation of characters and glyphs outside Unicode, and on the evolution of the ODD system were all created in the same way: a group of enthusiastic and largely self-selected experts worked up and tested proposals, which were then consolidated within the new Guidelines for ratification by the TEI Council. For the most part, the work of such groups was funded by specific grants, by institutional benevolence, or by individuals working on their own time. As of TEI P5, the TEI consciously transformed itself into a classic open source project, not only making all of its inner workings openly accessible to the public at large (or at least those parts of the public unintimidated by the Sourceforge user interface) but also becoming reliant on community input both to define and to execute all of its future development.

29 For example, at release 1.2.0 of the TEI P5 Guidelines, a substantial set of proposals was introduced to support encoders wishing to represent the visual appearance or physical makeup of source documents, to align portions of a digitized page image with exact transcriptions of the text on it, to assign parts of a transcription to documented stages in the evolution of a text, and a number of other facilities typifying the discipline of documentary editing, or *édition génétique*. These extensions did not come about because the TEI Consortium identified a need for them, applied for a grant, chartered a working group, and managed the process. On the contrary, they came into being because a number of existing TEI users, dissatisfied with the current state of the TEI, identified a need, obtained funding for a wide-ranging consultative exercise, formulated proposals, and then worked with the TEI Council to ensure their eventual inclusion into the Guidelines. This is not an isolated example, but typifies the way in which

the TEI has now become better able to evolve in response to the changing priorities of the research community.

30 That plasticity has also been facilitated by a number of technical developments, in particular in the evolution of the current TEI processing architecture. Even in its original SGML manifestation, the TEI was conceived as a modular system, which could be customized to suit the needs of particular research communities, enabling them to define, for example, a schema containing all and only the elements of importance to them, to add new elements, or to modify existing ones. However, to carry out such modifications successfully required a degree of technical knowledge significantly beyond that of the average digital humanist, even with the availability of a web interface (the so-called Pizza Chef) designed to simplify the task. With the shift to XML, the creation of such tools became both simpler and more essential. With the advent of TEI P5 in particular, users of the TEI are increasingly expected to desire to engage with the full sophistication of the TEI architecture rather than to rely on precompiled one-size-fits-all subsets such as TEI Lite. This in turn has led to the development of more sophisticated schema generation tools such as Roma, which greatly reduce the complexity of developing a customization.

31 The TEI Guidelines are, of course, written and maintained as a large TEI-conformant document, from which schemas in various formal languages (RELAXNG, DTD, and W3C Schema) and documentation in various formats, ranging currently from PDF to EPUB, are all automatically generated using a suite of XSLT stylesheets developed and maintained by Sebastian Rahtz. Although developed primarily for the TEI's own internal needs, this same suite of XSLT stylesheets is sufficiently generically designed that it has come to be regarded as an integral part of the TEI—even though the TEI was originally designed as an abstraction independent of any particular processing model. The free availability of this suite of stylesheets, and its necessary maintenance in tandem with the Guidelines themselves, has greatly facilitated the development of TEI-aware systems by a new generation of web developers and programmers. To take two particular examples: oXygen, a popular commercial XML development tool, uses them to offer on-the-fly visualization of any TEI document; and the Oxgarage web application packages them to provide basic conversions (in both directions) between many different formats, including several popular word-processing formats and TEI XML.

32 Each new version of TEI P5 (since 2007, there has been a new release of P5 twice a year) is the consequence of an ongoing revision process, in which outstanding errors or feature requests submitted by the general public are considered and acted upon by the Council. As with many other open-source projects, the number of different people active at any one time is relatively small, but the essential point is that the TEI itself is not the only agency managing the review or proposing extensions. A TEI orthodoxy undoubtedly exists, but it does not control the evolution of the Guidelines, and is constantly under critical review by the user community, via mechanisms such as the TEI discussion list, the Sourceforge Trackers, and debate on the TEI Council list, all of which are publicly accessible. A quick glance at the Sourceforge site shows that this activity is increasing: between 2005 and 2010, visits to the site averaged a million page impressions per year, while between 2010 and 2013, the figure is nearer four million a year. Looking at feature requests and bug reports alone, in 2011 the Council reviewed a total of 170 tickets; in 2012 the number increased to 245.[11] The annual meeting of the TEI membership, initially a legal obligation for the TEI Consortium, has of recent years been transformed into an open TEI Conference, showcasing current debate about the use of the TEI within the digital humanities research community, much of which eventually finds its way to a larger audience via journals such as the *Journal of the Text Encoding Initiative*.

3. TEI Today

33 There are, it has been noted, two kinds of unsuccessful standardization effort. One kind fails because it is based on a theory which is not yet sufficiently mature for widespread use. The other kind fails because it addresses a community which has not yet achieved consensus and within which any proposal therefore seems alien to at least some of its intended beneficiaries.

Today's TEI is engineered to avoid both kinds of problem. An immature theory—for example a new element definition—can be incorporated into the TEI model without perturbing the rest: that is one of many advantages inherent in the TEI architecture. Hence, it is easy for ideas developed for the benefit of one project to be made visible to a wider community, tested, evaluated, and, if appropriate, eventually included in the standard. The notion of TEI conformance now forming part of the TEI incorporates precisely this model: conformance is defined as expressing clearly which parts of a model are to be understood in the TEI sense, and which parts are not. This does not preclude the addition of new features or the modification of old ones (mutations in the evolutionary sense); it requires only that they be clearly identified as such.

34 As to the problems of diversity within the target audience, the TEI has always sought to be a truly international enterprise. In its first phase, this desire was manifested by the careful geographical balance of working groups membership; more recently it has been shown in the support for multilinguality built into the Guidelines themselves. More significantly perhaps, the Guidelines today are hospitable to other standards. Where an existing XML vocabulary such as SVG or MathML exists, the TEI schema generally prefers to incorporate it (within its own namespace) rather than to reinvent it. And, at the conceptual level, there is scope for reexpressing the semantics of any TEI element set using other conceptual formalisms (for example, the CIDOC Conceptual Reference Model has been used to provide a mapping for almost all TEI elements).

35 For longevity, however, a standards initiative must also address the need for financial and administrative support. The TEI has increasingly adopted an open-source–style business model, in which its users are the key agents providing that support. The frontiers of the TEI user community are, however, both ill-defined and large. Within the community, to date, a sufficiently large number of institutions have proved willing to provide small amounts of regular funding to keep the infrastructure itself in existence. Opinions differ notably on either side of the Atlantic as to whether that regular funding should be provided by individuals or by collectivities, and consequently how its continuation should best be ensured or its functions promoted; ensuring that continuity is the function of the TEI's Board of Directors, who are elected by the members of the Consortium, with responsibility to them and also the (currently non-voting) subscribers.

36 Fortunately, at least as far as concerns the maintenance and development of its core technical deliverables, the TEI organizational model is increasingly independent of funding. Final editorial decisions as to what does or does not change in the Guidelines are taken by an elected body of experts, the TEI Council, but this is carried out in response to reports of bugs and suggestions for change which may come from individuals, from special interest groups, or from working groups specially chartered by the TEI or others, and which may have members even beyond the TEI user community. Although there are individuals who sit on both the TEI Board and the TEI Council to ensure good communication between the two, it is clear that the TEI needs both administrative and technical expertise, and that there is an advantage in providing distinct forums for the two.

37 Arguably, with this change the TEI ceased being an academic research project and became something different. To talk of it as a "community-based research project" (to use the officially sanctioned phrase) conveys something of that difference but also obscures a part of it. Research projects come to an end: they have a goal, which may or may not be achieved, and a fixed term. The TEI, however, has transformed itself into an open-ended activity, which will continue to be usable for as long as it is felt to be useful. Its strengths are enhanced by its community of users; its weaknesses can be corrected by them. In a strictly Darwinian sense, the TEI has evolved by fostering and profiting from mutations that are considered beneficial to the user community, while ignoring those which are not. Because it is felt to be useful to so many, and because its products underpin so much of current research activity, it seems not unreasonable to regard it as constituting in itself a research infrastructure. If nothing else, its organizational model seems highly appropriate for all such infrastructural institutions or initiatives.

Bibliography

Burnard, Lou D. 1988. "Report on Workshop on Text Encoding Guidelines." *Literary and Linguistic Computing* 3(2): 131–33. doi: 10.1093/llc/3.2.131.

Coombs, James H., Allen H. Renear, and Steven J. DeRose. 1987. "Markup Systems and The Future of Scholarly Text Processing." *Communications of the ACM* 30(11): 933–47. doi:10.1145/32206.32209.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. 1990. "What is Text, Really?" *Journal of Computing in Higher Education* 1(2): 3–26. doi:10.1007/BF02941632.

Genet, Jean-Philippe, ed. 1988. *Standardisation et échange des bases de données historiques: actes de la troisième table ronde internationale, Paris, 15–16 mai 1987*. Paris: Editions du CNRS.

Ide, Nancy, and Jean Véronis, eds. 1995. *Text Encoding Initiative: Background and Context*. Dordrecht: Kluwer.

Langendoen, Terry. 1990. Plan of Action for Remaining Work of A and I Committee. http://www.tei-c.org/Vault/AI/air03.gml.

Langendoen, D. Terence, and Gary F. Simons. 1995. "Rationale for the TEI Recommendations for Feature-Structure Markup." In Ide and Véronis 1995, 191–210.

Sperberg-McQueen, C. M. 2001. "The TEI is Dead: Long Live the TEI." Paper presented at the TEI Members Meeting, Pisa, Italy, November 2011. http://www.tei-c.org/Membership/Meetings/2001/tei1116alt.html.

TEI (Text Encoding Initiative). 1988. Design Principles for Text Encoding Guidelines. TEI ED P1. Last revised January 9, 1990. http://www.tei-c.org/Vault/ED/edp01.htm.

———. 1999. An Agreement to Establish a Consortium for the Maintenance of the Text Encoding Initiative. http://www.tei-c.org/About/consortium.html.

———. 2005. TCR03: Report of the TEI Council to the Members Meeting. http://www.tei-c.org/Activities/Council/Reports/tcr03.xml.

Thaller, Manfred. 1986. "A Draft Proposal for a Standard for the Coding of Machine Readable Sources." *Historical Social Research/Historische Sozialforschung* 40: 3–46, repr. in *Modelling Historical Data*, ed. Daniel Greenstein, Halbgraue Reihe zur historischen Fachinformatik, A11 (St. Katharinen: Max-Planck-Institut für Geschichte in Kommission bei Scripta Mercaturae Verlag, 1991), 19–64

Wittern, Christian, Arianna Ciula, and Conal Tuohy. 2009. "The making of TEI P5." *Literary and Linguistic Computing* 24(3): 281–96. doi:10.1093/llc/fqp017.

Notes

1 See Burnard 1988 for a contemporary report on the event.

2 See TEI 1988.

3 Many resources are available providing information about SGML (Standard Generalized Markup Language) and its successor language XML (Extensible Markup Language); Wikipedia is as good a starting point as any.

4 For example DeRose et al. 1990.

5 See Coombs, Renear, and DeRose 1987.

6 A good introduction to the TEI's use of this formalism is provided by Langendoen and Simons (1995).

7 The two chapters of P5 which define feature structure representation and feature system definition are copublished by ISO as ISO 24610-1:2006 and ISO 24610-2:2011 respectively.

8 Ide and Veronis 1995.

9 See TEI 2005.

10 See Wittern, Ciula, and Tuohy 2009 for a summary of the work undertaken.

11 These and other statistics on Sourceforge usage are available at https://sourceforge.net/projects/tei/stats/.

Cite this article

Electronic reference

Lou Burnard, « The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure », *Journal of the Text Encoding Initiative* [Online], Issue 5 | 2013, Online since 21 June 2013, connection on 25 July 2013. URL : http://jtei.revues.org/811 ; DOI : 10.4000/jtei.811

Author

Lou Burnard
Lou Burnard has been closely associated with the TEI since its inception, initially as European editor, and more recently as an elected member of the TEI Board. He also plays an active part on the TEI Council. Since retirement from Oxford University Computing Services, where he was responsible for the development of the Oxford Text Archive, the British National Corpus, and several other key projects in the Digital Humanities, he has been working as a private consultant, most recently with the French infrastructural organization TGE Adonis.

Copyright

Abstract

It is twenty-five years since the Text Encoding Initiative was first launched as a research project following an international conference funded by the US National Endowment for the Humanities. This article describes some key stages in its subsequent evolution from research project into research infrastructure. The TEI's changing nature, we suggest, is partly a consequence of its close and highly responsive relation with an active user community, which may also explain both its longevity and its effectiveness as a part of the digital humanities research infrastructure.

Index terms

Keywords : humanities computing, history, infrastructures, text encoding