How many standards do we need to model reality?

Lou Burnard *2016-03-17*

1 Standards...

There is a very old joke about standards which says 'Standards are a good thing because there are so many to choose from', recently given a new lease of life for example by a popular xkcd cartoon. Like many old jokes, this plays on an internal contradiction (the structuralist might say 'opposition'). On the 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 seem thus necessarily to be instruments of control, managed or even imposed by a centralising authority. This contradiction is particularly noticeable when the process of standardisation has been protracted, whether because the technologies concerned are only gradually establishing themselves, or because of disagreements amongst the decision-making parties. We see this contradiction particularly clearly in consumer electronics: there is a financial market-driven imperative to establish standards as rapidly as possible so that all may benefit, 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.

In the academic research community, similar tensions underly 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, and hence by the specific tools and methods which support their activities. (The opposition often made between *methodology* and *discipline* is thus at best debatable — as witness 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.

Standardisation has a frivolous younger sibling called *fashion*, whose role in determining the ways in which particular modes of thought become institutionalised (or standardised) 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 *modelling* 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 which seem to give results consistent with our own internal model. The currently fashionable technique of topic-modelling 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 article 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 modelling in the real world

The word *modelling* as used throughout this book is naturally inseparable from any kind of semiotic process, but began to be applied in a self-conscious and conscientious way in the 1960s and 1970s during 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 magnetised ferric surfaces, on glass or metal disk, or spools of plastic tape. For our present purposes, the salient feature of those now superceded 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 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. 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, 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, standardisation bodies, etc. Typical also of the period was the tension between standards such as COBOL or ALGOL, developed as a result of discussion amongst representatives of a number of interested but competitive parties, and standards such as FORTRAN imposed 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-flavoured 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 standardisation (and hence the focus on modelling techniques) across the data processing departments of corporations and administrations world-wide was the rise of the corporate database. As both commercial and government organisations surveyed their information processing activities, the need to integrate previously discrete systems (many of them not yet digital) became more and more evident. It was 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 mechanics of a particular computer's filing system before permitting access to the information it contained. Neither should the computerised 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 amongst them. This process was commonly dignified with the name of *conceptual analysis*: 'the work of philosophers, lawyers, lexicographers, systems analysts and database administrators.'(Sowa 1984) 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 which 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 organised nodes, as a network of nodes, or as normalised 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, and Tschritzis were slugging it out over the relative merits of the relational, network, and binary database models and the abstractions they supposedly modelled. Kent's quietly subversive message is that this is a struggle predominantly over terminology. He notes 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 modelled reality in any complete sense is easy to explode; it's 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', 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 should be represented in a unified way, the basic idea of what we now call object-oriented processing; yet it also reminds us of some fundamental

ambiguities and assumptions swept under the carpet even within that paradigm. 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.' 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: 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.' 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'. *[Note: The acronym first appears in Haugeland 1985]* 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, analyse documents, or retrieve information from documents') his argument throughout the book reminds us that data is really a special kind of text, subject to all the hermeneutical issues we tend (wrongly) 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 which 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 eighties and nineties 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 realise 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 made to SGML was the ability to express both a schema and the objects it describes using the same language. Where what are usually called the semantics of an XML schema should be described and how remains a matter which only a few current XML systems (notably the TEI) explicitly consider.

2.2 Data modelling 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 realised (some years later), with the aid of Thomas Watson of IBM, and you can see it still working today at `http://www.corpusthomisticum.org/it/index.age`

Of course, as Busa himself points out in a characteristically self-deprecating article published in 1980 [Busa 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, or of the Federalist Papers, (a set of anonymously published pre-American revolutionary pamphlets), and even the disputed works of the Russian novelist Sholokhov. At the same time, many research groups began to contemplate a more ambitious project which 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. *[Note: Holmes 1994 provides a good bibliography of earlier work; Juola 2006 reviews more recent thinking on the topic.]*

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 *[Note: For a persuasive historical analysis of this tradition and its development, see Leon 2008]* with the aid of massive quantities of machine-readable text. The emergence of the Brown Corpus in 1960 and its successors *[Note: 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/`*]* 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 which 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 modelling 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 lemmatisation — 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 looked for *ngrams* — patterns such as recurrent word (or token) sequences — but were not for the most part interested in indications of textual organisation 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, etc.) but the annotation of multi-word sequences, for example to indicate syntactic function, was more problematic and hence less standardised.

Nevertheless, the development of corpus linguistics as a defined area of research (a discipline even) owes much to the clear consensus amongst its practitioners concerning both core principles, methods, and objects which define the discipline, and those concerning which multiple points of view were recognised. For example, the Brown corpus instantiated a surprisingly long-lived model for the construction of language corpora which was based on fixed-size synchronic sampling of language production according to explicit selection criteria. In developing the Cobuild corpus 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 one American linguist as 'there's

no data like more data', 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, and because a consensus as to their 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...) 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 which might be both adequate to the immediate needs of one research project and generic enough to facilitate its re-use 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 laboured (and gained recognition for their labour) 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, *[Note: See for example, Denley and Hopkin 1987 or Denley et al 1989]* 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 pre-defined matrix, easily analysable by tools such as SPSS or SIR . Many others accepted uncritically the database orthodoxy proposed by their local computing centre (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 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. Nichols 1987). For the most part, however, few historians thought to follow their literary or linguistic colleagues in prefering 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, etc.) despite the reductionist view of the complexities of historical sources which 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 (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 new form of *Quellenkritik*. *[Note: See Greenstein 1991 for*

a collection of essays on the problems of modelling historical textual data sources.] And it is surely the realisation that a focus on how text itself should be modelled which resolved this dichotomy and showed the way forward for subsequent digitally-assisted humanities research.

3 The apotheosis of textual modelling

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 which 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 amongst 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, the essential remains to define precisely the bounds 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 which intelligent readers claim to perceive when reading a text ; the explicit claim was that by modelling 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 favour of a more profound consideration of the meaning and purpose of written texts. This opposition is most clearly stated in Coombs and Renear's foundational text on *The future of scholarly communication* (Coombs 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 *[Note: 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]* shows clearly the influence of contemporary database design orthodoxies. For example, a working paper from 1989 called 'Notes on Features and Tags,' still available from `http://www.tei-c.org/Vault/ED/edw05.htm`,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 labelled 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 determinate 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 XML schema currently used to define TEI components. This model, for which the TEI editors coined the name ODD (One Document Does it all), has 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. It is well documented elsewhere *[Note: The current system is of course 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.]* we highlight here a few of its salient characteristics, in particular those which 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 resembles a somewhat rambling collection of independent components rather than a single construct.

Each of the objects defined by the TEI has essentially the same set of properties : a canonical identifier, a description of its intended semantics in one or several natural languages, an indication of its equivalent in other ontologies, its classification within the TEI's own ontology, a list of its associated attributes (each of which is defined in a similar way), a formal model of its possible constituent components, usage notes and examples. None of this is inextricably linked to any particular enabling technology: although the first version of the TEI was expressed using the Standard Generalised Markup Language, later versions have used the W3C's Extensible Markup Language, and several experiments have shown the feasibly of re-mapping 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.

The constellation of textual features or objects defined by the TEI Guidelines corresponds with the set of 'significant particularitites' originally identified by the members of the TEI working groups, 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 not a fixed entity but one which has steadily evolved and developed in response to the changing needs and priorities of its user-community. This 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 élite, 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. Although the TEI predates 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 succesful user-focussed internet-mediated projects to have been created without benefit of today's 'social media'. As an internationally-funded research project, the TEI project conscientiously strove to give equal importance to researchers separated not only by discipline, but also by language and geography.

The interdisciplinary nature of the TEI model is 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 contained much material unlikely to be of interest to every reader , while at the same time every chapter contains material of importance to some reader. This 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, etc), 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 tagsets providing more specialised components, each reflecting a particular style of analysis (linguistic, hypertextual, text-critical etc.) Finally, a user might elect to suppress some of the components provided by a tagset, 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 favourite dish is typically constructed, also seems in retrospect to reflect something of the deeply balkanised 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 instantiated in TEI P3, the construction of a personalised 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 and the way that its parameter entities interacted as well as a good grasp of the structures within which the TEI declarations were organised. Unsurprisingly, many early adopters preferred to use a generic pre-defined subset such as TEI Lite (TEI 2012) or to rely on a view of the TEI provided by their own research community, such as the Corpus Encoding Standard (Ide 2000), or more recently the Epidoc Guidelines. Yet the existence of many such customizations, even those which 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 for the encoding of all texts for all time, but rather an architecture within which such standards could be developed in an interoperable or at least interchangeable way, a kind of agreed lexicon from which individual dialects could be derived. The architecture also envisaged a mechanism, known as a TEI customization file, which would enable one to determine how exactly that dialect had been derived, by specifying the TEI tagsets used, and setting parameter entity values to control which parts of their content would be included.

The major developments undertaken during the transition from TEI P3 to TEI P5 reflect a desire to maintain this objective. The transition from TEI P3 to TEI P4 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 *[Note: Technical details of the transition from P3 to P5 are provided in Burnard 2006 inter alia.]* was a far more ambitious process. Necessarily, it involved the addition of much new material and the updating of some outdated 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. 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. 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. At the same time, simple subsetting was made very much easier, and a simple web interface provided to achieve it. (Roma). A further change was to ensure that as far as possible a single language (ODD) was to be used for every aspect of the modelling process — no more embedded schema language fragments or *ad hoc* rules about how differently named objects should be processed.

4 Explicitness and coercion

We suggested initially that there is a long-running tension within all standardisation efforts consequent on an opposition 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 to have many different kinds of cake from the same messy dough. For this to work, there is a need for standards which not limit choice, but rather allow for accurate presentation of the choices made. Such an approach is also essential for a modelling standard which hopes to be effective in a domain where the objects of discourse, the components of the model, are constantly shifting and being remade, or remain the subject of controversy.

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 the expression is (say) a date associated with some event, and we wish to perform automatic analyses comparing many such. Simplifying somewhat, a TEI document may choose to normalise 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 calendar system entirely if the material concerned is all derived from some cultural context (ancient China, for example) in which dates are traditionally normalised in some other way. All three options are provided for by different sets of attributes in the TEI (@when, @when-iso, @when-custom etc.), each set being provided by a different attribute class (att.datable.w3c, att.datable.iso, att.datable.custom). These three attribute classes are subclasses of the class att.datable, as a consequence of which all the elements which are members of that class inherit all three sets of option by default. Of course, since it is most probable that in a given model only one of these sets will be used, that class may be customised to provide the attributes inherited from only one of the three schemes. And 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 associated with the documents in question.

Ever since its first publication, the TEI has been criticised 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 life is consequently 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 1996). The availability of tools like ODD cannot entirely remove those ontological anxieties, but at least they facilitate ways of coming to terms with them, of making them explicit.

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 favour of the one that interfaces most neatly with the latest most fashionable tool set. A similar 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, re-instating 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 standardisation; 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 so much that it has formalised and rendered explicit so many parts of the digital landscape, but rather that it has done so in a consistent and expandable way. Its value as a meta-model is essential to its usefulness as a modelling tool.

Bibliography

[1] Etienne Brunet 'Qui lemmatise dilemmes attise' in *Lexicometrica* 2 (2000) http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.html

[2] Lou Burnard 'New tricks from an old dog: An overview of TEI P5' in Lou Burnard, Milena Dobreva, Norbert Fuhr, Anke Lüdeling (eds) *Digital Historical Corpora - Architecture, Annotation, and Retrieval, 03.12. - 08.12.2006* Internationales Begegnungs und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl. 2006

[3] Lou Burnard 'The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure' in Journal of the Text Encoding Initiative Issue 5 2013 http://jtei.revues.org/811 DOI : 10.4000/jtei.811

[4] Lou Burnard and Sebastian Rahtz 'RelaxNG with Son of ODD' http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html in Proceedings of Extreme Markup Languages 2004

[5] Lou Burnard and Sebastian Rahtz 'Reviewing the TEI ODD system' in Proceedings of the 2013 ACM symposium on Document engineering 193–196 ACM New York, NY, USA 2013 ISBN: 978 1 4503 1789 4 doi>10.1145/2494266.2494321

[6] Roberto S. Busa, SJ 'The annals of humanities Computing: the index thomasticus', *Computers and the Humanities*, xiv , 83–90

[7] James H. Coombs, Allen H. Renear, and Steven J. DeRose 'Markup systems and the future of scholarly text processing' in *Communications of the ACM* Nov 1987 Vol30, No11, pp 933–47

[8] Peter Denley and Deian Hopkin (eds) *History and computing.* Manchester: Manchester University Press. 1987

[9] Peter Denley, Stefan Fogelvik and Charles Harvey(eds) *History and computing II.* Manchester: Manchester University Press. 1989

[10] Jean-Claude Gardin *Archaeological constructs*, Cambridge, 1980.

[11] Daniel Greenstein 'Modelling Historical Data: towards a standard for encoding and exchanging machine-readable texts' St Katherinen: Scripta Mercaturae Verlag. Halbgraue Reihe zur Historischen Fachinformatik herausg. von Manfred Thaller, serie A, band 11.

[12] Haugeland, John (1985), *Artificial Intelligence: The Very Idea*, Cambridge, Mass: MIT Press, ISBN 0262 08153 9

[13] D. I. Holmes, 'Authorship attribution', *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[14] Nancy Ide Jean Véronis *The Text Encoding Initiative: background and context* Dordrecht Boston Kluwer Academic Publisher 1995

[15] Nancy Ide and Greg Priest-Dorman *Corpus Encoding Standard.* Last modified 20 March 2000 http://www.cs.vassar.edu/CES/CES1.html

[16] Tom Elliott, Gabriel Bodard, Elli Mylonas, Simona Stoyanova, Charlotte Tupman, Scott Vanderbilt, et al. (2007–2014), *EpiDoc Guidelines: Ancient documents in TEI XML* (Version 8). Available: http://www.stoa.org/epidoc/gl/latest/.

[17] Steven E. Jones *The Emergence of the Digital Humanities* Routledge 2014.

[18] Patrick Juola, 'Authorship Attribution' in Foundations and Trends in Information Retrieval Vol. 1, No. 3 233–334 2006 DOI: 10.1561/1500000005

[19] William Kent *Data and Reality: Basic Assumptions in Data Processing Reconsidered* North-Holland Publishing 1978

[20] Timothy J. King 'The use of computers for storing records in historical research', *Historical Methods* 14 (1981), 59–64.

[21] Jacqueline Léon 'Aux sources de la « Corpus Linguistics » : Firth et la London School' *Langages* 2008/3 (n° 171) DOI : 10.3917/lang.171.0012 (`http://www.cairn.info/revue-langages-2008-3-page-12.htm`)

[22] Alan Macfarlane *Reconstructing Historical Communities*. Cambridge: 1977

[23] Thomas C Mendenhall 'The characteristic curves of composition' in *Science — supplement.* vol IX no 214 pp 237–246 11 March 1887 (online at `https://archive.org/details/jstor-1764604`)

[24] Jon Nichols et al 'Logic programming and historical research' in Denley and Hopkin 1987

[25] Andrew Prescott, 'The Imaging of Historical Documents'. In: Greengrass, M. and Hughes, L. (eds.) *The Virtual Representation of the Past*. Aldershot: Ashgate, 2008 7–22. ISBN 9780754672883

[26] Darrell Raymond, Frank Tompa and Derick Wood. 'From Data Representation to Data Model: Meta-Semantic Issues in the Evolution of SGML.' in *Computer Standards & Interfaces* 18 (1996): 25–36. doi:10.1016/0920–5489(96)00033–5

[27] J M Sinclair *Looking Up: an account of the COBUILD Project in lexical computing*. Collins ELT. 1987

[28] John F. Sowa *Conceptual structures*, Reading (Mass): Addison-Wesley, 1984.

[29] Text Encoding Initiative *Design Principles for Text Encoding Guidelines* Working Paper ED P1. 1988, revised 1990. `http://www.tei-c.org/Vault/ED/edp01.htm`

[30] Text Encoding Initiative *TEI Lite: Encoding for Interchange: an introduction to the TEI* `http://www.tei-c.org/Guidelines/Customization/Lite/`

[31] Manfred Thaller *: A Data Base System for Historical Research* Göttingen: Max-Planck-Institut für Geschichte 1987