TEI in a crystal ball[1]

Fotis Jannidis
Institut für Deutsche Philologie, Universität Würzburg, Würzburg, Germany

Correspondence:
Fotis Jannidis, Institut für Deutsche Philologie, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany.
E-mail:
fotis.jannidis@uni-wuerzburg.de

Abstract

Text Encoding Initiative (TEI) is an organization, a research community, and a markup language. Looking back into the history of these three TEIs, this article tries to describe what has been achieved and what its future challenges will be. The historical analysis is based on a closer look at the development of the TEI-L and topics covered by the Guidelines. A final section outlines possible roles of the TEI as an infrastructure for digital libraries and disciplinary virtual environments.

The title of my talk was inspired by a remark by David Birnbaum 17 years ago on the TEI-L discussion list: ‘I have no crystal ball, but 16-bit coding looks very much like the future for multilingual texts’.[2] Even without a crystal ball he was right and I cannot even hope to foresee into the future as well as he did. Looking into a crystal ball to see the future is much easier if you know your customer. Doing this for an institution is much more difficult and requires research into the past: how was the TEI conceived when it was first developed and what it has become over the past 20 years? What is missing from the Guidelines and what is superfluous? Are there large trends in the digital world sweeping TEI along with it; are there smaller trends in humanities computing? Or are there specific subjects that will assert the most influence on the TEI in future? Talking about TEI in the future implies that we know what TEI is. Our answers to the aforementioned questions will be different for different aspects of TEI. Very obviously the three letters stand for a growing collection of tags, or rather the concepts behind these tags. As Elly Mylonas and Allen Renear pointed out at the last birthday celebration at Brown University, the TEI has also become the synonym for a research community in the digital humanities (Mylonas and Renear, 1999). Last but not least, the TEI is the name of an international consortium with subscribers all over the world—one of the very few humanities organizations that successfully operates at that scale. In this article, I will examine the development of the TEI from these perspectives and attempt to extrapolate some future trends.

1 The Text Encoding Initiative (TEI) as an Organization

The TEI began as a project financed by granting agencies.[3] After the grants that made the initial stage of TEI development possible were exhausted, those involved created the Consortium. Financed by its members, the legal, financial, and technical framework of the Consortium affords an important element guaranteeing the sustainability of the TEI.[4] As with other ‘standards’ in the Internet age, these have to be developed to accommodate new factors in their environment, new user wishes, and new technical possibilities. They have to be promoted to the ever increasing number of people working on the web and the digital world in general; and they have to be supported to close the gap between the technical definition, which is general, terse, and often difficult to access, and the user who wants to solve a special problem and to avoid a steep learning curve. In this context, standards are not everlasting but allow temporal stability and a

doi:10.1093/llc/fqp015 Advance Access Published on 14 May 2009

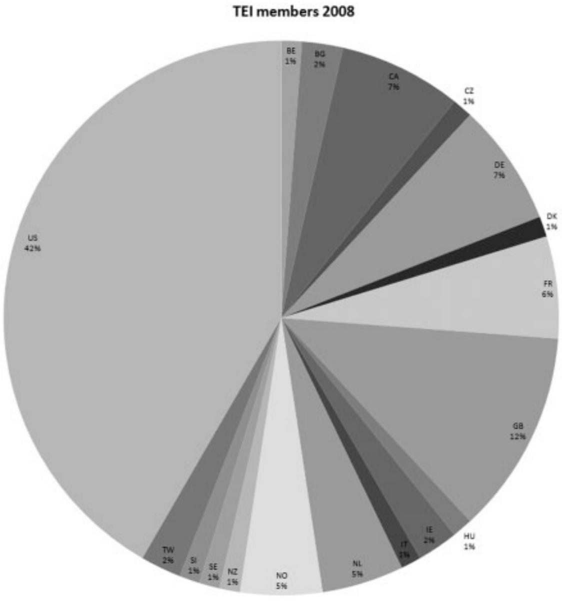

Fig. 1 International distribution of TEI members

controlled transition from one point to the next. For all these tasks, there is the need for an institution to fulfil them, and the TEI Consortium looks like a good solution to do so. For the humanities, this is a new situation and the creation of a consortium an innovative solution, at least from the German perspective. In the USA and the UK, there seems to be a greater familiarity with this kind of solution, judging from the chart (Fig. 1) that shows where the countries' TEI members come from.

Today the TEI has eighty-four institutional members from eighteen different countries on four continents (Fig. 1).[5] This is impressive but the distribution of the institutional members doesn't represent—to my knowledge—the adoption of TEI in these countries.

The fact that more than a half of all institutional TEI members (61%) are from the USA, Canada, and the UK certainly reflects differences in the way and degree humanities computing is integrated into traditional humanities; in the way libraries in the different countries picture their role in the digital age; and differences in the speed some libraries in some countries understand their new role. Most libraries and archives don't have the tradition of taking an active part in developing standards. Moreover, during periods when they have to cancel subscriptions and buy fewer books, they shy away from spending substantial amounts of money for something that doesn't look like their core activity.

As short-sighted this may be, a change of attitude will take time and it will involve a new definition not only of the role of libraries and archives but also for the development of new research institutions for digital humanities. It also means that scholars have to include activities like the development of humanities related IT standards into their job descriptions and reflect this fact in their decisions on tenure track positions. Probably it is even more important to communicate these new institutional needs to those agencies that usually provide scholars with funding. In Germany, funding agencies tend to limit themselves to supporting primary research and we have to come up with new ways to finance the infrastructure for the digital humanities, an infrastructure that encompasses services, tools, data, support, and standards like the TEI.

This is a learning process for all involved, for the libraries and archives, the funding agencies, the universities, and the scholarly communities, and it may take some time before a chart like the one above looks a bit more like a birthday cake.

Nevertheless it may be useful—in addition to the existing structure—to describe clearly defined research goals in the context of the TEI and to try to obtain funding for them. As this is only a temporary commitment, in many countries it probably will be easier to obtain this funding and it will ease the financial burden of the Consortium that have to maintain, adopt, promote, and support one of the first international standards in the humanities.

2 TEI—A Research Community

It has been pointed out repeatedly that TEI is not only a technical success story but also a social one, and that the technical success is closely related to the social network. We all know this social network exists but how can it be demonstrated? A very

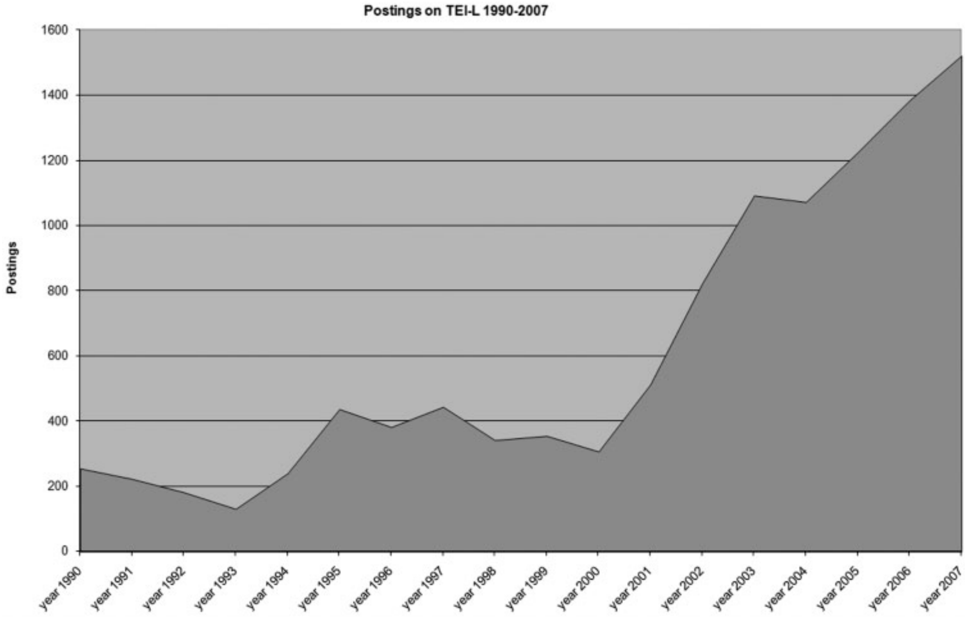

Fig. 2 Postings on TEI-L 1990–2007

crude way is to have a look at the discussion lists, TEI-L and the now defunct TEI-Tech.[6]

Figure 2 represents the numbers of postings on the TEI-L list between 1990 and 2007, in total 10,928. We see that around the release of P3, there was a peak of activity but afterwards it stayed at a comparatively high level and started to increase steeply after 2000. But this is just the activity, not the people. Exactly who contributes to the lists?

The numbers in these figures are just rough approximations. Some very unsophisticated perl scripts collected the data and I did not post-process them in any way—so Lou's laptop is a rather active member of the list next to Lou Burnard himself. The number of people present in the community rose around the release of P3 (Fig. 3). Actually it seems that the release of P3 triggered significant activity and attracted new people but then the numbers decreased; you could say, normalized, until 2001 when a new rise began. But the number of active posters is not nearly rising as rapidly as the number of postings. This means that people already involved in the TEI are writing more about it. This could be understood as a sign that TEI is becoming a full-time job for more people. But upon closer inspection, another reading is possible: some people are investing more and more time into helping others, talking about, and discussing TEI.[7]

If my figures are accurate, more postings are written by fewer people (Lou's laptop included). Figure 4 shows how many of the postings are created by the 10% of most active posters.[8] Between 1998 and 2003, there has been a steady rise in the amount of emails of the most active members of the TEI-L community. So the chart clearly suggests that the burden on them is growing instead of being reduced as more people become involved in the TEI.

But what does this mean? Let us assume for a minute that the TEI is being used more widely and more people have started adopting it. If this is true, then we can safely say there seems to be a problem integrating all these users into the TEI research community (if we accept activity on the TEI-L discussion list as an indicator).

But possibly calling the TEI a research community is a misnomer. If you look at the postings, you will see many of them relate to issues that are not really about what we would call 'research'. Most of the time practical questions are discussed, questions like how to use a specific element, attribute, etc.

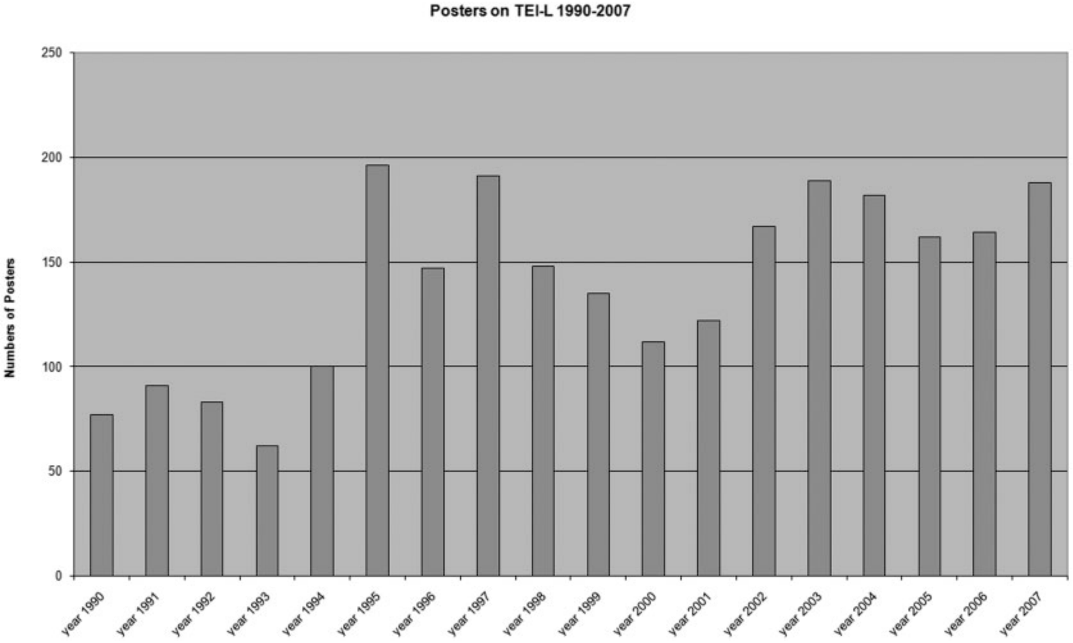

Fig. 3 Numbers of posters on TEI-L 1990–2007

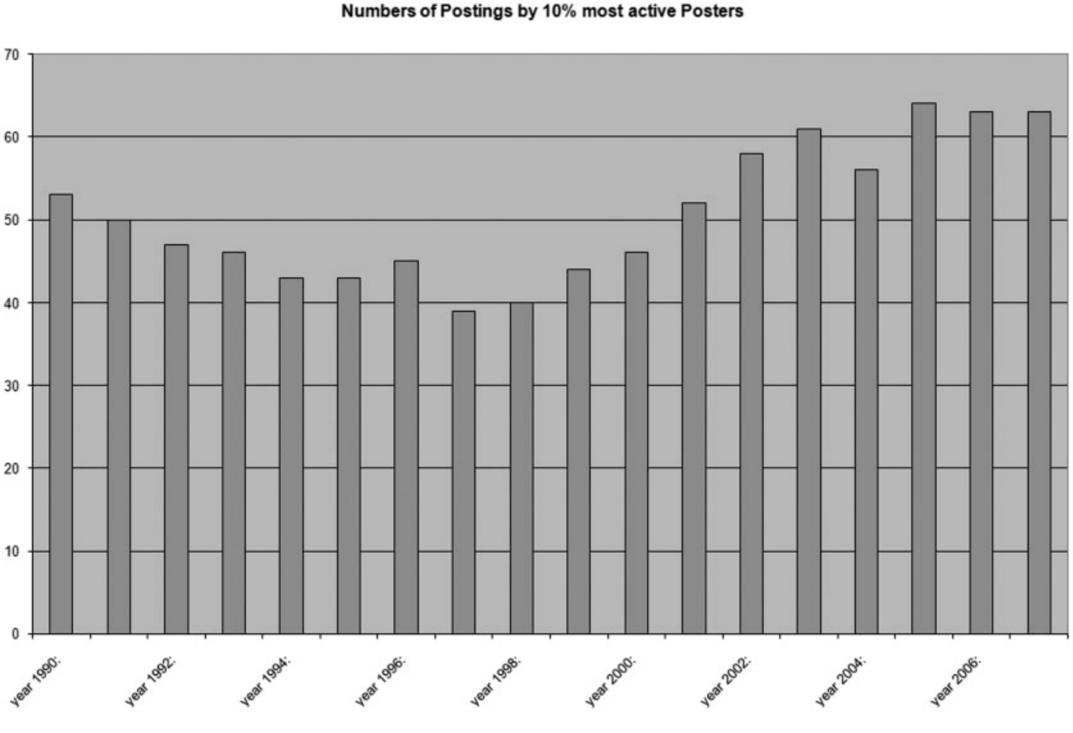

Fig. 4 Numbers of postings by the most active contributors (highest 10%)

Many of the answers come from people encoding similar texts with a similar perspective—what we usually call a research field. It is also interesting to note that some names of scholars well known for their contributions to the theory of textual markup, names like Dino Buzetti, Jerome McGann, Allen Renear, or Peter Shillingsburg are rarely to be found or are missing altogether as active posters.[9] So maybe TEI is not a research community itself but rather a place where many research communities such as corpus linguistics, medieval studies, modern literary studies, and others overlap to discuss more practical problems.[10] Does this tell us anything about the future of the TEI?

If there is something like a research community, then an attempt to integrate researchers and to mobilize them into participating actively should be successful. And the numbers of active members should significantly increase in the coming years.

If—on the other hand—TEI is not a research community but mainly the intersection of different communities, a market place for (mostly practical) solutions to markup problems, then we probably won't see such a significant rise in the numbers of active members because most people will use TEI to achieve specific goals and will only come to the TEI community to solve problems they can't find a solution to themselves. In my experience of many different projects in the digital humanities over the last 10–15 years (mostly in Germany), I have had the impression that many of them hesitate to contact the TEI community and sometimes even choose a home-grown solution destined to be less stabile and never to be maintained. There are some practical reasons for this; many scholars are not very familiar with email-based discussion lists and the language barrier is an impediment, but the main problem seems to be one of perception: many people perceive the TEI as a community to which their respective communities do not intersect.

The TEI may be thought of in many ways as being in a similar position as that of textual criticism: it is part of many different disciplines, such as literary studies, history, philosophy, but its role differs in those disciplines and although textual criticism in these disciplines has a lot in common, there are also very marked differences. Textual criticism in German philology is in many important ways quite different from textual criticism in English studies, and the same is true for the relation of text encoding and its relation to the scholarly field making use of it: text encoding—even text encoding using the TEI—inherits basic assumptions that are quite different depending on the fields making use of it.

So if we define a research field by a common object—or a common construction of this object—and a set of methods shared by all in this field, if a research field is also defined by a common research infrastructure including university institutes and positions, conferences, book series, journals, and other forms of organized exchange, then the TEI is not a research community yet, but it certainly is a community.

Nevertheless some practical questions remain: how do we harvest the knowledge of those working with digital texts to find the best concepts to describe and encode them? The most effective ways of aggregating knowledge at the moment are the strategies used by open source communities and Wikipedia—both are in use now for the further development of the TEI.[11] The archives of the discussion lists alone are a treasure trove for anyone new to the mysteries of scholarly encoding.[12] All this makes TEI an important role model for new forms of knowledge production that are in use elsewhere but which the humanities are slow to adopt. The technical complexities of using these tools still overwhelm many newcomers and it is easy to foresee that an expert knowledge base will be created in the near future. Many of the elements of this knowledge base and its distribution mechanisms are either already in existence or planned like organized TEI training courses, a content-rich TEI FAQ and *TEI by example*.

3 TEI as a Set of Concepts and a Set of Tags

The most important meaning of TEI, the main reason the community and the organization developed and blossomed is without any doubt due to the set of tags and their underlying concepts.

In the last 20 years, TEI has become a *de facto* standard for literary computing. It is used in many digital editions and in some corpora.[13] Its importance for textual markup in the digital humanities is also codified by its prominent place in modern introductions into this field,[14] and the fact that the TEI is endorsed by many agencies involved in funding digital libraries and electronic text projects.[15]

The history of the TEI as a set of concepts and their expression in a specific markup will be rich as a source for future historians. The way this history reflects the changing understanding of markup in general is one of special note. The same is true for the instruments to handle the growing complexity of the guidelines as a set of texts, schemas, and software to produce output formats.[16]

In the following, I can only point to some of the more obvious trends of this multilayered and complex development over the last twenty years and I will concentrate on the way TEI uses text types as a means to articulate the complexity of text, and the relation of TEI to other standards.

3.1 Text types

The following list shows the TEI versions and the respective number of tags:

TEI version	Number of tags
P1 (1990)	163
P2 (1992–93)	354
P3 (1994)	440
P4 (2002)	441
P5 (2007)	504

At first glance, there seems to be a continuous, but not too excessive growth of knowledge, of concepts, and of the numbers of tags. Based on these numbers, P10 should have about 1,000 elements. But numbers can be misleading. Between P4 and P5, only about sixty new elements were added. But the numbers belie the significance of the changes as approximately eighty elements that are part of P4 were not incorporated into P5. In other words, the TEI tag set has not only grown but it has also changed significantly over time and will, no doubt, continue to do so.

The main reason for this change is the fact that TEI can be seen as the intersection of the perspectives of different disciplines engaging with textual markup. In most of them, the needs users have are becoming more diverse, new needs arise, and old ones become more complex. One of the resulting problems of the TEI is its complexity: it needs to accommodate many different disciplines in encoding their knowledge in a way that is granular enough for their specific semantic and textual practices. But its complexity is also the main reason many people would rather learn how to write their own schema than learn the TEI.

This problem seems to have been obvious to many of those actively engaged in the TEI. The pizza model, Roma, teiLite, all these were and are ingenious solutions to this problem. But there is a difference: using the Pizza Chef or Roma demands a basic understanding of the whole of the TEI. You have to know the outline of the TEI to decide what you don't need. teiLite, on the other hand, was a small and self-contained subset. This is, no doubt, one of the reasons it has been so successful. Defining subsets for more sophisticated scholarly text types together with a subset of the documentation could make the TEI more accessible. And referencing these subsets in the TEI header so that people can use them to access the texts, for example in online searches, would make them even more useful.

I want to focus on just one aspect, i.e. grouping elements by function drawn from the P5 version of the Guidelines

Metadata
Organization of the TEI header

Text types
Elements available in all TEI documents
Default text structure
Prose
Verse
Performance texts
Dictionaries

Meta text types
Transcriptions of speech
Manuscript description
Representation of primary sources
Critical apparatus
Language corpora

Textual aspects/elements
Representation of non-standard characters and glyphs
Names, dates, people, and places
Tables, formulae, and graphics
Linking, segmentation, and alignment
Simple analytic mechanisms
Feature structures
Graphs, networks, and trees

Text types and tag groups in TEI P5

There are text types like drama, prose, etc. (and in this category I included all tags shared by all text types). But we also have scholarly text types like critical editions, dictionaries, or language corpora that function a bit like a wrapper for the first. They communicate the subject-specific approach to those simple text types. And there is a last group that encodes specific aspects of the text, such as persons, glyphs, tables, etc.

A look at these categories in the P1 version of the Guidelines shows that early on the TEI as a community had more ambitious goals.

Metadata

Text types
Features common to many text types
Literary texts
Office texts
Dictionaries and lexica

Meta text types
Critical apparatus and parallel texts
Language corpora and other collections
Transcripts of spoken texts

Textual aspects/elements
Characters and character sets
Bibliographic control
Basic non-structural features
Appearance features
Figures, tables, formulas, and diagrams
Bibliographic references
Editorial comment and emendation
Ambiguous punctuation
Reference systems
Cross references and text links
Analytic and interpretive features

Text types and tag groups in TEI P1

We can see that the list of scholarly text types handled by the TEI already exists in P1. The list of textual aspects is longer here, and some of them are less current such as the issues surrounding complex character encoding. Other aspects seem to have too much weight such as ambiguous punctuation that is currently treated in a short subchapter. Similar to bibliographic references, it is now part of the long list of elements available in all TEI documents. Probably of most interest is the list of text types that the TEI intended to cover: not only literary texts but also office documents.

If we look at the structure of P2 we see that ‘literary text’ has been substituted by a list: prose, verse, drama, and letters, while the office documents were dropped.

Metadata
TEI header

Text types
Tags available in all TEI DTDs
Default text structure
Base tag set for prose
Base tag set for verse
Base tag set for drama
Base tag set for letters and memoranda
Base tag set for printed dictionaries

Meta text types
Composite texts and combining bases
Base tag set for transcriptions of spoken texts
Base tag set for terminological data
Additional tag set for language corpora
Manuscripts

Textual aspects/elements
Segmentation and alignment
Simple analytic mechanisms
Feature structure analysis
Certainty
Analytic bibliography and physical description
Text criticism and apparatus
Names and dates
Graphs, digraphs, and trees
Graphics, figures, and illustrations
Formulae and tables

Text types and tag groups in TEI P2

Looking back to the P5 chart, we see that the list of explicitly supported text types today is quite short in comparison: no office documents, and still, after 20 years of the TEI, no letters and no tag set for the physical description of books.[17] When John Lavagnino points out that many text types that are not explicitly mentioned in the TEI Guidelines can be encoded using the TEI by using a similar text type (mostly prose) and by adding the elements necessary for the purpose but not included in the Guidelines, he certainly presents a position shared by many in the TEI community (Lavagnino, 2006). But this happens at the cost of interoperability: there is thus some urgency to support these text types and the markup of textual aspects important to scholars in these fields.

Thus, one possible future for the TEI is to include more of these text types. Without any doubt, this will make the TEI even more complex and will keep the rate of growth steep. Is this necessary? Well, in the extreme, you could express almost anything with just one block and one inline element and put all other information into attributes. On the other hand, you could have thousands of elements. Finding the most sensible tradeoff is not easy. And this is not only a technical decision. The naming of elements and their description in the TEI guidelines is most of the time informed by a special perspective on texts: quite frequently you hear a linguist or a literary studies person talking.[18] To actively embrace other disciplines in the humanities and to invite them to collaborate in future development is the obvious thing to do, and looking at the list of speakers for the TEI conference in 2007 this has already begun.

Another possible future for the TEI could be that in the long run, it is only responsible for the meta types because there are so many different text types and the active management of all of them is too complex. Another possibility is that the TEI is responsible for everything, or, more probable, the TEI is responsible for many meta types in the Humanities and for those text types that draw enough interest from within the TEI community. But that seems to me to be the real problem: how does the TEI community embrace newcomers who want to handle new text types? There is always a tension between a subject- or usage-specific approach to a text—like the way linguists handle texts or librarians—and a general markup system aimed at all texts relevant for the humanities.

3.2 Neighbours

Early in 2007, there was a meeting of the TEI Council in Berlin. A TEI Day accompanying the meeting afforded the opportunity to present to the Council a number of German projects that utilize the TEI. During the discussion of one of these projects, a question arose; a question that had obviously baffled more than one of the German presenters: How do I encode a collection of texts like an anthology with different text authors? The collective wisdom of the Council smiled and said: ‘ah—that’s an easy one: you use’, and at that moment the collective voice suddenly split into two: ‘you use METS’[19] said one of them, ‘you use teiCorpus’ said the other. Everyone who has ever read the TEI lists knows about the many possibilities that exist for encoding most textual features using the TEI. But having a collection of texts by different authors and in different text types is so common that you would have expected that there has to be a generally acknowledged way to handle this. But what was even more striking for any onlooker was that there was no consensus about whether this was a job that should be handled by the TEI or by another standard.

The relation of TEI to other standards—as far as I can judge—has always been a very open one. If there is a standard managing something in a way useful to the humanities community, it was taken up and integrated.[20] Actually, the very first posting archived on the TEI-L discussion list is a question about another standard. Robin Cover asked:

> Does anyone have access to ISO 639, and would you be willing to summarize the information on language codes? ISO 639 bears the title (ca.) _Codes for the representation of names of languages_, and according to the ANSI office in New York, was updated in 1988. [...] A summary of language codes would probably be worth posting publicly.[21]

The email was answered by the Chairman of the Registration Authority for this standard, who pointed out that it 'is considered inadequate for bibliographic purposes' and a working group has been set up to create a new standard. In view of the anecdote related above, it is a bit ironic that the very next email asks: 'Could someone please explain the TEI approach to compound documents and images? Will SGML be used here, and if so, how?' [22]

But to return to the issue of TEI and other standards: Unicode, Xlink, and Xpointer, just to mention a few, all solve problems that the TEI had solved in earlier versions of the Guidelines. Some of these developments are inspired or directly influenced by the TEI or are the direct product of a collaboration as the joint activity of the TEI together with an ISO workgroup.[23] In the former case, the TEI delegates specific tasks to other standards. In the last case, there is an overlap of TEI and another standard.

Sometimes the collaboration with another standard is easier said than done. One example is the question how to relate the TEI header to existing bibliographical standards. At the end of the chapter on the teiHeader there is a short section, a 'Note for Library Cataloguers':

> It is the intention of the developers, however, to ensure that the information required for a catalogue record be retrievable from the TEI file header, and moreover that the mapping from the one to the other be as simple and straightforward as possible. Where the correspondence is not obvious, it may prove useful to consult one of the works which were influential in developing the content of the TEI file header.[24]

Then there are references to The International Standard Book Description, the Anglo-American Cataloguing Rules, and the ANSI Z.39.29, concluding with the sentence: 'Other relevant standards include BS 1629:1989, BS 5605:1978, and BS 6371:1983.'

The combination of 'simple and straightforward' and the reference to six standards just by their names not only have some comic aspects for non-librarians but it also shows a clash of different cultures. Even if this is really as simple and straightforward as possible for librarians, it certainly is a deep mystery for most scholars who want to make their TEI headers accessible to catalogues in the most effective way, and who want to provide the necessary information without becoming librarians. Standardized headers and best practice examples are probably all we need here and maybe a mechanism to inform others which profile has been chosen.

The relation of METS to the TEI is probably a bit more difficult. There are good arguments to choose one over the other, but a literary scholar new to the guidelines will be puzzled: in the guidelines, compound texts are handled by 'teiCorpus', an element which by its name signals its provenance from linguistics (and the literary scholar has yet to find his way to this section); while the use of METS isn't mentioned at all—quite in contrast to the way XLink is referenced. And METS is in itself quite complex that may prove too high a barrier for many smaller digital edition projects to implement.

Despite these inconsistencies, the relation of TEI to other standards is probably as explicit as it can be considering the state of flux of most standards and the ongoing development of new ones. In the long run, TEI will probably leave other tasks to new standards. References to people, to named entities may be such a case, but perhaps the new mechanism to encode names in P5 will also be successful in other fields not interested in textual markup. As the people managing the TEI have done a remarkably good job in providing a mechanism for integrating new standards into the TEI, TEI's future seems to be bright from this perspective.

3.3 Sidenotes

Two small remarks at the end of this section: we still await a TEI module to encode rhetorical aspects of texts or narrative features as they are studied by narratology. Although there is a mechanism for encoding these kinds of features, i.e. 'Simple Analytical Mechanisms' (chap. 17) and 'Feature Structures' (chap. 18). These mechanisms are very powerful but they have an interesting unintentional result that the subject-specific knowledge (linguistic language analyses, stylistic view on language, narratological view on a text, etc.) is not part of the Guidelines and would need another home to be standardized and documented. Is this really the best path forward?

The TEI has done a significant amount to make it possible to create digital editions. In some way, the TEI is like a promise that our data will be useful in 100 years if we chose the right format. But we all know that there is more to an edition than to encode a text. In designing an edition, we also create ways to approach the text and to interact with it, in short visualization and functionality, i.e. algorithms. We all know enough examples of digital editions that demonstrate very clearly that editors spend a lot of time thinking about the right interface for their data, how to communicate their insights and their special points of view in the most effective ways, and how to offer special ways of retrieving the text.[25]

Most of this information will be lost in a few years. It would be very useful if we could archive together with our text an abstract description of the user interface and its functionality. The TEI header would be a good place for something like this if there were a useful way to describe this kind of information in such a way that it can be utilized mechanically. There is a lot of room and need for development in this area.

4 Web x.o—or the Global Library and Its Books

Most people involved in the TEI community know that the information docuverse is changing dramatically, but we don't know the exact nature of this change nor how to bring our less tech-savvy colleagues into these new modes of scholarly communication. In many countries, there are very ambitious projects to build up a national cyberinfrastructure. In the UK, there was the Art and the Humanities Data Service[26]—a showcase ehumanities center *avant la lettre*. In the Netherlands, there is DANS[27] in Germany where there are large Projects like eSciDoc[28] or Textgrid[29] that aim to be the building blocks for such a structure. In the USA, there are similar endeavours under way as the report of the American Council of the Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences demonstrates.[30] Basically this means, after the client–server architecture of the first computer generation, the local applications of the PC generation, and the webserver applications of the WWW generation, we now have a service-based architecture allowing people to plug together their own ensemble of data sources and functions via web services.

One of the main goals of these initiatives is to set up an infrastructure that allows easy access to data and services without bothering where the software and where the content (text, film, images) is. Until now, the main goal of a digital humanities project was to put its data online. TEI was mainly used as an archive format to provide sustainability or as an exchange format for the larger collections in digital libraries. Some even used it as working format creating their editions in TEI from the beginning. Only in very rare cases the TEI encoded texts were accessible to the reader and user of an edition. Interoperability was important only for projects or digital libraries providing a common interface to their resources.

The basic idea of building up the new infrastructure is to have interoperability between different projects, digital libraries, etc. And this interoperability does not only exist at the data level but also on the level of functions and of services handling the data. We can expect to see a common interface to data and services that will allow users to access them as needed to solve specific problems. From the point of a linguist or a literary scholar, the architecture might look like this:

The working environment of the scholar would show him in a transparent way[31] the available resources and service and allow him to assemble the component parts as per his needs. One of the interesting aspects of this infrastructure would be the fact that every time a scholar has invested some time and knowledge into adding information to a text, this could be the starting point for someone else. Maybe the first has created catalogue data and digitized the text, the second scholar may then provide a transcription of a facsimile, or to add markup to an existing digital source and to

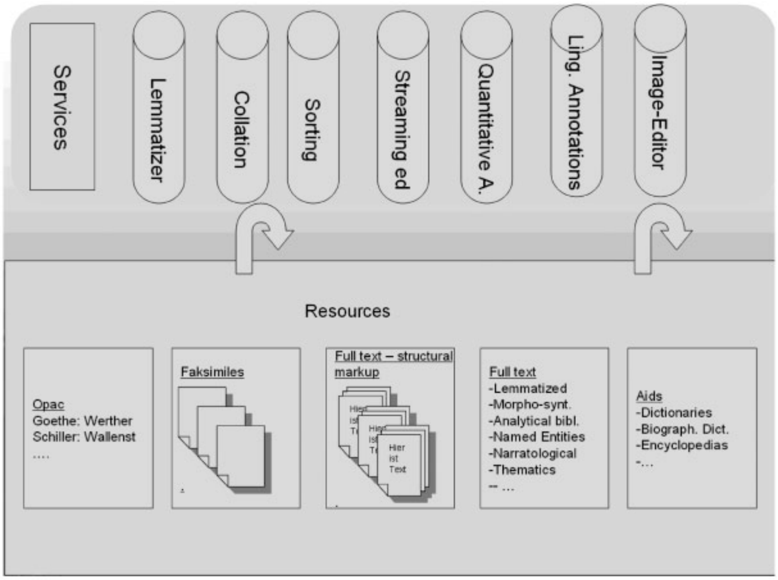

Fig. 5 Schema of a future working environment of a scholar

analyse the text from some discipline-specific point of view (Fig. 5).

The term 'ratchet effect' has been used in many disciplines to describe the 'phenomenon that some processes cannot go backwards once certain things have happened'.[32] The ratchet allows the wheel to move in one direction but every time it clicks it prevents it from moving backwards. Like a ratchet, so evolutionary psychologists assume, culture allows the accumulation of information. People born into a culture contribute to the movement of the wheel but culture itself stops the loss of information. The famous notion of *a dwarf standing on the shoulders of a giant* and thus seeing more than the giant has often been applied to the field of science. It expresses a similar concept of an accrual of information. Humanities find it notoriously difficult to accrue knowledge in a similar way. It is my belief that with an architecture, as these projects have, it will be possible to achieve similar effects in the realm of textual studies.

4.1 What does this mean for the TEI?

First of all (and there is nothing new here), many projects and libraries will use TEI. Second, one can use the knowledge in TEI to solve some of the problems in the approach outlined. What could such an interface possibly provide which isn't done better and faster by Google? Google, at the moment, does exceedingly well in providing ranked results of full text searches including all the millions of books Google has digitized. As the Internet consists of text coming from all kind of sources, this may be for some time the best we can expect. But this full-text search can't differentiate between names and places; it doesn't know whether something is the title of a book, or just mentioned in a text. But in an environment that is much smaller but controlled, the world of the digital libraries and archives, this seems to be an unnecessary loss of information. There are many interesting ideas as to how to change this situation, for example the vision of the semantic web. But maybe the more complex proposals demand too much even for an e-humanities environment. In the Textgrid project, we use text types as a concept to provide additional information that allows the reuse of texts in a more intelligent way.[33] Let's take the example of a dictionary: a full text search, mainly oriented along the physical document structure, can be very useful. But searches can be improved dramatically based on an understanding of the way a dictionary is structured. For example, you can limit the search to those entries that encode the search term as the lemma.

This sounds like something that has been done before: creating a subset of TEI, declaring it to be if not best, at least a good practice, and building software to interface with this data structure.[34] The only difference this time is that we dismissed the idea of the best practice. Rather, we defined a format that will not be at any time the project format but just an internal representation of the lowest common denominator. On the other hand, it supports those forms of access to a text type that are most often used. Basically from all texts a dumbed down version or view is created that allows us to access them using the basic text type information.

Another problem is how to provide links in such an environment. The main question seems to be whether we can find a way to target text below the level of the whole text in a way that is not technically defined but by the conventions people use for texts. It would be very nice if we could link to the first page of the first edition of Franz Kafka's *The Trial* without having to bother how this page has been encoded.

Coming from this application of the TEI, the integration of more text types into the Guidelines seems to be a promising way to extend the uses of the TEI.

5 Conclusion

The TEI Community has quite a lot to do in the coming years—at all levels of meaning which represent these three letters. But it has an impressive track record and if there is a similar conference in 20 years—and I would bet that there will be a similar one—and things continue to progress, TEI will still be the role model for 'doing' humanities in the digital world.

References

Burnard, L. and Sperberg-McQueen, C. M. (1995). The Design of the TEI Encoding Scheme. *Computers and the Humanities*, **29**(1): 17–39.

Cummings, J. (2008). The Text Encoding Initiative and the Study of Literature. In Schreibman, S. and Siemens, R. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 451–476. Available at <http://www.digitalhumanities.org/companion/DLS/> (accessed 10 March 2009).

Hockey, S. (2000). *Electronic Texts in the Humanities*. Oxford: OUP.

Lavagnino, J. (2006). When not to use TEI. In Burnard, L., O'Brien O'Keefe, K., and Unsworth, J. (eds), *Electronic Textual Editing*. New York: Modern Language Association of America, pp. 334–38.

Mylonas, E. and Renear, A. (1999). The Text Encoding Initiative at 10: Not just an interchange format anymore—but a new research community. *Computers and the Humanities*, **33**(1–2): 1–9.

Renear, A. (2004). Text encoding. In Schreibman, S., Siemens, R., and Unsworth, J (eds), *A Companion to Digital Humanities*. Malden, MA; Oxford, Carlton, Victoria: Blackwell, pp. 218–39. Available at <http://www.digitalhumanities.org/companion/> (accessed 10 March 2009).

Vanhoutte, E. and Van den Branden, R. (eds) (2003). *Dalf Guidelines for the Description and Encoding of Modern Correspondence Material Version 1.0*. Gent: CTB-KANTL. Available at <http://www.kantl.be/ctb/project/dalf/dalfdoc/> (accessed 10 March 2009).

Notes

[1] This text is based on a plenary talk given at the TEI meeting in Maryland 2007. Although it has been revised, it keeps the characteristics of a talk. Many thanks to Susan Schreibman for a thorough revision of the text.

[2] David Birnbaum, email on TEI-L, Fri, 5 October 1990 18:50:00 EDT. The archive of the TEI-L discussion list is accessible at http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L. (accessed 10 January 2009).

[3] The history of textual markup is a desideratum; on the history of TEI, see the short outline at http://www.tei-c.org/About/history.xml.

[4] Other elements that are also quite important include publishing the TEI under a well-known open source licence and using an open source process of information gathering.

[5] Source of the numbers is the TEI website: http://www.tei-c.org/Membership/current.xml (accessed 3 January 2009).

[6] My thanks to Syd Baumann, the maintainer of the TEI-L list, for making the archive available.

[7] The numbers of postings and posters are difficult to determine because TEI-Tech was established in 1992 but became inactive in 2003. Currently all discussion takes place on TEI-L. Since there were not that many postings on TEI-TECH, it was ignored for this study.

[8] I don't need to list their names as all involved in the TEI know them. And we are very thankful for their dedication.

[9] That doesn't imply in any way that some people whose names can be found repeatedly on TEI-L have not contributed significantly to the theory of markup.

[10] The same could be said for the field of textual markup in general (and not only concerning scholarly text types like the TEI), which has a specialized journal, first Markup Languages (1999–2001) and later the Proceedings of Extreme Markup Languages (2001 ff.); see http://www.idealliance.org/papers/extreme/proceedings/index.html.

[11] But the community hasn't really adopted these new tools yet; see for example the short histories of the wiki http://www.tei-c.org/wiki/index.php/Main_Page or the comparatively short list of bugs and feature requests on the sourceforge website: http://sourceforge.net/projects/tei/(accessed 18 September 2008)

[12] See http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-L.

[13] The TEI website lists 104 projects (August 2008), many of them being quite substantial in volume and visibility; see http://www.tei-c.org/Activities/Projects/index.xml.

[14] c.f., Hockey (2000, 36 ff.), Renear (2004), and Cummings (2008).

[15] For example, the US National Endowment for the Humanities, the UK's Arts and Humanities Research Board, the Modern Language Association, the European Union's Expert Advisory Group for Language Engineering Standards; see *Impact of the TEI* http://www.tei-c.org/About/history.xml.

[16] For the design principles of ODD (one document does it all) cf. Burnard and Sperberg-McQueen, 1995. In P5, the language to encode this document became a module of the TEI; see chapter 22 of the guidelines (P5) on 'documentation elements'.

[17] There are, however, now special interest groups dedicated to developing these modules for these and others which can base its work on forerunners such as the DALF project for the markup of letters; see Vanhoutte and Van den Branden (2003).

[18] 'teiCorpus', for example, is supposed to be used to also encode anthologies and similar text groups, but the name for the tag very clearly points to the linguistic background of the concepts behind it.

[19] METS: The Metadata Encoding and Transistion Standard; see http://www.loc.gov/standards/mets/.

[20] The W3C standards XLink and Xpointer substitute the linking mechanism provided by the TEI in earlier versions. The same is true for the id and the lang attribute, which are now replaced by the corresponding attributes from the xml specification xml:id and xml:lang.

[21] Robin C. Cover on TEI-L, 8 January 1990 16:08:22 CST

[22] Robert Philip Weber on TEI-L, Wednesday, 31 January 1990 09:33:26 CST

[23] The list of experts contributing to the ISO/TC37/SC4 working group on 'Terminology and other language and content resources' includes many names known for their work in the TEI.

[24] TEI P5 Chapter 5.7 Note for Library Cataloguers; see http://www.tei-c.org.uk/Vault/GL/P3/HD.htm#HD8

[25] Quite typically the book *Electronic textual editing* (see footnote 15) does include screenshots of the digital editions but the editors didn't see the need to include a section on the visualization of textual data, timelines, etc. A collection of impressive example how a collaboration of a information visualization and a textual criticism could look like can be found here; see http://www.item.ens.fr/index.php?id=173027.

[26] AHDS, see http://ahds.ac.uk/. For some obscure reason the AHDS stopped to receive funding when in many other nations people start to build up similar structures.

[27] DATA Archiving and Networked Services; see http://www.dans.knaw.nl/en/.

[28] eSciDoc; see http://www.escidoc-project.de/JSPWiki/en/Startpage.

[29] Textgrid; see http://www.textgrid.de/.

[30] See http://www.acls.org/programs/Default.aspx?id=644. C.f. now also the Bamboo project funded by the Andrew W. Mellon Foundation; see http://projectbamboo.org/.

[31] 'Transparent' refers here to the fact that it is immaterial to the user where the resource 'really' is.

[32] See http://en.wikipedia.org/wiki/Ratchet_effect (accessed 1 August 2008).

[33] A first report on this baseline encoding can be found here (only in German) http://www.textgrid.de/fileadmin/TextGrid/reports/TextGrid_Kerncodierung_070615.pdf.

[34] On a more formal level you could describe the problem as a mapping between two formats, a source and a target vocabulary; c.f. Michael Sperberg-McQueen: Thinking about schema mappings. http://people.w3.org/~cmsmcq/blog/?p=88 (accessed 18 September 2008).