3 How modeling standards evolve

The case of the TEI

Lou Burnard

1 Standards . . .

There is a very old joke about standards which says: “The nice thing about standards is that you have so many to choose from.” It is attributed by Wikiquotes to Andrew Tanenbaum (1981, p. 168) and has also recently been given a new lease of life by a popular xkcd cartoon. Like many old jokes, it plays on an internal contradiction (a structuralist might say “opposition”). On one hand, the world is a complicated place in which we value diversity and complexity; on the other, we value standards as a means of controlling that diversity. Standards may be considered to be instruments of control, managed or even imposed by a centralizing authority, or randomly spread out for our delight as if on a buffet. This contradiction is particularly noticeable when the process of standardization has been protracted, whether because the technologies concerned are only gradually establishing themselves, or because of disagreements amongst the decision-making parties, but is a tension inherent to the process. In the world of consumer electronics, for example, there is a financial market-driven imperative to establish standards as rapidly as possible so that new products may be developed more cheaply and efficiently, and at the same time an equally strong market-driven imperative not to standardize at all, so long as one’s own product has significant market share in comparison with those of the would-be standardizers. Above all, successful standardization requires the existence of a shared perception, or model of how a product should look or behave, before any consensus can emerge. For this reason, it seems useful to consider how the concept of information modeling has emerged, and has itself been the subject of standardization.

In the academic research community, similar tensions underlie the gradual evolution of individual ways of thought into communities of practice, and the gradual consensus-based emergence from these of de facto and (eventually) “real” standards. Scientific research communities are tribal both by temperament and in their practice for a variety of reasons, both good and bad. Tribes define themselves by shared perceptions and priorities, by shared models of reality, and by the specific tools or methods that support their activities. (The opposition often made between methodology and discipline is thus at best debatable—as witnessed by the fact that polemical articles entitled “What is digital humanities?” generally debate it). The adoption of a particular set of assumptions about what objects and methods are fruitful and pertinent can become deeply entwined with a research community's sense of its own identity, jealously guarded, aggressively promoted, and coercively imposed on the agnostic. At the same time, if such assumptions are to be adopted by the wider community, their proponents must seek to establish a consensus. If their model is to achieve recognition, it will not be by fiat from any central body or establishment, though such entities may well play a role in facilitating a context in which consensus and (perhaps) standardization can be achieved—for example, by specific research funding policies.

Standardization has a frivolous younger sibling called fashion, whose role in determining the ways in which particular modes of thought become institutionalized (or standardized) should not be neglected. Fashion reflects and (occasionally) affects broader socio-technological changes in ways that are hard to determine. Is the uptake of Twitter within the research community cause, effect, or symptom of shifts in the way we perceive the humanities' central role of explaining ourselves and our surroundings to ourselves? If we agree with, for example, Jones (2014) that the eversion of the digital world into the "real world" has been entirely transformative, does it make any sense to insist on a continuity in the models we apply, and the discourse derived from their application? And contrariwise, if we think that nothing fundamental has changed, and hence that the nature of the devices we use for communication is largely a matter of fashion, are we comfortable with the implication that there is a clear continuity between (say) clay tablet and mobile phone, such that the model we apply to describe messages on one will also be useful to describe the other? The higher one advances up the mountain, the easier it becomes to see the world as simply brown, blue, or green, but the harder it becomes to see the nuances in the shadows.

A good definition of modeling is that it is the process by which we construct meaning from observed data. The classic scientific procedure is to form a hypothesis and then search for observed data, either to support or to contradict it. Living now in an over-instrumented world of data-excess, we tend to do the reverse: that is, we look at the data and try to construct a hypothesis to support it, using the best tools at hand, or the tools that seem to give results consistent with our own internal model. The currently fashionable technique of topic-modeling is a case in point. Yet we do well to remember that the only reason we are now in a world awash with comparable data is precisely because standards for the representation of that data have now become reasonably pervasive and effective.

2 Data versus text

Our focus in this chapter is the evolution of standardized data models in the humanities and social sciences, and we therefore take a historical perspective. Nevertheless, much of what we discuss seems applicable more widely, both across other scientific disciplines, and even perhaps within a synchronic framework. One does not have to be a historian to suspect that the kinds of story we tell now about what our predecessors thought are likely to have been determined as a consequence of that body of tradition as much as they are by autonomous reflection.

2.1 Data modeling in the real world

The word "modeling" as used throughout this book is naturally inseparable from any kind of semiotic process, but in the domain of informatics began to be applied in a self-conscious and conscientious way in the 1960s and 1970s. This was the first period of massive expansion of digital technologies into the "real world" of business, public service, the research community, and, of course, the military.

This was the age of the mainframe computer, those massive power-hungry, water-cooled assemblies of transistors and storage systems based on huge magnetized ferric surfaces, on glass or metal disk, or spools of plastic tape. For our present purposes, the salient feature of those now superseded machines was not so much that they needed to be maintained in special air-conditioned environments or attended to by serious people in white coats—the same, after all, is true of the server farms deployed by Amazon or Google which have replaced them in today's world—but rather that they came in so many radically different forms. In many respects, of course, an IBM 370 and an ICL 1906, a CDC 6400, or a Univac 1100 machine all did much the same thing, relying on essentially the same set of mathematical and physical principles: a central processing unit, data storage, a set of predefined instructions for manipulating discrete pieces of data, input and output peripherals, and so on. But wherever there was scope for divergence—in the number of bits used to represent a single unit of storage, in the assembly code used to generate sequences of instructions, in the software libraries and operating systems built on top of all these things—they diverged. For this reason, as much as because of the significant amount of effort needed to keep these monolithic machines functioning at all, software developers and users alike rapidly began to focus on questions of interoperability of data and (to a lesser extent) software, and hence to participate in a variety of industry-led forums, user groups, standardization bodies, and so on. Typical also of the period was the tension between standardized programming languages such as COBOL or ALGOL, developed as a result of discussion amongst representatives of a number of interested but competitive parties, and imposed standards such as FORTRAN developed by a dominant manufacturer (in those days, IBM) or user group (in those days, the hard sciences). This applied even to such an arbitrary matter as the internal representation of character sets: IBM continued to support only EBCDIC, its own multi-flavored 8 bit code, for thirty years after the US government had mandated use of the industry-developed 7 bit ASCII code, the ancestor of today's Unicode. Again, this kind of tension does not seem entirely alien to contemporary experience.

A key driver in the impetus towards more and more standardization (and hence the focus on modeling techniques) across the data-processing departments of corporations and administrations worldwide was the rise of the corporate database.

As both commercial and government organizations surveyed their information-processing activities, the need to integrate previously discrete systems (many of them not yet digital) became more and more evident. Evangelists for data analysis, such as John Sowa (Sowa, 1984), argued that integrated database systems would offer an escape from existing preconceptions and from the design constraints inherent in pre-electronic systems. Existing manual methods were not designed to facilitate either the sharing of data or multiple ways of accessing subsets of it. When converting manual systems to electronic form, therefore, it was correspondingly important that these constraints should not be perpetuated in a new and more insidious form by requiring of the user, for example, a detailed knowledge of the minutiae of a particular computer's filing system before permitting access to the information it contained. Neither should the computerized system simply mimic the manual system it was designed to replace. The manual system had been a means to an end, not an end in itself. To achieve these objectives, deep ontological questions about the goal of an enterprise and the information it processed had to be confronted and resolved. Hence, we find database designers confidently asserting that their task was to abstract away from the mundane world of order forms, invoices, and customer address lists, in order to create a structure representing the information of which those documents were the physical trace, by which they meant the formal identification of real world entities and relationships among them. Sowa dignified this process with the name of conceptual analysis: "the work of philosophers, lawyers, lexicographers, systems analysts and database administrators" (Sowa, 1984, p. 294; see also http://ontolog.cim3.net/forum/ontolog-forum/2009-10/msg00165.html), but it would not have been an entirely strange concept for any medieval philosopher familiar with Plato.

By the early 1980s, several competing "standard methodologies" (note the plural) were being marketed for the process of defining reality in a business context—that is, those portions of reality that mattered to an enterprise, along with a wide range of complex (and expensive) software tools to simplify both that task, and the semi-automatic generation and implementation of actual data systems corresponding with the model so painstakingly arrived at. These systems naturally implemented a range of different data models. IBM, still a player at this time, had invested too much in its hierarchic system IMS not to see this as the only natural way of working; the business community, on the other hand, had worked hard in its CODASYL committee to develop what was called a network model; while in the rapidly expanding computer science research community, the relational model developed by ex-IBM staff Codd and Date was clearly the way of the future. Whether you regarded your data as hierarchically organized nodes, as a network of nodes, or as normalized relations, there was software to support you, and a community of practice to talk up the differences amongst these orthodoxies and their implications for data representation rather than their similarities.

A book called *Data and Reality* (Kent, 1978), first published in 1978, comes from that heroic age of database design and development, when such giants as Astrahan, Chen, Chamberlin, Codd, Date, Nijssen, Senko, Tschritzis, and others were slugging it out over the relative merits of the relational, network, and binary database models and the abstractions they supposedly modeled. Kent's quietly subversive message was that this was a struggle predominantly over terminology. He noted that almost all of these passionately advocated models were fundamentally very similar, differing only in their names, and in the specific compromises they chose when confronted by the messiness of reality. Whether you call them relations or objects or records, the globs of storage handled by every database system were still combinations of fields containing binary representations of perceptions of reality, chosen and combined, for their utility in a specific context. The claim that such systems modeled reality in any complete sense is easy to explode; it is remarkable, though, that we still need to be reminded, again and again, that such systems model only what it is (or has been) useful for their creators to believe. Kent is sanguine about this epistemological lacuna: "I can buy food from the grocer, and ask a policeman to chase a burglar, without sharing these people's view of truth and beauty" (Kent, 1978, p. 202), but for us, living in an age of massively interconnected knowledge repositories, which has developed almost accidentally from the world of more or less well-regulated corporate database systems, close attention to their differing underlying assumptions should be a major concern. This applies to the differently constructed communities of practice and knowledge which we call "academic disciplines," just as much as it does to the mechanical information systems those communities use in support of their activities.

In its time, Kent's book was also remarkable for introducing the idea that data representations and the processes carried out with them might be represented in a unified way. At a period when the processes carried out by computer programs were thought of as belonging to an entirely different conceptual domain from the data on which they operated, the notion that it might be convenient to consider as a single entity both a piece of data and the processes that might be associated with it was distinctly innovative. Kent's work is thus an important precursor of what we now call object-oriented processing, which is characterized by this unified approach. An object-oriented programmer defines objects that combine data structures with the methods appropriate to them, rather than defining data structures and data processes independently, as the dominant programming styles of the 1970s required. Kent's work also reminds us of some fundamental ambiguities and assumptions often swept under the carpet during conceptual analysis of any period. Are objects really uniquely identifiable? "What does 'catching the same plane every Friday' really mean? It may or may not be the same physical airplane. But if a mechanic is scheduled to service the same plane every Friday, it had better be the same physical airplane" (Kent, 1978, p. 7). The way an object is used is not just part of its definition. It may also determine its existence as a distinct object.

Kent's understanding of the way language works is clearly based on the Sapir-Whorf hypothesis of linguistic relativity: indeed, he quotes Whorf approvingly: "Language has an enormous influence on our perception of reality. Not only does it affect how and what we think about, but also how we perceive things in the first place" (Kent, 1978, p. 200). There is an odd overlap between his reminders about the mocking dance that words and their meanings perform together and contemporaneous debates within the emerging field now known as GOFAI, or "Good Old Fashioned Artificial Intelligence."[1] And we can also see echoes of similar concerns within what was in the 1970s regarded as a new and different scientific discipline called Information Retrieval, concerned with the extraction of facts from documents. Although Kent explicitly rules text out of discussion ("We are not attempting to understand natural language, analyze documents, or retrieve information from documents," Kent, 1978, p. vi) his argument throughout the book reminds us that data is really a special kind of text, subject to all the hermeneutical issues we tend mistakenly to consider relevant only in the textual domain.

This is particularly true at the meta-level, of how we talk about our data models, and the systems we use to manipulate them. Because they were designed for the specific rather the general, and because they were largely developed in commercially competitive contexts, the database systems of the 1970s and 1980s proliferated terms and distinctions amongst many different kinds of entity, to an extent that Kent (like Occam before him) argues goes well beyond necessity. This applies to such comparatively arcane distinctions as those between entity, attribute, and relationship, or between type and domain, all of which terms have subtly different connotations in different contexts, though all are reducible to a more precise set of simple primitives. It applies also to the distinction between data and metadata. Many of the database systems of the 1980s and 1990s insisted that you should abstract away all the metadata for your systems into a special kind of database variously called a data dictionary, catalogue, or schema, using entirely different tools and techniques from those used to manipulate the data itself. This is a needless obfuscation once you realize that you cannot do much with your data without also processing its metadata. In more recent times, one of the more striking improvements that XML (Extensible Markup Language: the W3C-defined de facto standard for representing information on the web) made to SGML (Standard Generalized Markup Language: the ISO standard for markup languages from which XML was derived) was the ability to express both a schema and the objects it describes using the same language. The representations of real world objects manipulated by an information system are themselves objects in the real world, and should therefore be modeled in the same way. How best to document the intended meaning of those representations—what is usually called the semantics of an XML schema—remains a matter that only a few current XML systems (notably the TEI) explicitly consider.

2.2 Data modeling in the humanities

According to the foundational myth of the digital humanities, it all began in 1950 or thereabouts when a Jesuit father called Roberto Busa conceived the idea of using a machine to tabulate every occurrence of every word, and the lemmas associated with the words, and the senses of those lemmas, in the works of St Thomas Aquinas. His vision was realized (some years later), with the aid of Thomas Watson of IBM, and you can see it still working today at: www.corpusthomisticum.org/it/index.age.

Of course, as Busa himself points out in a characteristically self-deprecating article published in 1980, he was far from having been the first person to have considered using mechanical or statistical methods in the investigation of an author's writing: for example, in the nineteenth century, the British statistician August De Morgan, and in particular a student of his, an American scientist called T.C. Mendenhall had speculated that the frequency of occurrence of certain words might be used to distinguish the writing of one person from that of another (Mendenhall, 1887). Clearly, human beings do write differently from one another, and certainly human readers claim to be able to distinguish one writing style from another. Since all they have to go on when processing writing is the words on the page, it seems not entirely implausible that the calculation of an author's "characteristic curve of composition" (as Mendenhall called it) might serve in cases of disputed authorship.

With the advent of automatic computing systems, and in particular of more sophisticated statistical models of how words are distributed across a text, it became possible to test this hypothesis on a larger scale than Mendenhall had done (he relied on the services of a large number of female assistants to do the counting drudgery), and a number of research papers began to appear on such vexed topics as the authorship of the Pauline epistles, the disputed works of the Russian novelist Sholokhov, or the Federalist Papers (a set of anonymously published pamphlets of the American Revolutionary War period). At the same time, many research groups began to contemplate a more ambitious project that might develop a new form of stylistic studies, based on empirical evidence rather than impressionistic belief or dogma. Stylometry, as this was called, and authorship studies dominated this first heroic period of the digital humanities, and continue to fascinate many researchers.[2]

At the same time, but in another part of the forest, a new tribe of linguists was emerging, re-energizing an empirical tradition going back to J.R. Firth[3] with the aid of massive quantities of machine-readable text. The emergence of the Brown Corpus in 1960 and its successors[4] represents an important moment in the evolution of the digital humanities for several reasons. The "corpus linguists," as they called themselves, were probably the first humanities researchers of whom it might plausibly be said that their research was simply not feasible without the use of digital technologies. The model of language praxis and linguistic patterning that emerged from their research was also fundamentally innovative, not to say controversial with regard to the prevailing Chomskyan orthodoxy of the time. The insights gained from their approach have radically changed the way in which such traditional activities as dictionary-making or language-teaching and learning are now carried out. And, with hindsight, we can detect in their methods a distinctive approach to the modeling and analysis of textual data.

As with the stylisticians and the authorship hackers, however, the corpus linguists' shared model of text was neither formally defined nor structurally ambitious. Its focus was something called the word, variously defined as an orthographic unit, or a lexical one, even though the process of lemmatization—the grouping of individual tokens under a single lexical form—remained problematic; as the title of an article by Brunet memorably reminds us: *Qui lemmatise dilemmes attise* . . . (Brunet, 2000). Corpus linguists studied ngrams—recurrent sequences of words or tokens—but were less interested in indications of macro-textual organization or structure, except where these could be derived from an analysis of the constituent words. Individual tokens in a text were often annotated by codes indicative of their word-class (noun, preposition, and so on) but the annotation of multi-word sequences, for example, to indicate syntactic function, was more problematic and hence less standardized.

Nevertheless, the development of corpus linguistics as a defined area of research (a discipline even) owes much to the clear consensus among its practitioners concerning both core principles, methods, and objects that define the discipline, and those concerning which multiple points of view were recognized. For example, the Brown corpus instantiated a surprisingly long-lived model for the construction of language corpora that was based on fixed-size synchronic sampling of language production according to explicit selection criteria. In developing the Cobuild corpus (Sinclair, 1987) by contrast, Sinclair was one of the first to propose a model of continuous sampling from an ever-expanding and diachronic base of reference materials, and may be thought of as having initiated the perspective memorably phrased by more than one American linguist as “there’s no data like more data,”[5] anticipating today’s gigaword corpora, and the “web as corpus” concept. The theoretical model underlying both these projects and the many others that followed them was, however, just the same: the function of linguistic research was to identify regularities in the way language is used, and to construct a view of how language functions solely in terms of that empirically derived data, rather than from a priori theorizing about postulated linguistic systems.

If stylometrics and corpus linguistics alike thrived in the digital environment, it was perhaps because their objects of study, the raw material of text, seemed easy to model, because a consensus as to its significant particularities had long been established. The same could hardly be said of other areas of the humanities, in which the primary object of interest was not the text but the subject matter of the text, not its form but its intention, not the medium but the message. And yet it was obvious (as Manfred Thaller, Jean-Philippe Genet and others argued persuasively in the 1980s) that there was much to gain if only consensus could be achieved as to the best way of transferring the written records that constitute the primary sources for historical research into a digital form. Running through the proceedings of, for example, the annual conference of the Association for History and Computing, is a constant argument between text analysis and text representation. For those whose methods were entirely contingent on the use of particular pieces of software (statistical packages, logic programming systems, relational database systems, etc.) the source existed only to be pillaged, annotated, or reduced to some more computationally tractable form. For those with a broader perspective, wishing to produce resources that might be both adequate to the immediate needs of one research project and generic enough to facilitate its reuse and integration with other resources, the absence (or multiplicity) of standard models for their representation seemed insurmountable. In the nineteenth century, historical scholars had frequently labored (and gained recognition for their labor) to codify, transcribe, and standardize collections of medieval and early modern records from many sources in print form. How should that effort be replicated and continued into the digital age?

We can see also in those conference proceedings,[6] and in the journals of the period, a tendency for researchers in history to adopt whatever computational solutions the market was throwing up, without much effort to truly appropriate it to their perspective. Social historians in particular often embraced uncritically the methods of sociology, which required the reduction of historical data to vectors of scores in a predefined matrix, easily analyzable by tools such as SPSS or SIR, popular statistical packages that had been developed originally to aid in the analysis of survey data, rather than archival records. Many others accepted uncritically the database orthodoxy proposed by their local computing center (in those distant days, many universities provided computing services and support for them centrally) which, in practice, meant adjusting their data to the hierarchic, network, or relational model, as the case might be. Others, perhaps more surprisingly, attempted to apply the methods of logic programming, reducing historical data to sets of assertions in predicate logic: the pioneering work of the French archaeologist Jean-Claude Gardin (Gardin, 1980) was often cited in support of this idea. In the UK, there was even a short-lived vogue for recommending logic programming in secondary school teaching (see, e.g., Nichol et al., 1987). For the most part, however, few historians thought to follow their literary or linguistic colleagues in preferring to develop their own tools of analysis which might reflect models closer to their discipline's view of its data.

With a few notable exceptions, it seems that most historical researchers were content simply to adopt technical standards established by the wider data-processing community (relational databases, information retrieval systems, and so on) despite the reductionist view of the complexities of historical sources that such systems required. Amongst the exceptions we should, however, note pioneering experiments such as those of Macfarlane, 1977, or King, 1981, as well as more mature and influential systems such as Thaller's κλειο (Thaller, 1987), which demonstrated that it was possible to use the new technology to combine faithfulness to the source with faithfulness to the historian's understanding, in a kind of re-evaluation of the German tradition of Quellenkritik or "source criticism" pioneered by historians such as Leopold Ranke and Berthold Niebuhr.[7] That re-evaluation, by focusing on ways of modeling in an integrated way, both the text itself and the historian's reading of it, showed the way forward for subsequent digitally assisted humanities research in many disciplines, just as Quellenkritik originally benefited from the insights of traditional philology.

3 The apotheosis of textual modeling

What happens when a non-digital text is transformed to a digital form? If the goal is no more than to re-present that source, it is very likely that the job will be considered accomplished by a reasonable quality digital image, perhaps accompanied by a transcription of (most of) the words on the page, in a form that will facilitate a reasonably close simulation of the original to be displayed when the digital version is presented on screen or paper. Self-evidently, this approach prioritizes the visual aspects of the source at the expense of its semantics, except in so far as those are intrinsically tied to its visual aspects. It requires but does not impose the addition of metadata to contextualize and describe a source, which may or may not be stored along with the digital surrogate itself.

Nevertheless, presumably for largely practical and economic reasons, page-imaging, or facsimile production remains the common denominator of the majority of current digitization initiatives, as it has done for the past few decades. For today's digital library, in fact, we may say that the predominant model is one in which digital surrogates approximate as closely as possible a subset of the visual characteristics of a source. Note that this remains a subset: Prescott, 2008 among others has pointed out how even the most fastidiously prepared and executed digital imaging of an ancient manuscript can fail to capture all of its properties of interest. Digitization is an inherently reductive process and nothing is likely to change that. As in database design, therefore, it is essential to define precisely the limitations of the model to which one is reducing the source.

In explicitly rejecting that model of textual essence, the Text Encoding Initiative (TEI) attempted something rather more ambitious. From the start, its intention was to create an explicit model of the objects and structures that intelligent readers claim to perceive when reading a text; the explicit claim was that by modeling those readings, and assigning a secondary role to the rendition of actual source documents, the goals of integration and preservation of digital surrogates would be greatly simplified; perhaps implicitly there was also an attempt to redirect the energies of scholarly discourse away from the accidental trivia of word processing in favor of a more profound consideration of the meaning and purpose of written texts. This opposition is most clearly stated in Coombs, Renear and DeRose's foundational text on the future of scholarly communication (Coombs et al., 1987) and it is also explicit in the original design goals of the TEI as enumerated in the so-called Poughkeepsie Principles: "Descriptive markup will be preferred to procedural markup. The tags should typically describe structural or other fundamental textual features, independently of their representation on the page" (TEI, 1988).

A reading of the TEI's original design documents[8] shows clearly the influence of contemporary database design orthodoxies. For example, a working paper from 1989 called "Notes on Features and Tags"[9] defines a conceptual model in which entities such as tags are considered independently from both the abstract features they denote and the textual data strings to which they are attached, before proceeding to define a data structure to hold all the features of a given mark-up tag. This latter definition is labeled as "Design for a TAGS Database," and a mapping to a simple RDBMS provided for it. The assumption behind the model described here is that the well-attested variation in the many ways texts were converted for use by computer might be overcome by treating those variations as accidental quirks of the software in use. Essentially, this model says, there is a determinable collection of textual features of interest on which scholars agree, many of which are differently expressed by different pieces of software, but which could all be potentially be mapped to a single interchange format. The TEI was conceived of originally as an interchange or pivotal format; not necessarily as something to replace existing systems of markup, but as something to enable them to communicate, by appealing to a higher level abstract model of the common set of textual features that individual markup systems were deemed to denote.

This same working paper includes a suggested SGML DTD which might be used to organize the components of that higher level abstract model, and which is in many ways the ancestor of the schema currently used to define TEI components. The fundamental concepts of this model, for which the TEI editors coined the name ODD (One Document Does it all), have clear antecedents both in the work of Donald Knuth and in contemporary SGML documentation systems, such as that developed for a major European publishing initiative called Majour, and have not fundamentally changed since. The model is well documented elsewhere;[10] we highlight here a few of its salient characteristics, in particular those that qualify it for consideration as a meta-model, a tool for the construction of models.

There has long been a perception that the TEI is a prescriptive model, as indeed in some respects it is: it prescribes a number of very specific constraints for documents claiming to be TEI conformant, for example. However, the prescriptive part of the TEI is concerned only with how the TEI definitions are to be deployed; very few prescriptions are provided as to which of the many hundreds of TEI-defined concepts should be selected in a given context, although, of course, each choice of component has implications for subsequent choices. In this respect, the TEI system resembles a somewhat disorganized collection of independent components rather than a single construct.

Each of these components is, however, defined in a standardized way, using essentially the same set of properties: a name or canonical identifier; a description of its intended meaning (supplied in one or several natural languages); where possible, an indication of equivalent objects in other systems; its classification within the TEI's conceptual model;[11] a formal model of its possible constituent components and attributes, usage notes and illustrative examples. None of this documentation is inextricably linked to any particular enabling technology: although the first version of the TEI was expressed using SGML, later versions have used XML, and several experiments have shown the feasibly of remapping its definitions to other currently fashionable technologies such as JSON or OWL. This also is in line with (though not identical to) the original goals of the project.

As noted above, those original project goals make clear that the TEI was not originally conceived of as a standards-making exercise, but rather as a way of defining a convenient interchange format, which might perhaps be generalized to serve as an encoding format in its own right.[12] To define its interchange format, however, the TEI necessarily had to define an interlingua in which existing models of textual structure might be re-expressed without loss of information, and thus found itself inevitably working towards the definition of a meta-standard: a framework for the definition of standards. The disorganized constellation of textual features or objects found in the TEI Guidelines corresponds with the set of "significant particularities" originally identified by the members of the TEI working groups, which has been refined and revised over a period of several decades, during which new objects have been added and existing ones revised for consistency and clarity. As noted elsewhere (Burnard, 2013), the TEI system as a whole is thus not a fixed entity, but one that has evolved and developed in response to the changing needs and priorities of its user community. In this respect, it has been created in a very different way from most standards.

This shape-shifting is a continuation and intensification of a principle adopted very early on and manifest in the conspicuously consultative manner by which the TEI Guidelines were originally constructed. They do not represent the views of a small technical self-appointed elite, but rather the distillation of a consensus formulated by combining input from specialists from many academic disciplines, having in common only an interest in the application of digital technologies within those disciplines. As an internationally funded research project, the TEI project also conscientiously strove to pay equal attention to the needs of researchers separated by language and geography. Although the TEI pre-dates the World Wide Web, it was born into a world in which virtual internet-based communities were already emerging and remains, perhaps, one of the first and most successful user-focused internet-mediated projects to have been created, even without benefit of today's "social media."

The interdisciplinary nature of the TEI model is also reflected in the way the Guidelines themselves are organized and in the way that its formal definitions are intended to be used. Inevitably, most of the individual chapters of the reference manual known as TEI P3 (TEI, 1994), which constituted the first public release of the TEI Guidelines in 1994, contained much material unlikely to be of interest to every user. At the same time, every chapter contains material of importance to some user. The material combined rigorous prose definition and exemplification with formal specifications, initially expressed as a "tagset": a set of declarations expressed in the DTD language used by the SGML standard. The expectation was that the skilled user would (having read and understood the documentation) select one of a small set of "base" tagsets (prose, verse, drama, dictionaries, speech, and so on), together with a set of elements common to all kinds of text (the "core") and the metadata associated with them (the "header"). This combination could then be enriched further by the addition of any number of "additional" tagsets providing more specialized components, each reflecting a particular style of analysis (linguistic, hypertextual, text-critical, and so on). Finally, a user might elect to suppress some of the components provided, modify some of their properties, or even to add new components not provided by the TEI model at all.

This model, humorously referred to as the "pizza model" by analogy with the way that Chicago's favorite dish is typically constructed, also seems in retrospect to reflect something of the deeply balkanized intellectual and social milieu of its time. For all its good intentions and practicality, the tidiness of the pizza model seems at odds with the gradual blurring of the well-fenced frontiers between linguistics and literature, history and sociology, science and the humanities, which characterizes our current intellectual landscape, in which humanities research ranges far and wide across old disciplinary frontiers, grabbing methods from evolutionary biology to explore textual traditions, or deploying complex mathematical models to trace the evolution of literary style.

As first instantiated, the construction of a personalized model from the huge (and occasionally overlapping) range of possibilities defined by the TEI Guidelines was a relatively complicated task, requiring fairly detailed technical knowledge about SGML, as well as a good grasp of the way in which the TEI tagsets were organized. Unsurprisingly, many early adopters preferred to use a generic predefined model such as TEI Lite (TEI, 1990) or to rely on one provided by their own research community, such as the Corpus Encoding Standard (Ide and Priest-Dorman, 2000), or, more recently, the Epidoc Guidelines (Elliott et al., 2007–14). Yet the existence of many such customizations, even those that were not always entirely TEI conformant as the term was understood at the time, clearly vindicated the basic design of the project, which was to construct not a single standard model for the encoding of all texts for all time, but rather an architecture within which such models could be developed in an interoperable or at least interchangeable way, a kind of agreed lexicon from which individual dialects could be derived. The same mechanism (the ODD system mentioned above) is used to define both the TEI itself and customizations of appropriate to a given project; it is thus easy to determine the correspondence between a project-specific model and the whole of the TEI from which it was derived by specifying the TEI tagsets used, selectively choosing parts of each tagset and (where judged necessary) adding new declarations to complement or replace those provided by the TEI.

The transition from TEI P3 to TEI P4 carried out in 1999 was a largely automatic process of re-expressing the same objects in XML rather than SGML, with little of significance being changed. However, the development of TEI P5[13] was a more ambitious process. Necessarily, it involved the addition of much new material and the updating of some no longer relevant recommendations such as those concerning character encoding, but it also included changes introduced specifically to simplify and render more accessible the hitherto rather arcane customization mechanism. Firstly, the overall architecture was simplified by abolishing the distinction amongst types of tagset: in TEI P5, each P3 tagset becomes a simple collection of specifications known as a module, and any combination of modules is feasible. It is even possible (within limits) to select elements for inclusion in a model without specifying the module in which they are defined. Secondly, the class mechanism used to group elements together by their semantics, their structural role, or their shared attributes (independently of their module) was made both more pervasive and more apparent; indeed, any customization of TEI P5 beyond simply creating a subset now requires some understanding of the class system. Thirdly, simple subsetting was made very much easier, and a simple web interface called Roma was provided to achieve it.

This short review of the TEI's technical evolution suggests that the project, which was initially intended to define a basic interchange format into which any other kind of textual markup might be transformed, has instead become a framework for the definition of such markup systems. What began as a simple exercise in string processing has of necessity developed into a higher-level system, using more sophisticated and more general-purpose methods and processors. Today's TEI user is less interested in defining their own markup syntax than in finding a standard way of expressing their own textual model. We suggest that by abstracting away from the specifics of any particular markup syntax to focus on the conceptual model underlying it, the TEI designers paved the way for this change.

4 Explicitness and coercion

Perhaps there is a long-running tension within all standardization efforts between generality and customization. The more generally applicable a standard, the harder it may be to use productively in a given context; the more tailored it is to a given context, the less useful it is likely to be elsewhere. Yet surely one of the main drivers behind the urge to go digital has always been the ability not just to have one's cake and eat it, but also to produce many different kinds of cake from the same messy dough. For this to work, there is a need for standards that do not limit choice, but rather facilitate an accurate presentation of the choices made. Such an approach is also essential for a modeling standard that hopes to be effective in a domain where the objects of discourse, the components of the model, are constantly shifting and being remade, and consequently remain controversial.

Consider, for example, the common requirement to annotate a stretch of text believed to indicate a temporal expression with some normalized representation of it. This has obvious utility if we believe the expression represents the date of some event, and we wish to perform automatic analyses comparing many such—for example, to determine the chronological sequence of a collection of documents. One document says simply "Wednesday," another says "Saint Martin's day," yet another says "the 12th Sunday after Lammas Tide." Some kind of normalization is clearly essential if these are to be compared, but the norms for temporal reference vary considerably both across cultures (dates in the Islamic, Aztec, Roman, Chinese, or Jewish calendars are not all easily interconvertible), across time (the Gregorian versus the Julian calendar, for example) and even across international standards (a W3C date is not the same thing as an ISO date). Simplifying somewhat, a TEI document may choose to normalize dates using the international standard for representation of temporal expressions (ISO 3601), or the profile (subset) of that standard recommended by the W3C, or it may choose to use some other user-defined calendar system entirely. The price of this liberty is that all three options must somehow be provided for within the TEI architecture, even though in any given case it is likely that only one normalization method will be used. Leaving aside the technical detail, the TEI class system provides exactly such a mechanism: although attributes appropriate to each normalization method are defined, in any given customization only a subset will be made available. Hence, while the developer of a generic TEI processor needs to be aware that all three options are feasible, in a given case they can reliably infer which has actually been deployed by processing the ODD specification associated with the documents in question.

Ever since its first publication, the TEI has been criticized for providing too much choice, giving rise to too many different ways of doing more or less the same thing. At the same time (and even occasionally by the same people), it has been criticized for limiting the encoder's freedom to represent all the concepts of their model in just the way they please. Neither criticism is without foundation, of course: despite the best efforts of the original TEI editors, Occam's razor has not been applied as vigorously throughout the Guidelines as it might have been, and as a result life is complicated for both the would-be software developer and the conscientious digital author. Darrell Raymond remarked in a very early critique of SGML that "descriptive markup rescues authors from the frying pan of typography only to hurl them headlong into the hellfire of ontology" (Raymond et al., 1996). Standardized modeling tools such as ODD cannot entirely remove those ontological anxieties, but at least they facilitate ways of coming to terms with them, by providing a neutral space in which the system designer can make explicit their views, in particular a vehicle for them to express the degree and nature of any dissent between their model and that elaborated by the TEI. The TEI provides no tags for the description of unicorns, nor even (as yet) for botanical names, but it does provide a standardized way of defining such tags, and relating their definitions to concepts already existing in the TEI model.

At the same time, the very success of particular TEI customizations increases the risk that the TEI may eventually begin to compromise on its design principles—for example, by downgrading support for the generic solution in favor of the one that interfaces most neatly with the latest most fashionable tool set. This risk of fragmentation needs to be confronted: do we want to see a world in which various different "TEI-inspired" models for editors of manuscripts, cataloguers, linguists, lexicographers, epigraphers, or users of digital libraries of early print separate themselves from the generic TEI framework and begin to drift apart, reinstating the babel of encoding formats that inspired the creation of the TEI in the first place?

A balance must be maintained between "do it like this" and "describe it like this" schools of standardization; while the former matters most to those charged with delivering real results in the short term, the latter is our only hope of preserving the inner logic of our models in the long term. For that reason, the importance of the TEI is not only that it has formalized and rendered explicit so many parts of the digital landscape, but also that it has done so in a consistent and expandable way. Its value as a meta-model is essential to its usefulness as a modeling tool.

All spheres of standardization activity, we suggested initially, demonstrate a tension between a centralized dirigiste urge and a decentralized desire for consensus. Attempts to provide standardized conceptual models are no exception to this generalization, but the most effective and long-lived such standards seem to require a powerful meta-modeling component. This enables the modeling standard to evolve in response to changing perceptions, priorities, and technologies without losing its identity. Standards may fail for a variety of reasons, but the most common is that no genuine consensus can be established among practitioners or theoreticians of the domain concerned; a standard that facilitates diversity of theory by reserving its constraints to the meta-model level is less likely to fall foul of this problem. A standardized meta-model enables diverse models to co-exist fruitfully, by providing a channel for mutual interchange and mutual comprehension.

Notes

[1] The acronym first appears in Haugeland, 1985.

[2] Holmes, 1994, provides a good bibliography of earlier work; Juola, 2006, reviews more recent thinking on the topic.

[3] For a persuasive historical analysis of this tradition and its development, see Léon, 2008.

[4] For links to documentation of this influential corpus and its imitations, including an impressive bibliography of research derived from it, see http://clu.uni.no/icame/manuals/

[5] The phrase is often credited to Robert Mercer: see www.lrec-conf.org/lrec2004/doc/jelinek.pdf (Jelinek, 2004).

[6] See, for example, Denley and Hopkin, 1987, or Denley et al., 1989.

[7] See Greenstein, 1991, for a collection of essays on the problems of modeling historical textual data sources.

[8] Many of the TEI's original working documents are preserved in its online archive; some of them have also been published, notably in Ide and Véronis, 1995.

[9] A lightly revised version is available from: www.tei-c.org/Vault/ED/edw05.htm

[10] The current system is fully described in Chapter 22 of the TEI Guidelines; for an early article outlining its architecture, see Burnard and Rahtz, 2004; for recent technical developments, see Burnard and Rahtz, 2013.

[11] The TEI architecture combines a notion of hierarchically organized element classes, similar to that found in many formal systems, with a loosely defined semantic model: the interested reader is referred to Chapter 2 of the Guidelines for further information.

[12] The first of the Poughkeepsie Principles mentioned above is "The guidelines are intended to provide a standard format for data interchange in humanities research"; the second is "The guidelines are also intended to suggest principles for the encoding of texts in the same format."

[13] Technical details of the transition from P3 to P5 are provided in Burnard, 2006, *inter alia*.

References

Brunet, E., 2000. *Qui lemmatise dilemmes attise*. *Lexicometrica*. Available at: http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.html (accessed January 24, 2016).

Burnard, L., 2006. New Tricks from an Old Dog: An Overview of TEI P5. In: Burnard, L., Dobreva, M., Fuhr, N., and Lüdeling, A. (Eds.). *Digital Historical Corpora – Architecture, Annotation, and Retrieval*. Dagstuhl, Deutschland 3–12 December 2006. Dagstuhl: IBFI.

Burnard, L., 2013. The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure. *Journal of the Text Encoding Initiative Issue*, 5. Available at: http://jtei.revues.org/811

Burnard, L. and Rahtz, S., 2004. *RelaxNG with Son of ODD*. Available at: Proceedings of Extreme Markup Languages: http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html.

Burnard, L. and Rahtz, S. 2013. Reviewing the TEI ODD System. In: *Proceedings of the 2013 ACM Symposium on Document Engineering*. New York: ACM, pp. 193–196.

Busa, R., SJ, 1980. The Annals of Humanities Computing: The Index Thomisticus. *Computers and the Humanities*, xiv, pp. 83–90.

Coombs J.H., Renear, A.H., and DeRose, S.J., 1987. Markup Systems and the Future of Scholarly Text Processing. *Communications of the ACM*, 30(11), pp. 933–47.

Denley, P. and Hopkin, D. (Eds.) 1987. *History and Computing*. Manchester: Manchester University Press.

Denley, P., Fogelvik, S., and Harvey, C. (Eds.) 1989. *History and Computing II*. Manchester: Manchester University Press.

Elliott, T., Bodard, G., Mylonas, E. and Stoyanova, S., 2007–14. EpiDoc Guidelines: Ancient Documents in TEI XML (Version 8). Available at: www.stoa.org/epidoc/gl/latest/.

Gardin, J.-C., 1980. *Archaeological Constructs*. Cambridge: Cambridge University Press.

Greenstein, D., 1991. Modelling Historical Data: Towards a Standard for Encoding and Exchanging Machine-Readable Texts. In: Thaller, M. (Ed.) 1991: *Halbgraue Reihe zur Historischen Fachinformatik,* A(11). St Katherinen: Scripta Mercaturae Verlag.

Haugeland, J., 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Holmes, D. I., 1994. Authorship Attribution. *Computers and the Humanities*, 28(2). Pp. 87–106.

Ide, N. and Véronis, J., 1995. *The Text Encoding Initiative: Background and Context*. Dordrecht/Boston, MA: Kluwer Academic Publisher.

Ide, N. and Priest-Dorman, G., 2000. Corpus Encoding Standard. Available at: www.cs.vassar.edu/CES/CES1.html.

Jelinek, F., 2004. Some of My Best Friends are Linguists. Paper presented at LREC 2004, Johns Hopkins University.

Jones, S.E., 2014. *The Emergence of the Digital Humanities*. London: Routledge.

Juola, P., 2006. Authorship Attribution. In: *Foundations and Trends in Information Retrieval*, 1(3), pp. 233–4.

Kent, W., 1978. *Data and Reality: Basic Assumptions in Data Processing Reconsidered*. Amsterdam: North-Holland Publishing.

King, T.J., 1981. The Use of Computers for Storing Records in Historical Research. *Historical Methods*, 14, pp. 59–64.

Léon, J., 2008. *Aux sources de la "Corpus Linguistics": Firth et la London School*. Langages, 3(171). Available at: https://www.cairn.info/revue-langages-2008-3-page-12.htm.

Macfarlane, A., 1977. *Reconstructing Historical Communities*. Cambridge: Cambridge University Press.

Mendenhall, T. C., 1887. The Characteristic Curves of Composition. In: *Science* – supplement, IX(214). Available at: https://archive.org/details/jstor-1764604.

Nichol, J., Dean, J., and Briggs, J., 1987. Logic Programming and Historical Research. In: Denley, P. and Hopkin, D. (Eds.) *History and Computing*. Manchester: Manchester University Press, pp. 198–205.

Prescott, A., 2008. The imaging of historical documents. In: Greengrass, M. and Hughes, L. (Eds.) *The Virtual Representation of the Past*. Aldershot: Ashgate, pp. 7–22.

Raymond, D., Tompa, F., and Wood, D., 1996. From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML. *Computer Standards & Interfaces*, 18, pp. 25–36.

Sinclair, J.M., 1987. Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary. *Computers and Translation*, 3(3/4), pp. 263–6.

Sowa, J.F., 1984. *Conceptual Structures*. Reading, MA: Addison-Wesley.

Tanenbaum, A.S., 1981. *Computer Networks.* Englewood Cliffs, New Jersey: Prentice-Hall.

Text Encoding Initiative (TEI), 1988, revised 1990. *Design Principles for Text Encoding Guidelines* Working Paper ED P1. Available at: www.tei-c.org/Vault/ED/edp01.htm.

Text Encoding Initiative (TEI), 1990. *TEI Lite: Encoding for Interchange: An Introduction to the TEI*. Available at: www.tei-c.org/Guidelines/Customization/Lite/.

Text Encoding Initiative (TEI), 1994. *Guidelines for the Encoding and Interchange of Machine-Readable Texts: Draft P3*. Chicago, Oxford: Text Encoding Initiative.

Thaller, M., 1987. *κλειο: A Data Base System for Historical Research*. Göttingen: Max-Planck-Institut für Geschichte.