The Present and Future of the TEI Community for Manuscript Encoding

Marjorie Burghart, Malte Rehbein

Journal of the Text Encoding Initiative

Issue 2 (February 2012)
Selected Papers from the 2010 TEI Conference

Marjorie Burghart and Malte Rehbein

The Present and Future of the TEI Community for Manuscript Encoding

Electronic reference
Marjorie Burghart and Malte Rehbein, « The Present and Future of the TEI Community for Manuscript Encoding », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 01 April 2012. URL : http://jtei.revues.org/372 ; DOI : 10.4000/jtei.372

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

Document available online on:
http://jtei.revues.org/372
Document automatically generated on 01 April 2012.

Marjorie Burghart and Malte Rehbein

The Present and Future of the TEI Community for Manuscript Encoding

1. Introduction

1 The purpose of this article is to highlight current trends that are prevalent in the community of scholars interested in manuscripts and their representation utilizing the TEI Guidelines. To achieve this goal, we invited scholars to take an online survey we designed, the results of which are reported and discussed here.[1]

2 Unsurprisingly, this study was initiated by members of the TEI Manuscript Special Interest Group (MS SIG), the goal of which is "to bring together users of the TEI who wish to improve the encoding strategies for marking up transcriptions and editions of manuscript materials," exploring "a range of issues common to editing manuscripts" (TEI Consortium 2011c). The MS SIG serves as a liaison between the community and the TEI Council for issues regarding the improvement of the Guidelines. Created in 2003, the MS SIG communicates using a mailing list[2] and holds regular meetings at the annual TEI conference and members' meeting.

3 However, reflecting the interests of the large and varied community of TEI users working with manuscript material is a sensitive task. In a recent article, the TEI was defined to mean three different things: it is an organization (the TEI Consortium); a research community (the users); and a set of concepts and tags (the Guidelines) (Jannidis 2009). The community of users reaches beyond not only the formal membership of the Consortium but also beyond the circle of subscribers to the mailing lists of the TEI and of the MS SIG.

4 For this reason, we have tried to identify the main themes that emerged from this survey among users of the TEI more broadly, not only within the MS SIG. In this article, we will outline the methodology we used to gather data and discuss what the results and suggest about the practices of users, the issues and limitations they have to deal with, their assessment of the tools, techniques, and Guidelines available to them, and their hopes for the future of manuscript and text encoding.

2. Methodology

5 The data was anonymously collected through an online survey conducted between 20 September and 17 October 2010.

6 The announcement of this survey was circulated on mailing lists, both within the TEI community and within the wider Digital Humanities community. The majority of these lists were international but several French lists were chosen as an example of a national, non English-speaking sub-community.[3]

7 The mailing lists within the TEI community were:

- TEI-L (international): the main list of the TEI community, for discussing all things TEI. About 820 members.[4]
- TEI-MS-SIG (international): list of the MS SIG, theoretically the core target of this survey. About 180 members.[5]
- TEI-FR (French): a French declination of TEI-L, with the same scope but French language discussions. About 160 members.[6]

8 The mailing lists with a more general Digital Humanities scope were:

- DM-L (international): list of the Digital Medievalist community of practice, among which many TEI users and/or manuscript editors are found. About 580 members.[7]
- DH (French): list about the Digital Humanities in France. About 250 members.[8]

9 The survey itself was created with an online survey application,[9] which provided a suitable framework, with the ability to download the raw collected data in CVS format. This raw data was then exported to a spreadsheet for analysis and to create graphs and charts.

3. Results

3.1 The Manuscript Community

10 The community of "manuscript encoders" can be regarded as the largest sub-community within the TEI[10] and as an important clientele for the future development and policies of the TEI in general and of the Guidelines, tools, and infrastructure in particular. In order to take into consideration the concerns and requirements of this sub-community, from which some general conclusions may be drawn, it is important to understand the profile of the community itself, i.e. its members and the work they are doing.

3.1.1. Outreach of the Survey within the Community

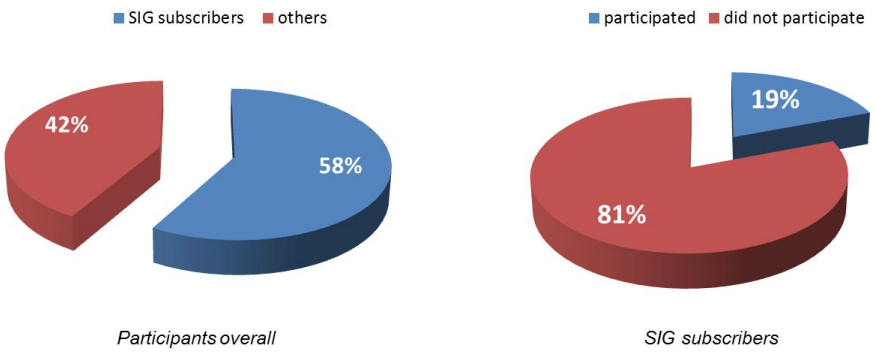

Figure 1: Response to the survey (60 participants overall) in regards to the subscribers of the MS SIG mailing list (total 180 subscribers).

11 The survey had an overall response of 60 participants, which we initially assessed as being a success considering that at the date of the survey the MS SIG mailing list had 180 subscribers. Since we felt that the list was the major forum for manuscript encoders and hence their subscribers as main clientele for the survey, a response rate of ⅓ of the list was impressive. However, closer examination of the data revealed that only 35 of the 180 subscribers to the list had actually participated, while the other 25 respondents (who are not members of the MS SIG list) were reached via other lists. This leads us to two immediate observations:

1. With only 19% of the SIG list subscribers responding to our call for participation in this (in our opinion) beneficial survey: is the MS SIG list (or the MS SIG in general) encouraging enough people to express their opinions?
2. Further, as 42% of the participants were not members of the MS SIG list,[11] how can we improve the awareness of the MS SIG (list)? Apparently a high number of scholars are using TEI for manuscript material but not presenting or discussing their results, suggestions and problems on this list.

12 These observations were discussed at the SIG meeting in Zadar the day after the presentation of the survey results, and steps were identified to improve marketing and public awareness of the SIG.[12]

13 One of the first questions asked about the educational level of the respondents: 63% of the have a PhD; 30% are post-graduates; and 7% state that they have no post-graduate degree. The positions currently held by respondents cover a wide range: research associate; research fellow or post-doctoral researcher (29%); full professor (19%); and assistant professor (8%). 25% of the participants could not find a suitable category in the system we devised. This also illustrates the diversity of academic positions across different countries.

14 For more on this and related topics, see Siemens et al. 2011. A good overview of academic ranks in different countries is provided by Wikipedia: http://en.wikipedia.org/wiki/List_of_academic_ranks.

15 Two out of three of the survey's participants were male. The age distribution ranges mainly between 25 and 54 years. Most of the participants consider themselves as either "Humanities scholars" and/or (multiple answers were allowed) "Digital Humanities professionals". Eight participants (13%) describe themselves as librarians, indicating that a large amount of work on manuscripts is undertaken in libraries or by librarians.[14] This might be explained by the role of libraries to preserve manuscripts and make them available via special collections.[15] Another seven participants (12%) were computer scientists — a number which illustrates the interdisciplinary nature of many encoding projects.

3.1.2. Geographic Distribution

16 A closer look at the geographic distribution of the participants revealed further useful information. Firstly, it was not a surprise that the vast majority of the respondents came from either Europe or North America.[16] Within Europe, France, the UK and Germany were the most represented countries, while no other European country had more than two participants. Again, this is a question of the outreach of the SIG and at the 2010 SIG meeting there was a lively discussion about how to involve more scholars from outside Europe and North America: for example, those from the Middle East or Asia. This is also a concern of the Digital Humanities community in general and regularly addressed, most recently in the call for papers for the DH2011 in Stanford, with its "Big Tent" theme, which particularly encouraged scholars from Latin America to participate.[17]

Figure 2: Geographic distribution of participants.[18]

17 But the unforeseen outcome of this part of the survey was the relatively high number of participants from France (12 respondents), especially in comparison to the other similar-sized Romance-language-speaking countries, Spain (one respondent) and Italy (two respondents). One reasonable explanation for this is that the survey was announced not only on the general MS SIG list, but on two French lists (TEI-FR and DH-FR).

18 We chose to compare the situation in France with that of Spain and Italy because unlike in Germanic-language-speaking countries, native speakers of Romance languages seem to be more reluctant to use English as a scientific lingua franca. The example of France shows that by addressing scholars in their native language, the outreach of the survey could be improved significantly—a fact that one should consider transferring to any issue of international community-building. We do not state here that non-English-speaking countries refuse to use TEI for encoding manuscripts but that there is more reluctance to share their experiences and results with the global community due to language constraints.[19]

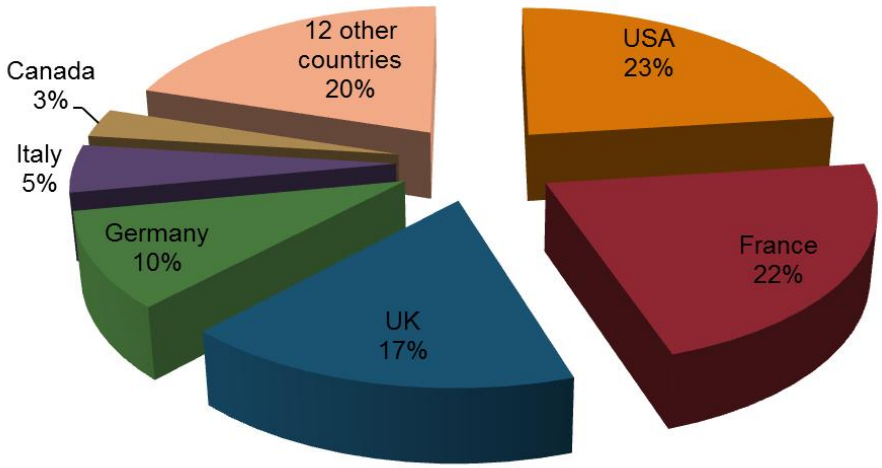

Figure 3: Place of residence.

19 The survey also revealed poor outreach to Central and Eastern Europe: only three participants in total responded from the Czech Republic, Slovenia and Russia. The same applies for countries outside Europe and North America. Organizing an internationally-active community with a global outreach is not easy. As one of the participants has pointed out, it is a drawback of the current activities of the MS SIG:

> It can be a little Eurocentric—if you can't make it to [face-to-face] meetings in Oxford or Paris it is easy to get shut out and there seems to me a little bit [is] going on behind the scenes among collaborators who know each other from other projects they are working on.[20]

3.1.3. Scope of Work

20 Answers to several questions give some insight into the scope of work of our community and into what sorts of source documents are encoded using the Guidelines.

21 According to the answers regarding the "historical era of your work", many respondents do not seem to specialize in one period alone as the total count of answers (89) to this question (in which multiple answers were allowed) exceeds the number of participants in the survey (60). Furthermore, the survey provides evidence that manuscript encoding is by no means the exclusive the domain of medievalists: the chronological distribution of the encoding projects before and after 1500AD is almost even (see fig. 4, left).

22 The same also applies (surprisingly to a larger extent) to the question about the "discipline (field) of your work". While manuscripts are typically thought of as source material for for historians and literature scholars, the fact that multiple answers were used (105 over 60) gives another indication as to the interdisciplinary of many of the projects (see fig. 4, right).

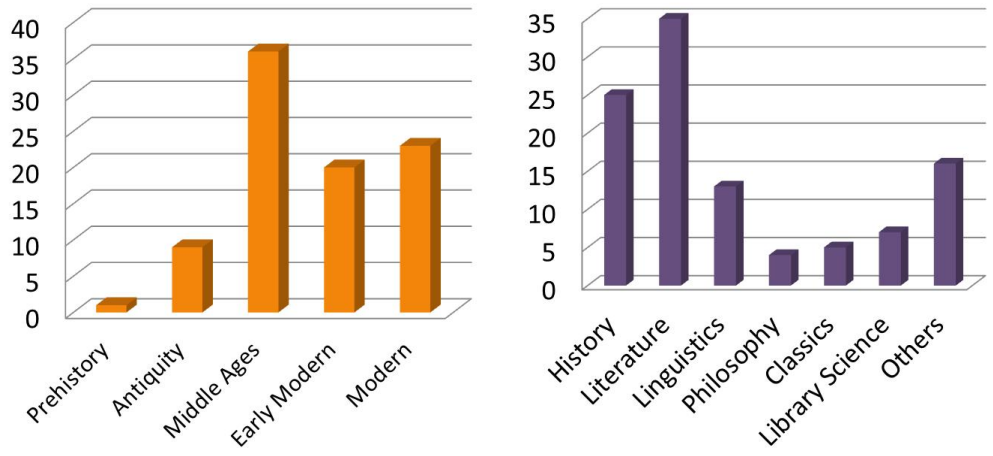

Figure 4: Historical era of work (left) and discipline/field of work (right). Figures are given in absolute numbers.

23 Given the high number of medieval texts from Europe, texts in Latin and other old European languages are most widely encoded by the community. In particular, French is represented by a good number while, for instance, Spanish or Italian are not. Although the pre-defined answers to "language(s) of the encoded material" could have been designed a bit more flexibly, one can conclude that result to this question (see fig. 5) also reflects the poor penetration of the survey into traditions with different writing systems (such as Chinese) or non-Latin characters (Arabic, Hebrew, Greek, Russian, etc.). On the other hand, one might question whether there is enough support in the TEI for writing systems that function differently (such as from right to left or using logograms)[21] or whether encoding issues or a lack of such support might be indeed a reason for the poor response in countries whose sources are mainly written in these languages.

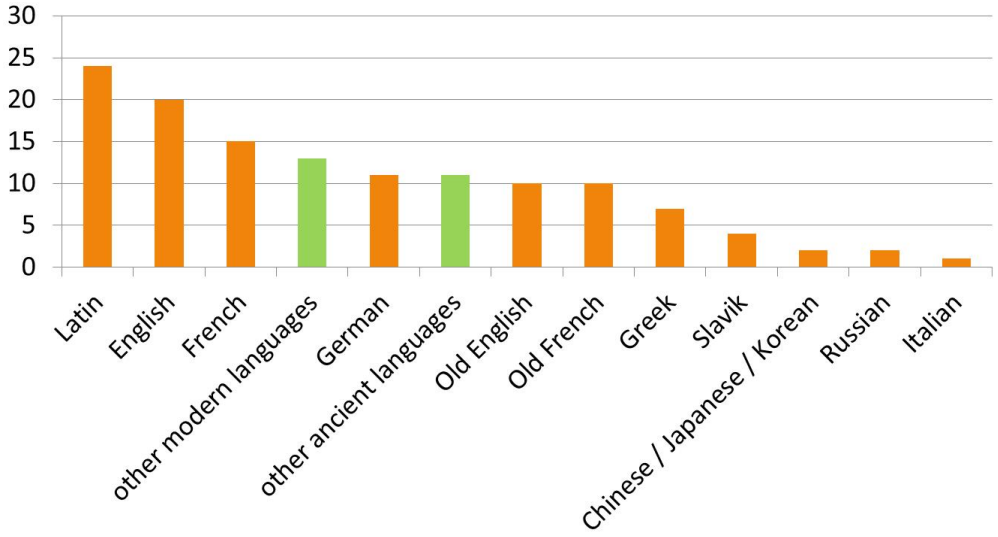

Figure 5: Languages of the encoded material. Figures are given in absolute numbers.

24 That is, the survey does *not* suggest that manuscripts are mainly (apart from Latin) written in (Old) English or French, but only that projects that are represented in this survey deal predominantly with manuscripts in these languages.

3.2. Learning and Using the TEI

25 The community of manuscript encoders is diverse. Certain common features are visible, but one can hardly define a typical scholar or object of study. In the second part of the survey we sought to better understand how people have got to grips with the TEI, what similarities exist in manuscript-related TEI projects, and how the TEI is utilized in these projects.

3.2.1. Learning

26 One set of questions forms the basis for examining when and how people learn the TEI, the main obstacles they encounter, and what supports them in their learning process.[22] Figures 6a and 6b summarize the answers to "for how long (rough estimate) have you been aware of the TEI?" and "for how long (rough estimate) have you been actually using the TEI for your work?". They show that there is, on the one hand, a good number of scholars relatively new to the TEI (with three years or less of awareness or practical experience) and, on the other hand, an equal number that have known about or used the TEI for quite a long time (more than ten years).

27 The results also suggest that it takes some time to go from knowing the TEI to actually using it (for example, 19 participants have known about the TEI for more than ten years, but only 11 have used it for that long). This is an interesting gap. An explanation for this learning curve might be the following explanation on obstacles to learning:

> I learned TEI in a specific […] environment, and when I learned it, I didn't quite understand the scope of the effort. But with time as I changed from a graduate student doing encoding to make money to a […] professional, I learned the full picture.[23]

28 But one should not necessarily draw the conclusion that the TEI has a steep learning curve just because of these figures. Equally plausible is the explanation that scholars are aware of the TEI and what it can do for their work even when they have no current use for it. They, however, return to it at a later stage when it is needed.

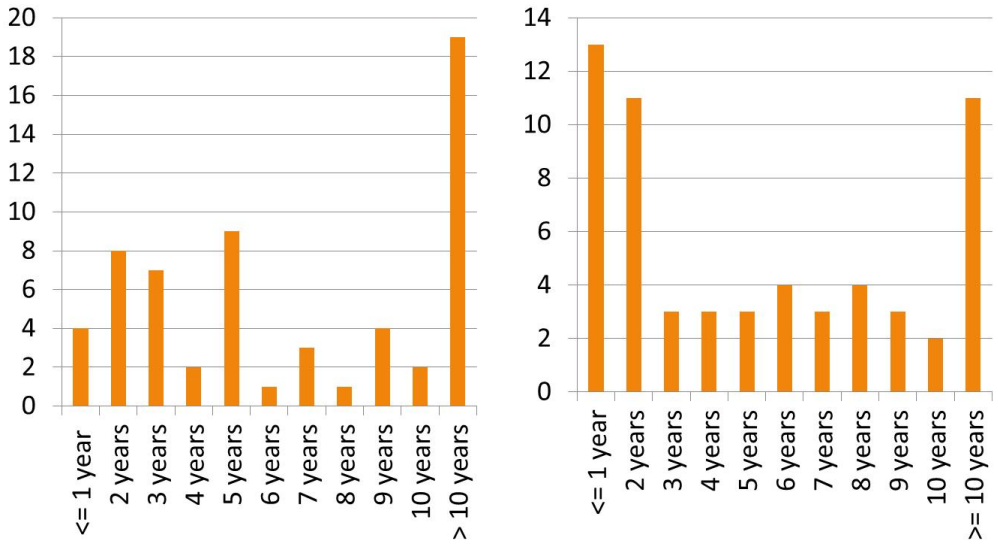

Figure 6a: Time of TEI awareness (left) and usage (right). Figures are given in absolute numbers.

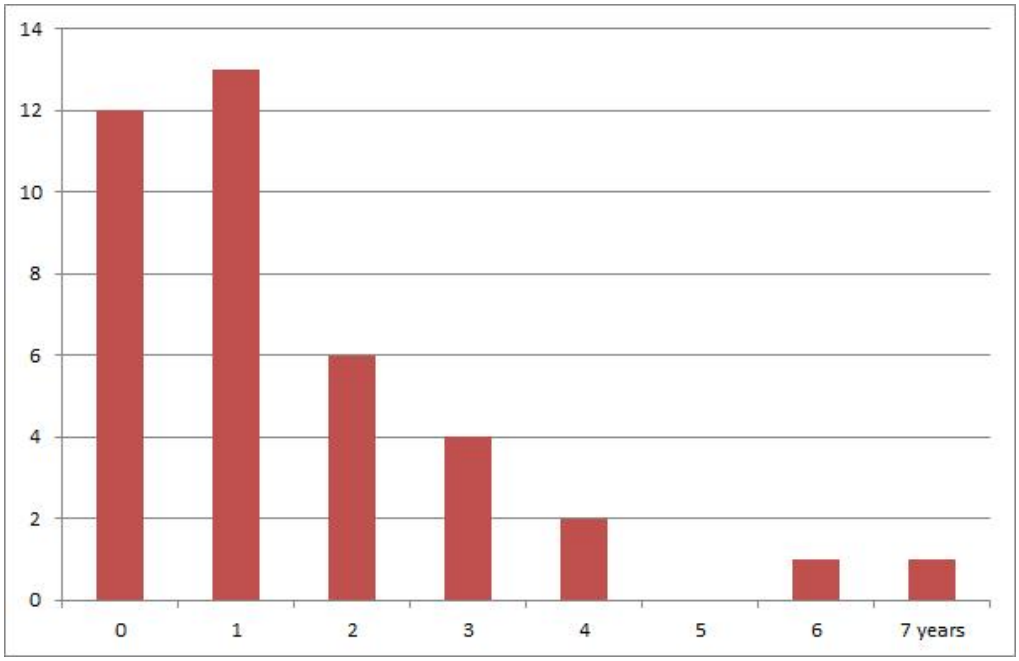

Figure 6b: length of the gap between TEI awareness and TEI usage. Figures are given in absolute numbers.[24]

29 A more interesting aspect of this part of the survey gives insight into how scholars learn the TEI. A vast majority (multiple answers were allowed) answered that they are self-taught or learned by doing. TEI courses are attended by less than one third of the respondents. However, course participation seems to have increased significantly (30% to 45%) within the last three years (see fig. 7).

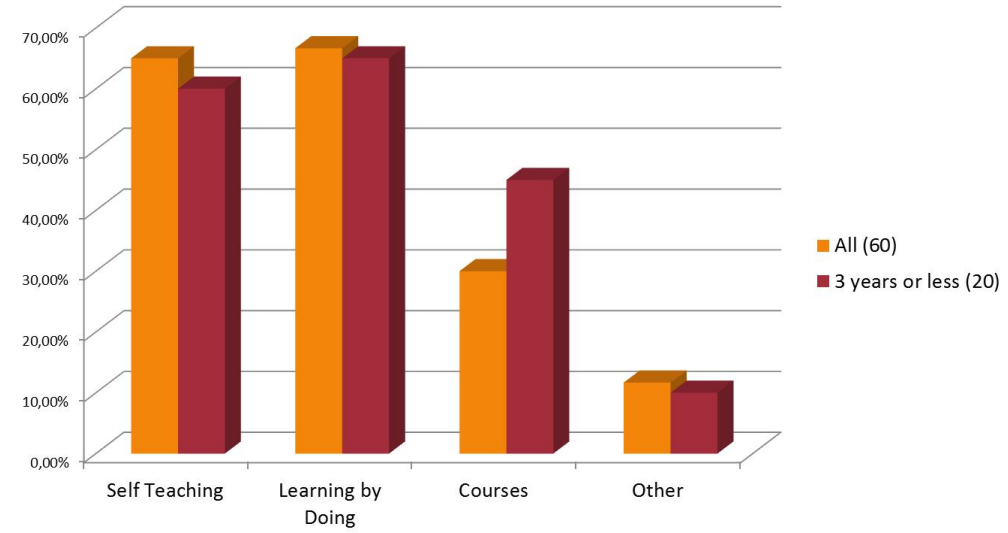

Figure 7: Learning the TEI. The orange bars on the left comprise all participants, the bars on the right only participants with three years or less experience with the TEI. Figures are given in percentages.

30 A further investigation into these figures gives insight into the types of learning methods (fig. 8). Although "learning by doing" and being self-taught might be difficult to separate from each other, 45% of the participants state that they rely on one method only, another 45% on a combination of two, and 10% on all three (self-taught, learning by doing, and course attendance). But the high percentage of learning by doing is striking and underpins the importance of not only practical experience while using the TEI for encoding manuscript material but also of practical approaches to teaching and learning.

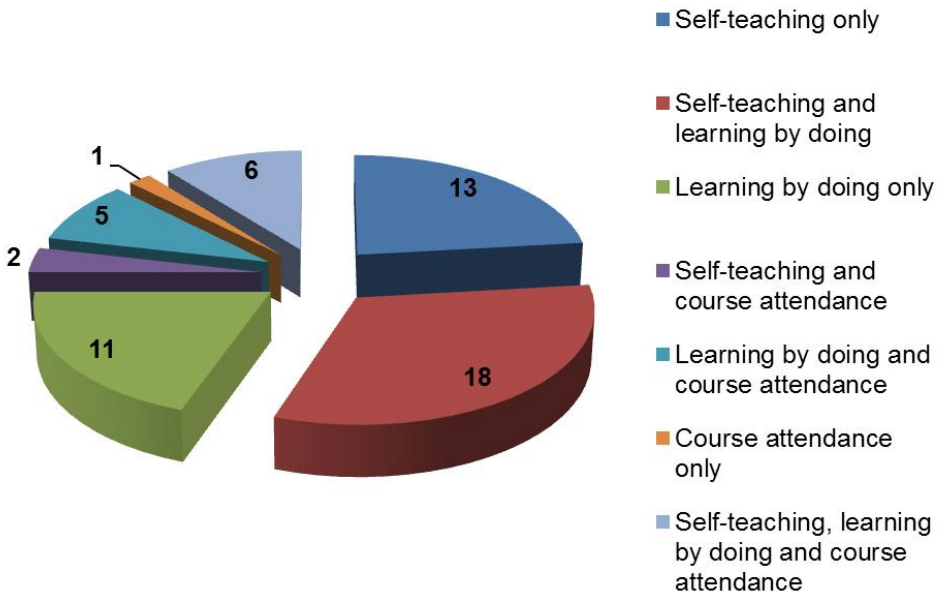

Figure 8: Combination of learning methods (in absolute numbers).

31 The survey also included a more qualitative assessment of learning practices. Two questions were asked of all participants regardless of how long they have used the TEI.

32 In the first question, we asked "when you decided to start using and/or learning the TEI, what were the main obstacles you had to overcome?". Respondents selected from a pre-populated list below (listed in order of ranking, giving the total numbers of answers in brackets):

- Lack of user-friendly tools (40);
- Problem understanding the guidelines (33);
- Lack of TEI training sessions in your district (28);
- Lack of support from your IT department (24);
- Reluctance of scholars to do encoding work (20);
- Problems finding TEI-competent collaborators for your project (16);
- Other (10).

33 When a respondent chose "other", we asked for more detail in free text form. Two problem areas were repeatedly addressed without the designers of the survey having biased the participants by pre-suggested answers. The first deals with difficulties in publishing TEI-encoded materials: both the lack of publication software and the complexity of publication, particularly in the application of XSLT. This was made explicit by one participant in a later question, when discussing his or her difficulties in TEI related work in general (discussed in section 3.4).

34 The second area touches upon the design of the TEI as such, the "breadth" of which is regarded as an obstacle to learning (and consequently using) it, as well as the "overwhelming options to choose from".

35 Apart from this, the lack of (user-friendly) tools is noticeable, as two thirds of the participants regard this as an issue, especially in the learning phase.[25]

36 The second qualitative question in this part of the survey tackled learning from the opposite angle: "when you decided to start using and/or learning the TEI, what were the most helpful elements?". Respondents again chose from a pre-defined list, given here according to rank:

- Guidelines (56)
- Consultancy / advice from TEI expert(s) (39)
- Questions to and answers from the TEI community (listserv etc.) (33)
- Possibility to attend TEI training sessions (22)[26]
- Support from your IT-department (8)
- Other (7)

37 It is noticeable that the Guidelines themselves are apparently most relevant as a reference document. Community support (experts and the mailing list) should also be mentioned, although only 55% of participants seem to refer regularly to TEI mailing lists for (see below for more on the use of the MS-SIG-specific mailing list).

38 Again, when stating "other", we asked for more detail. Worth mentioning here is that two respondents found examples particularly useful.[27] Unfortunately, we did not ask the question "what would you like to have as helpful elements in the learning process?" to get a more representative view on the usefulness of examples in learning and using TEI.[28] One participant makes this point explicit when answering a different question:

> Clear examples of the TEI in post-transcription/encoding action. Easier access to xquery tools or something to demonstrate how to put one's new corpus into good use.[29]

39 For both questions, we generally state that the answers by "Rookies" (less than three years of experience) do not differ significantly from answers by "Veterans" (more than ten years of experience).

3.2.2. TEI Projects in Manuscript Studies

40 In this section of the survey we wanted to learn about typical manuscript encoding projects. Figure 9 shows the size of projects in terms of project members. The question asked was: "apart from you, how many people contribute to your TEI-related work?". Projects or workgroups tend to be small- to medium-sized: the average (mean) is five, the mode (the value that occurs most frequently) is either one additional project member or it lies between three and nine. Typical roles in a project that has more than one member cover the whole process of creating digital resources, including project coordinator/leader, consultant, architect, encoder, and programmer.

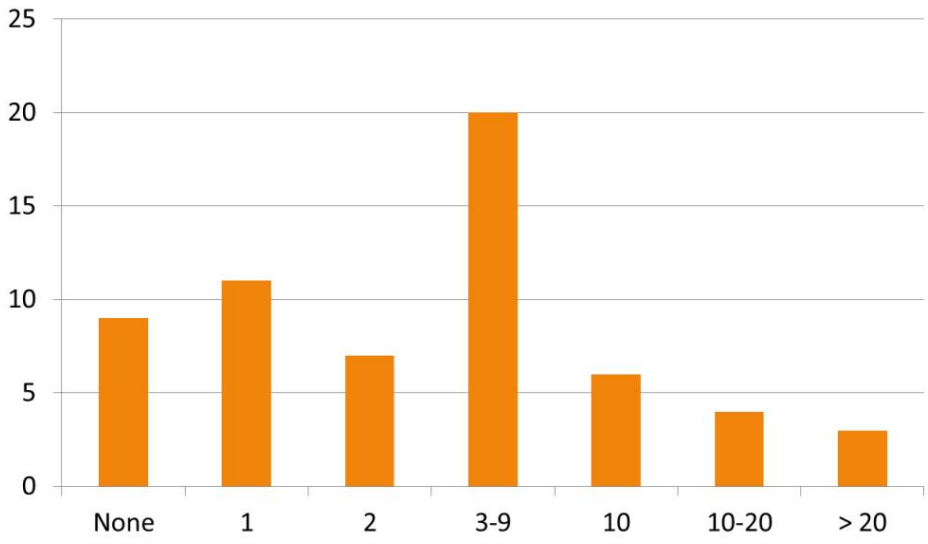

Figure 9: Number of project members in addition to the respondent. Figures are given in absolute numbers.

41 The medium for publication of encoded manuscript material is overwhelmingly digital: 97% of the respondents use the Internet as a publication medium.[30] Non-digital publication still plays a role, but a minor one: five respondents publish non-digitally. However, only one respondent will publish exclusively non-digitally; all others marked Internet and/or other digital publication).

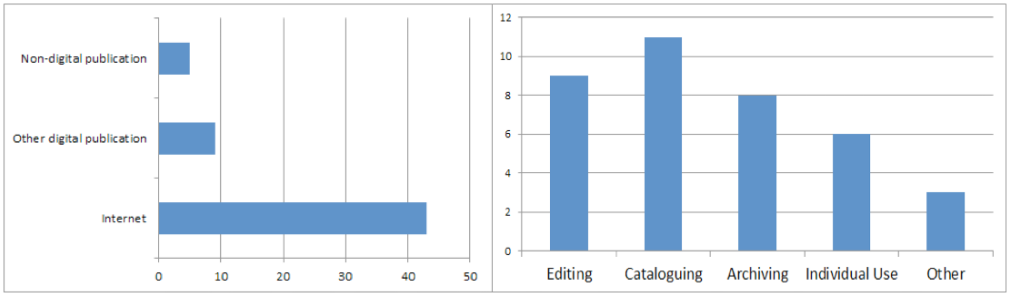

Figure 10: Publication medium (left) and final purpose of encoding (right). Figures are given in absolute numbers.

42 The final purpose of encoding manuscript material is more or less evenly distributed among editing, cataloging, archiving and individual use (i.e. not for publication); other usages were marked only by three participants and can be summarized as "individual use".[31]

3.2.3. Using the TEI

43 This section covers aspects of using the TEI in terms of designing customized schema as well as the software or tools for the encoding itself.

44 Figure 12 shows the take up of the various modules of the TEI[32] in manuscript encoding projects. The two top-ranked modules correspond to the chapters of the Guidelines that we anticipated to be the most important for manuscript encoding: msdescription (chapter 10: "Manuscript Description") and transcr (chapter 11: "Representation of Primary Sources"). The clear edge that msdescription has (85% overall) indicates that most of the projects include a manuscript description of some sort but not all (71%) provide a transcription of the text, maybe because they are pure cataloging projects[33] or because the texts do not have special features that need encoding beyond those available in the core module. Reflecting the fact that manuscripts are frequently encoded for the purposes of creating a scholarly edition, textcrit (chapter 12: "Critical Apparatus") also plays an important role (55%), but it is a module with a high demand for improvement (see below section 3.5). From the non-manuscript specific modules, namesdates (chapter 13: "Names, Dates, People, and Places") is ranked highest (64%), indicating that the TEI Consortium had reacted positively to the requirements of the community by expanding this chapter inversion P5 of the Guidelines.[34]

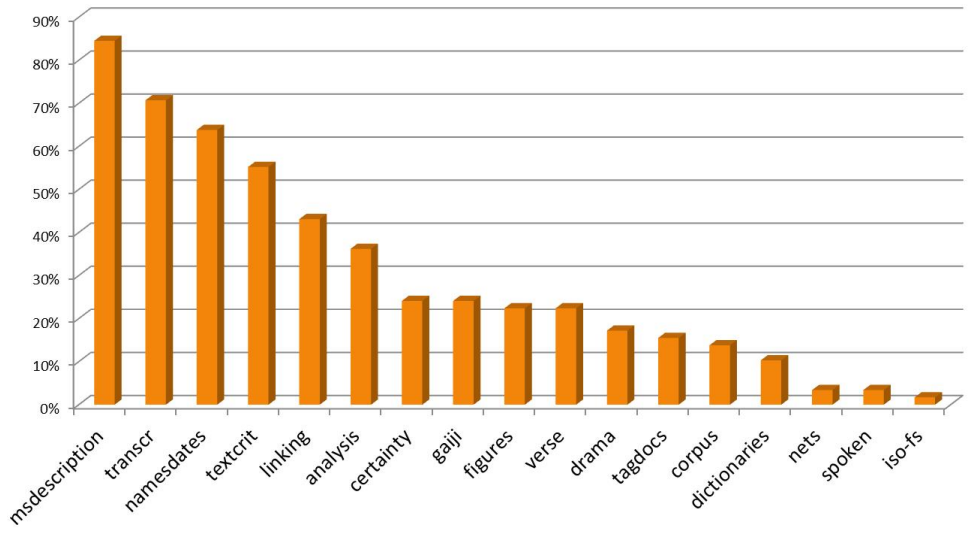

Figure 11: TEI-Modules used in manuscript encoding projects (apart from the four basic ones: core, tei, header, textstructure).

45 The TEI provides two options for schemas: the first is utilizing a pre-defined schema that can be downloaded from the TEI site; the second is to customize a schema with or without the support of a tool (TEI Consortium 2011a). A slight majority of 58% answered "yes" to "do you customize the TEI with ODD files for your projects?". This might be interpreted to mean that a large number of people are familiar with customization, but this number may be deceptive and may be due to only one person in a project team being responsible for schema creation

and design. However, most of the remaining respondents (42%) give a more detailed answer for what they use instead of customizing the schema with ODD: a vast majority state that they employ tei_all. TEI Lite does not seem to be an option for manuscript encoders despite the TEI Consortium's claim that it was "the most widely used TEI customization" (2011a) "designed to meet '90% of the needs of 90% of the TEI user community'" (TEI Consortium 2011b). It was discussed in the SIG meeting following the presentation of the survey that there is a strong need for something like a "TEI Lite for Manuscripts" that covers these 90% for manuscript-related projects in order to facilitate an easy start with the TEI. As one participant commented: "a 'digest' version of the guidelines, designed for specific uses, could help."[35] As no customization for manuscript encoding purposes is currently available, a task force for designing a manuscript-specific ODD was established (TEI MS SIG 2010).

46 A vast majority of the respondents support their encoding with specialized XML-software (90%); 27% use general text editors, and 17% use word-processing software (fig. 12). Almost a quarter of participants (23.3%) use "ad hoc software developed for your project"—encoding tools that are developed for a particular purpose with a project context, but only one respondent relies exclusively on tailor-made software for encoding. This illustrates an important need for generalizable software, the lack of which often leads projects to develop their own tools which for aspects of their workflow.

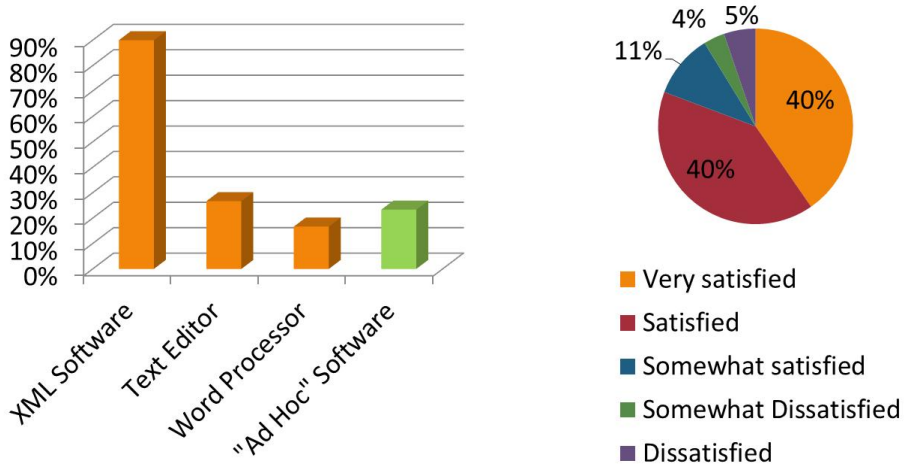

Figure 12: Software used for TEI encoding (left, multiple answers allowed) and grade of satisfaction with this software (right).

47 It is interesting to note that from the 9% lowest-ranked answers to a question about satisfaction with this software ("dissatisfied" or "somewhat dissatisfied"), only two indicated that they use specialized "XML software", four use text editors, and one uses word-processing software. On the other hand, in terms of the highest ranked grouping, XML Software, there is a clear preference for one product: 91% use oXygen.[36]

3.3. Involvement in the Community

48 Another part of the survey was aimed at assessing how (and *if*) the participants were involved in the TEI community at large, and more specifically in the TEI MS SIG.

3.3.1. The TEI Community at Large

49 Unfortunately, we did not include a question about participants' membership on the TEI-L discussion list, which we assumed was a given: but in light the unexpectedly low rate of membership on the MS SIG list, enquiring about membership on TEI-L list might have provided useful feedback.

50 Nevertheless, there were questions aimed at evaluating the level of participant interaction within the TEI community. Firstly, answers to the question "Have you attended annual TEI meetings?" revealed that 58.3% had never attended; 23.3% had attended once; and 18.3% had attended more than once. Maybe this can be explained by the large proportion of people relatively new to TEI[37] who have not yet found an oportunity or felt the need to attend a TEI conference.

51 Besides taking part in the discussions on the mailing list and attending annual TEI conferences, the participation in special interest groups is one of the main ways of getting involved in the community. Figure 13 shows the results of anwsers to the questions "Do you participate in other TEI Special Interest Groups? [apart from the TEI MS one]", and "If so, which one(s)?". Only 16 participants (26.7%) answered positively to the first question. It is interesting to note that among them two were not members of the TEI MS SIG, which means that 23 participants (38.3%) were not members of any TEI SIG[38]. Among the rest, 28 out of 37 (75.6%) were members of one or two SIGs at most, and only three people were subscribers of more than four SIGs. The "Scholarly publishing" SIG comes first with nine members, followed by the "Tools" and "Libraries" SIGs, tied for second place, with seven members each. It is worth noting out of the seven participants who were members of the Libraries SIG, only four defined their own background or job as "Librarian", demonstrating that participation in the SIGs breaks traditional professional boundaries.

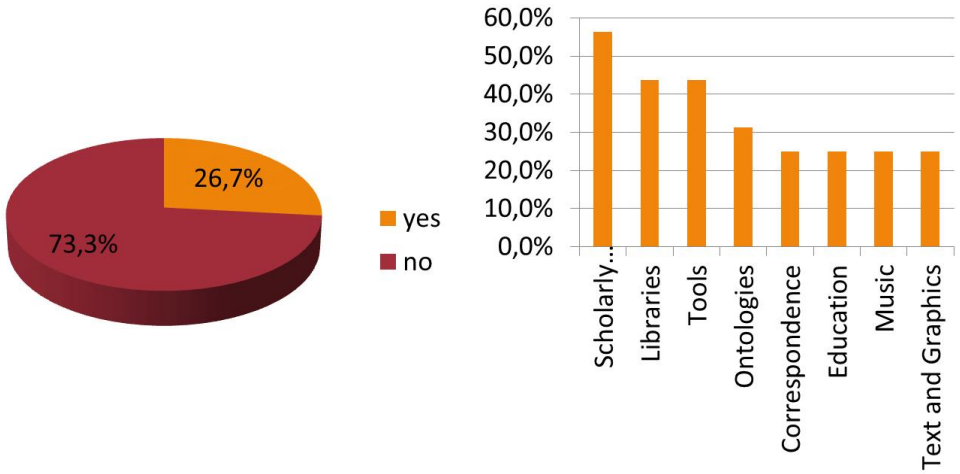

Figure 13: "Do you participate in other TEI Special Interest Groups?" (left) and if yes (16 answers): "If so, which one(s)" (right).

3.3.2. The TEI MS SIG

52 As has already been mentioned, the proportion of members of the TEI MS SIG list who participated in the survey was unexpectedly low (58%).[39] Nevertheless, a set of questions had been designed to understand the profile of those involved in the SIG. The first questions of this section dealt with the assessment by the MS SIG list members of the usefulness of the discussions posted: its results are rather encouraging since the vast majority (84%) rate the posts as occasionally or often useful to them while only a tiny minority (3%) deem them to never be useful (see fig. 13, left). But the enthusiasm needs to be put into perspective considering the responses to the next question: the participants were asked whether they post on the list, and this time what was revealed was a "silent majority" who follow discussions but do not post (54.1%), while only a minority of members (2.7%) define themselves as regular contributors (see fig. 13, right). These figures are consistent with the statistics of participation of the MS SIG list.[40]

53 It is interesting to examine the reasons for respondents' silence. Thirteen out of 20 the "silent members" provided an answer.[41] One member gave a lack of time as the reason;[42] two gave the reason as shyness or the feeling that they have not mastered the TEI well enough to participate in the discussion;[43] two others state that they are recent members who have not yet had the time to get involved,[44] and interestingly, four imply that they do not feel the need to post to the list.[45]

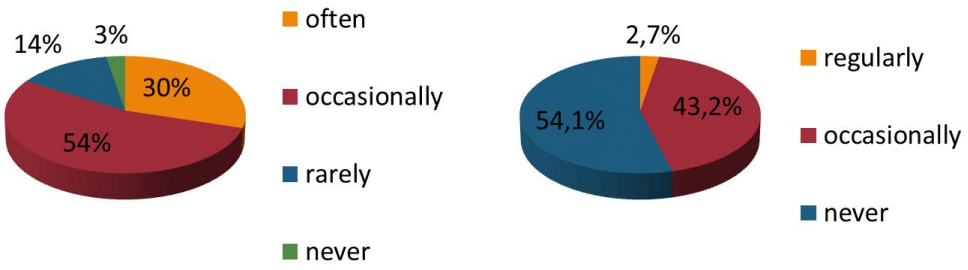

Figure 14: "Are the posts [of the MS-SIG-list] of interest for you/your work?" (left), "Do you post to the list yourself?" (right).

54 But participation on the discussion list is not the only yardstick by which the value of the MS SIG can be tested. The set of questions that followed was aimed at assessing the level of satisfaction of the participants in regard to the various activities of the SIG. They were asked to rate their satisfaction with six activities, from 1 (very dissatisfied) to 5 (very satisfied). Figure 15 provides a synthetic view of the results. The answers show that the participants appreciate the work of the SIG since the average mark of each activity ranged from 3.25 to 3.6. The three better-rated items were "Help with problems", "Improvement of the Guidelines", and "Representation of MS. related issues towards the TEI-C", showing that the role of the SIG as a means of support for problems and liaison with the TEI Consortium for manuscript-related issues is well received. On the other hand, the three items that received the lower rating were "Communication and information", "Community building" and "Involvement of users"; this is coherent with the relative lack of activity on the discussion list. It is certainly these areas of community-building and involvement of users that the MS SIG should develop.

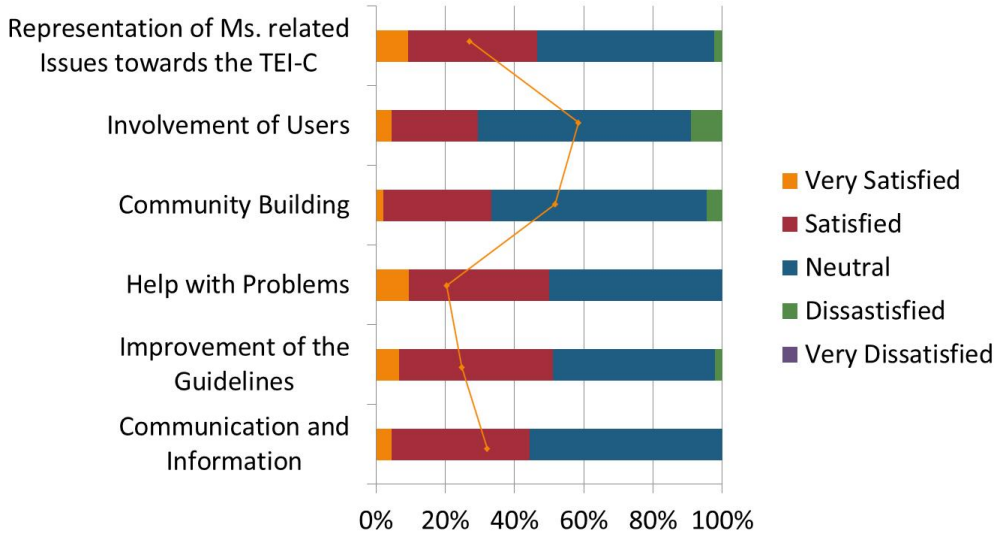

Figure 15: Grade of satisfaction with the activities of the MS SIG. The orange line in the center indicates the average (very satisfied = 5.0; very dissatisfied = 1.0). Ranges between 3.25 (Involvement of Users) and 3.60 (Help with Problems)

55 Fortunately, participants seem to be ready to become more engaged in the activities of the MS SIG: when asked "On what would you be most likely to contribute to this SIG?", 20 provided an answer.[46] A majority of respondents indicated that they would like to participate in terms of giving advice, sharing practices and animating the discussions (seven answers), echoing the will to develop the involvement of users and also the role of the SIG as a place where one can submit problems and get help,[47] but the rest were more technically-oriented. Among the possible tasks proposed by participants, some are being gradually addressed by the SIG: this

is the case, most notably, in the improvement of manuscript description for non-Latin scripts[48] and improvement in the Critical Apparatus module.[49] During the 2010 meeting of the SIG, both these issues were addressed and defined as a prominent part of the work plan for 2011.[50]

3.4. Issues

56 Participants were asked about the main issues they are faced with while working with the TEI in terms of workflow.

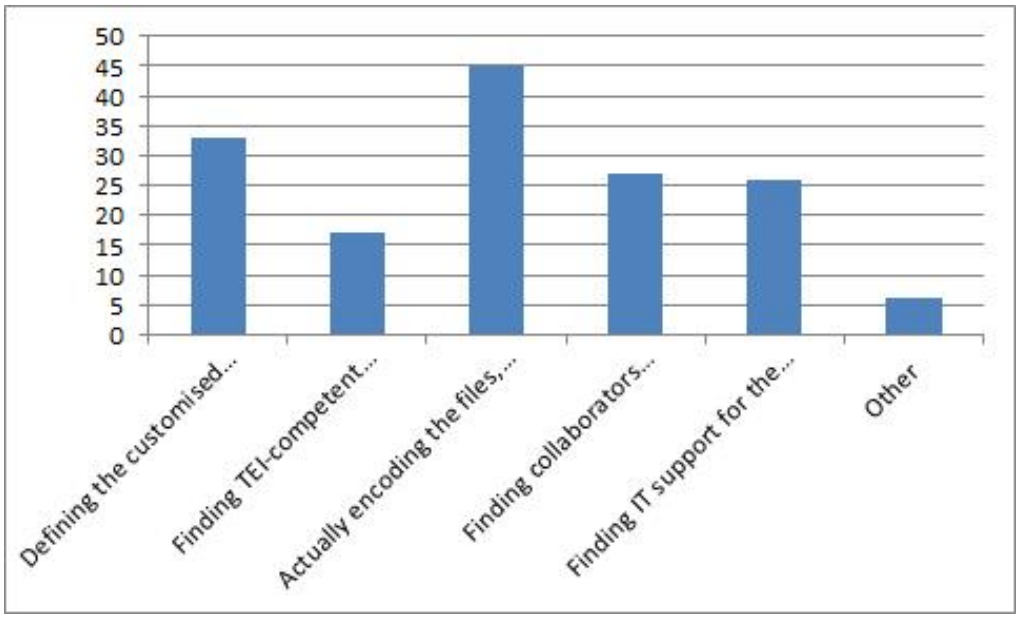

Figure 16: What would you define as the most difficult part of the TEI work in your project(s)? Figures are given in absolute numbers.

57 Answering the question "What would you define as the most difficult part of the TEI work in your own project(s)?" from a pre-populated list, respondents rated "Actually encoding the files, and if relevant making the practice of various encoders consistent" as the major impediment (45 responses). This may be linked to the lack of user-friendly tools. This is the main obstacle participants had to overcome when they started learning or using the TEI (40 responses). Participants were invited to expand with comments.[51] Ten comments were related to the encoding.[52] Beyond the usual throes of team-work,[53] some indicated limitations in the TEI, for instance:

> Different punctuation marks (developed in the Middle Ages) are difficult to encode along with text that is meaningfully displaced. And although TEI was not designed with images in mind, encoding images is, well, an adventure.[54]

58 Others emphasize the overwhelming scope of the TEI and the difficulty in making consistent encoding choices[55]—one even mentioned, not without humor—a situation that will probably sound familiar to most people who have experienced "TEI evangelization":

> [A]ny usage of TEI immediately leads to controversial fundamental and philosophical discussions on the sense and meaning of single elements/attributes as well as the question whether using the TEI makes any sense at all[56]

59 Fortunately, another answer gave a more optimistic insight into why the TEI has been adopted by so many scholars:

> Scholars contributing to this project surprisingly do like encoding and especially do like choosing their own way of using the tag library![57]

60 There is a noteworthy gap between the first and the second answer ("Defining the customized schema you are going to use (ODD file, etc.)"), which was mentioned by 33 participants, in comparison to 45 for the first. About this, it is interesting to compare the answers with those of the question "Do you customize the TEI with ODD files, for your projects?",[58] where 35 participants (almost the same number) declared that they did use ODD for TEI customization:

out of those 35 participants, only 20 rated ODD customization as an issue, meaning that 13 out of the 33 respondents who consider it a major difficulty do not use it themselves. This may be interpreted as an acknowledgement of the complexity of the TEI customization process, but on the other hand those figures may illustrate a knowledge gap, and in some way an apprehension, since people who are not involved in the customization process see it as a very complex and difficult part of a project.

61 The third and fourth choices ("Finding collaborators competent in technologies related to XML/TEI" and "Finding support for the installation/implementation/hosting of the specific applications needed to publish and use your TEI files on the Internet") followed in close succession (mentioned in respectively 27 and 26 answers). These two issues are both related to forms of technical support not directly linked to the TEI in itself, but to the various technologies and services needed to process and publish it. This demonstrates that TEI projects need to be considered in a holistic environment without restraining the focus to issues of modeling. When respondents provided further detail as to their choices, seven were related to these issues.[59] One of those respondents gives a particularly interesting explanation, insisting on the "great gap" between the modeling step of the TEI encoding and the actual possibility of publishing and processing the encoded files with performant tools:

62 The wors[t] thin[g] with TEI, in my opinion, is the GREAT gap between the text-encoding in XML and the final presentation of your results [o]n the internet. It is relatively easy to learn the basics of TEI-encoding, in order that a humanist is able to prepare his/her edition, BUT he/she needs a very specialised IT man to make his/her TEI document to be an actual on-line edition. This is a great obstacle to TEI; I have the impression that the TEI-C doesn't pay enough attention to this worst drawback.[60]

3.5. The Future

63 Addressing the issues raised in the preceding section is not always within the power of the TEI Consortium (this is most notably the case of the lack of adequate local IT support). But improving the Guidelines is one of its missions, Therefore a set of questions was designed to get feedback on the Guidelines and elicit possible improvements to them.

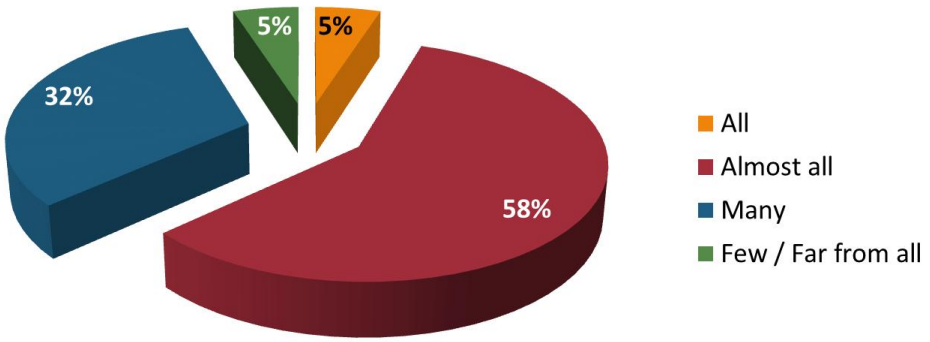

Figure 17: Do the TEI Guidelines fulfill the requirements of your projects?

64 Asked if the TEI Guidelines fulfill the requirements of their projects, 63% answered "All"[61] or "Almost all", showing an overall majority is satisfied with the Guidelines. However, 37% indicated a lower rate of satisfaction ("Many", "Few" or "Far from all"): for these respondents, it seems that the Guidelines would need small to large-scale improvements, such as the ones recently proposed by the MS SIG workgroup on genetic editing (Manuscripts SIG 2011; Workgroup on Genetic Editions 2011).

65 Participants were invited to give more details about the areas in which they wished the Guidelines were improved, with multiple choices corresponding to Guidelines chapters.

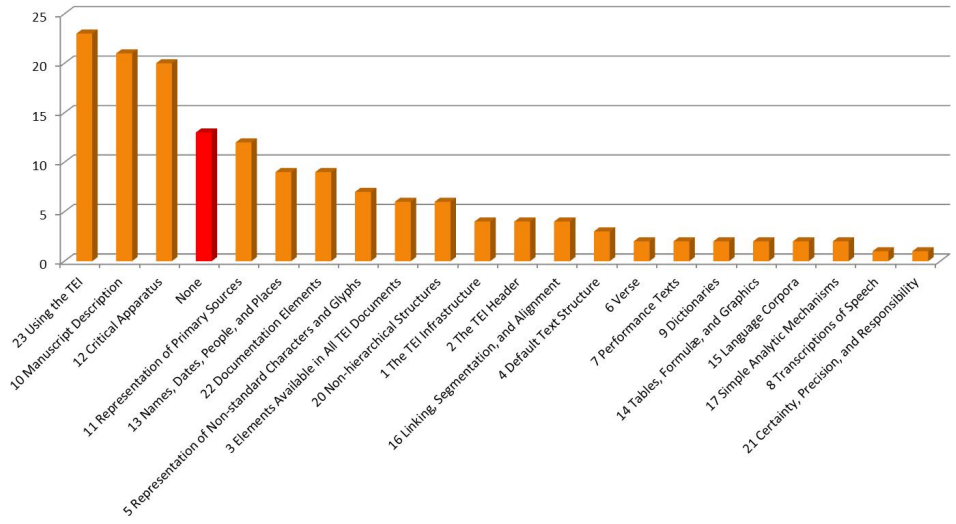

Figure 18: In what areas do you wish the Guidelines to be improved? . Figures are given in absolute numbers.

66 Interestingly, while five respondents declared that the Guidelines fulfilled "All" the requirements of their projects, significantly more (13 respondents) answered "none" to the question "In what areas do you wish the Guidelines to be improved?". Upon closer examination, four of the 13 had answered "All" to the question about the Guidelines fulfilling their requirements, eight "Almost all", and one only "Many". That is, among the five respondents who were fully satisfied with the Guidelines, only four answered "None". Presumably this discrepancy means that some participants acknowledge that some of the aspects of their encoding projects are outside the scope of the TEI Guidelines and therefore suggest no areas of improvement despite the fact that not all their needs are addressed; other users seem to take the opposite tact and suggest areas of improvement even though their own needs are fully addressed.[62]

67 Regarding individual chapters needing improvement, the chapter "Using the TEI" was ranked first by 23 respondents. This chapter is rather technical, dealing with the TEI schemas, their personalization and customization, ODD, and questions of conformance. Even though the Guidelines discussion of these advanced issues might very well be deemed to need some improvement, it is possible that this chapter was chosen because of its title rather than its content. This hypothesis is supported by the fact that when invited to comment further 10 out of 23[63] suggested improvements such as "user-friendly guidelines"[64] Another answer points to the difficulty in that too many options are offered by the Guidelines. The respondent called for stronger, more definitive recommendations about the way things should be encoded:

> Too many choices. The TEI is reluctant to make a decision (even if arbitrary) about best practice. As a result, alternative approaches proliferate because of personal preference (rather than technological necessity), which fractures the potential community and frustrates easy interchange.[65]

68 The chapters "Manuscript Description" and "Critical Apparatus" come second and third for improvement, with respectively 21 and 20 answers. That these chapters are so highly ranked as needing improvement does not come as a surprise since they are the most likely to be directly of use to people interested in the work of the MS SIG,. Understandably, the respondents have more suggestions for improvement for the parts of the Guidelines they use most and therefore know better. Consequently, the suggestions for improvement expressed by the participants in the next question were more precise, and informed by technical experience.

69 Seven answers suggest improvements to the Manuscript Description chapter,[66] indicating mainly on the need to broaden its scope, from a strict definition of the manuscript as a codex or part of codex to a more generic "object", as well as the need for more efficient ways to describe manuscripts. However, one respondent calls for a more elaborate system:

> MSDesc is still a bit too simple for the very complex cases, e.g. composite MSS with components scattered across libraries plus fragments, and so on. I understand work is being done on this, though.[67]

70 Another calls for more permissive manuscript descriptions:

> Manuscript Description tags are very specific. It would be useful to have some way to identify the variable information that one is not tagging systematically but acquires for different manuscripts. […][68]

71 Regarding the "Critical Apparatus" chapter and more generally the encoding of critical editions, seven respondents offered suggestions for improvement.[69] They address general issues (the difficulty in working with multiple witnesses in parallel segmentation,[70] the need for a better representation of variance,[71] or issues raised when one wants to encode both a critical and diplomatic edition of a document[72]), but also point out very precise issues, calling for a better way to handle uncertainty[73] or text transposition in different witnesses.[74]

4. Conclusions

72 The community of scholars interested in the encoding of manuscript material is broad and diverse. It reaches far beyond the institutional limits of the TEI Consortium or MS SIG members and even beyond the group of people who actively discuss all things TEI.

73 In this article, through the results of our survey, we have tried to draw a picture of this community and to understand its approach to the TEI, its way of implementing it in projects, and its wishes for the future. With sixty respondents, one can question how representative this survey is. In our view, even though the number of respondents clearly does not do justice to the number of people using the TEI for manuscript material the world over, one must keep in mind that there are considerably fewer people active in the community ready to voice their needs and feelings. Given the scope of the call for participation (covering the main mailing lists of the community) we consider that the results give a picture of the community that, if not comprehensive, is reasonably accurate.

74 The results demonstrate the existence of a steep learning curve (where self-teaching and learning-by-doing dominate) characterized by the long gap between the time when people become aware of the TEI, and the time they start to use it. Users have to overcome obstacles, such as the lack of user-friendly tools, the difficulty of coming to terms with the Guidelines, and the lack of advice from TEI experts.

75 But in digital projects, TEI encoding is only one of many technical aspects that must be considered. The TEI is neither a starting point nor an end in itself. It has to be considered within a holistic environment which goes beyond mere encoding but also concerns the use of the encoded data. In this respect, there is an important need for user-friendly, bespoke tools facilitating the processing, analysis and publishing of the material. Embedding TEI-encoded texts into a larger workflow is necessary to enhance to facilitate wider adoption of the TEI in digital editions This encompasses tools, of course, but also education and support with examples, best-practice guidelines, and how-to's.

76 The survey illustrates the existence of a strong, focused sub-community within the TEI, centered on one use (encoding manuscript material), with few people participating in more than one SIG. Such sub-communities have particular needs, and the assessment of the Guidelines by users in this community shows that these needs are not yet totally met. There is room for improvement, on a small scale as well as on a large scale, and the special interest groups appear to be the appropriate mediator between the TEI as a research community, the Guidelines, and the TEI as an organization. However, it appears that an significant proportion of TEI users do not take the first step of joining a SIG even though they would most probably benefit from it. This is partly due to the fact that SIGs are sometimes thought of as expert groups rather than interest groups, keeping less advanced users away. This could easily be addressed by better communication on the scope and role of the SIGs on the TEI website as well as on the main mailing-list, the TEI-L. We feel that the TEI would benefit from a strengthening of the role of the SIGs and their membership base, putting these Groups at the heart of the interactions between the formal organization, the Guidelines and the community of users.

Bibliography

Fischer, Franz, Christiane Fritze, Malte Rehbein, and Patrick Sahle, eds. Forthcoming. “Digitale Edition und Forschungsbibliothek”, *Bibliothek und Wissenschaft*.

Hirsch, Brett, ed., Forthcoming. *Digital Humanities Pedagogy: Principles, Practices, and Politics.* University of Michigan Press.

Jannidis, Fotis. 2009. “TEI in a Crystal Ball.” *Literary and Linguistic Computing* 24 (3): 253–265. doi:10.1093/llc/fqp015.

Manuscripts SIG. 2011. “Documents and Genetic Criticism TEI Style.” http://www.tei-c.org/SIG/Manuscripts/genetic.html.

Siemens, Lynne, Elisabeth Burr, Richard Cunningham, Wendy Duff, Dominic Forest, and Claire Warwick. 2011. “A Trip Around the World: Balancing Geographical Diversity in Academic Research Teams.” Abstract of paper presented at Digital Humanities 2011. http://dh2011abstracts.stanford.edu/xtf/view?docId=tei/ab-104.xml;query=;brand=default.

TEI Consortium. 2011a. “TEI: Customization”. http://www.tei-c.org/Guidelines/Customization/.

TEI Consortium. 2011b. “TEI Lite.” http://www.tei-c.org/Guidelines/Customization/Lite/.

TEI Consortium. 2011c. “TEI Manuscripts Special Interest Group.” http://www.tei-c.org/Activities/SIG/Manuscript/.

TEI MS SIG. 2010. “SIGMS Minutes 20101112.” http://wiki.tei-c.org/index.php/SIGMS_Minutes_20101112.

Terras, Melissa, Ron Van Den Branden, and Edward Vanhoutte. 2010. “Teaching TEI: The Need for TEI by Example.” *Literary and Linguistic Computing* 24(3): 297–306. doi: 10.1093/llc/fqp018.

Wittern, Christian, Arianna Ciula, and Conal Tuohy. 2009. “The making of TEI P5.” Literary and Linguistic Computing 24(3): 281–296. doi: 10.1093/llc/fqp017.

Workgroup on Genetic Editions. 2011. “An Encoding Model for Genetic Editions.” http://www.tei-c.org/Activities/Council/Working/tcw19.html.

Notes

1 We would like to thank James Connolly for his careful proofreading of this article.

2 See http://wiki.tei-c.org/index.php/SIG:MSS

3 It must be stressed that, even though the announcement was circulated with a French introduction, the survey itself was not translated.

4 There were 818 subscribers to TEI-L on 1 November 2010 according to http://listserv.brown.edu/archives/cgi-bin/wa?INDEX.

5 We thank Elena Pierazzo for providing this figure.

6 This figure is publicly available on the list’s homepage.

7 We thank Daniel O’Donnell for providing this figure.

8 This figure is publicly available on the list’s homepage.

9 SurveyGizmo: http://www.surveygizmo.com/.

10 The number of subscribers to other SIGs is the following: Ontologies: 94; Correspondence: 67; Text and Graphics: 50; Manuscripts: 182; Music: 84; Tools: 41; Libraries: 125; Education: unknown. Data collected on 1 and 2 November 2010.

11 This figure was quite a surprise (and was indeed not anticipated) so that we did not include in the survey the question “If applicable: why are you not a member of the SIG list?” which would have given us better data for improving awareness of the MS SIG.

12 See the minutes of the SIG meeting: http://wiki.tei-c.org/index.php/SIGMS_Minutes_20101112.

14 “Archivist” was, however, only named by two participants. This is, from our point of view, a surprise as archives are even more so a holder of manuscripts than libraries. The fact that librarians appear to be more active in the TEI community might be coincidental but is certainly worth a closer look.

15 On the role of libraries in editorial projects, see Fischer et al., forthcoming.
16 Data used for creating the map was provided automatically by SurveyGizmo from the IP addresses in which the participants were logged in during the survey. This might not in all cases be their actual institutional affiliation. The chart visualizing the ratio of countries represented, however, is based on a question answered by each participant.
17 See the call for papers on the conference website: "With the Big Tent theme in mind, we especially invite submissions from Latin American scholars" (https://dh2011.stanford.edu/?page_id=97).
18 Map created with communitywalk.com. For details, see the raw data collected during this survey which is available anonymously on the website of the TEI MS SIG: http://www.tei-c.org/Activities/SIG/Manuscript/.
19 The French TEI community has been increasingly active over the past few years. One of the outcomes has been the creation of the TEI-FR mailing list, and many local initiatives have provided opportunities for French-speaking TEI training sessions or scholarly discussions. Since Dec. 1st 2010, Lou Burnard has been in charge of a French structure, "Mutualisation d'Expériences pour l'Encodage des Textes" (MEET), within the framework of a larger institution, ADONIS. MEET's mission is to propose a roadmap for the development and promotion of the TEI in France. See http://meet.tge-adonis.fr/.
20 Answer #163 to the question "How can the work of the Special Interest Group for Manuscripts be improved?"
21 This is an issue that is particularly addressed by the participants in a different part of the survey, regarding possible contributions to the MS SIG: "Issues involving cyrillic and glagolitic [manuscripts]" (answer #129 to the question "On what would you be most likely to contribute to this SIG?") and "using TEI for the description of manuscripts in Arabic script" (answer #114 to the same question).
22 More insight into this topic is promised in Hirsch, forthcoming.
23 Answer #141 to question "If you chose 'other' as one of the 3 main obstacles, please elaborate".
24 The average length of the gap is 1.5 years. The gap length has been calculated from the data of the respondents, excluding those who had responded "More than 10 years" to one or both of the questions.
25 The SADE platform (Scalable Architecture for Digital Editing), successfully used in a class at the Leipzig European Summer University "Culture & Technology", addresses this problem. See Malte Rehbein and Christiane Fritze, "Hands-On Teaching Digital Humanities: A Didactic Analysis of a Summer School Course on Digital Editing," in Hirsch, forthcoming,) for an evaluation of this approach.
26 This number does not quite match course attendance as a learning method (18 versus 22; cf. above).
27 This was also expressed by several answers to the question "How can the work of the Special Interest Group for Manuscripts be improved?" (answers #109, #168, #191).
28 TEI by Example (http://tbe.kantl.be/TBE/) had just been launched by the date of the survey. See Terras, Van Den Branden, and Vanhoutte 2010.
29 Answer #124 to question "In what areas do you wish the Guidelines to be improved?"
30 However, 14 participants did not give an explicit answer regarding their publication medium, although "individual use (not for publication)" was marked only by six participants. This discrepancy is due to a deficiency in the design of the survey as we did not divide the question "for what purpose do you mostly use the TEI" into the publication medium on the one hand and purpose/product on the other. This was done only in the evaluation of the results but has the drawback in that participants were not forced to answer both aspects of this question.
31 Their specification is: "text corpora", "quantitative philological research", and "markup for use by analytical software".
32 Naming taken from Roma (http://www.tei-c.org/Roma/).
33 A deeper look into the data reveals that from the 11 answers that use msdesc but not transcr, six state "cataloging" as the purpose of their encoding work.

34 For a description of the revision of this section, see Wittern, Ciula, and Tuohy 2009.
35 Answer #106 to question “In what areas do you wish the Guidelines to be improved?”
36 13% (multiple answers were allowed) use Notepad. Other products are used by four participants or less.
37 Cf. section 3.2.1, particularly figure 6.
38 Cf. section 3.1.1.
39 Cf. section 3.1.1.
40 In 2010, an average of 6.5 mails per month were posted on the MS SIG list (total: 78), by 24 different posters (out of approximately 180 subscribers). Statistics drawn from the list’s archives: http://listserv.brown.edu/archives/cgi-bin/wa?A0=TEI-MS-SIG.
41 Out of the 14 answers, answer #135, “??”, was left aside.
42 Answer #153.
43 Answers #170 and 201.
44 Answers #176 and 181.
45 Answers #106: “I haven’t had the occasion or need yet”, #151: “Find solution without posting”, #167: “Our practices have stabilized”, and #209: “we have solved our problems inside the group”.
46 Out of the 22 answers, two were indecisive (#122: “Not sure.” and #167: “?”).
47 Answers #117, 119, 141, 170, 172, 195, and 201.
48 Answer #114 to the question “On what would you be most likely to contribute to this SIG?”: “using TEI for the description of manuscripts in Arabic script”, and answer #129: “Issues involving cyrillic and glagolitic MSS. […]”.
49 Answers #111 and 163.
50 A task force was set up for the revision of the critical apparatus, and the work on manuscript description was decided to be conducted in “three directions […]: objects (mss., early prints etc.), time (medieval, modern, …), space (western mss., arab mss., …)” (TEI MS SIG 2010).
51 Out of the 20 answers, one was left aside (#112: “no comment”).
52 Answers #114, #117, #119, #146, #153, #154, #155, #170, #195 and #212.
53 For instance, answer #119: “The other members of the team do not understand anything computer-related beyond using MS Word. I have difficulty explaining why things have to be done a certain way, and why it’s easier for me if we use with a different workflow.”
54 Answer #212.
55 Particularly answers #153, #154 and #195.
56 Answer #153.
57 Answer #170.
58 Cf. section 3.2.3.
59 Answers #141, #157, #163, #172, #194, #196, and #204.
60 Answer #204. As a side note, our Gender Studies colleagues might be interested to notice how, in this answer, the respondent referred to the humanist with “he/she” or “his/her”, but referred to the IT specialist as “a very specialised IT man”.
61 Unfortunately, a bug in the early stages of the survey was preventing people from selecting “All” as an answer. It was later corrected, and we updated the data for two users who had mentioned this problem.
62 Answer #169 to the question “In what areas do you wish the Guidelines to be improved?”: “23 Using the TEI”.
63 Answers #109, #114, #119, #127, #129, #131, #152, #169, #185 and #195.
64 See for instance answer #185.
65 Answer #127.
66 Answers #115, #157, #170, #172, #191, #194 and #203.
67 Answer #115.
68 Answer #194.
69 Answers #112, #122, #126, #146, #158, #165 and #172.
70 Answer #122.
71 Answer #126.

72 Answer #158.
73 Answer #146.
74 Answer #165.

Cite this article

Electronic reference

Marjorie Burghart and Malte Rehbein, « The Present and Future of the TEI Community for Manuscript Encoding », *Journal of the Text Encoding Initiative* [Online], Issue 2 | February 2012, Online since 03 February 2012, connection on 01 April 2012. URL : http://jtei.revues.org/372 ; DOI : 10.4000/jtei.372

Authors

Marjorie Burghart
Marjorie Burghart is a research officer at the Ecole des Hautes Etudes en Sciences Sociales (EHESS) in Lyon, France, head of the Digital Humanities programme of the CIHAM UMR 5648, a research centre in Medieval Studies. She is member of the executive board of the Digital Medievalist community, the representative of the ALLC on the Publications Committee of the ADHO, and has recently been elected to the Board of Directors of the TEI (term 2012–2013). Her main topics of interest include medieval sermon studies and digital scholarly editing.
Malte Rehbein
Malte Rehbein is research associate in the department of Computerphilology at Würzburg University, director of the Würzburg Research Centre for Digital Editing, and lecturer in Digital Humanities and Medieval History. He is co-chair of the TEI special interest group for manuscripts, member of the executive board of the Digital Medievalist and Editor-In-Chief of the Digital Medievalist Journal. He has published on digital scholarly editing, textual variation, didactics, medieval history and digital palaeography.

Abstract

This article provides a detailed analysis of the current state, needs, and desires of members of the TEI community working with manuscript material, based on the results of a survey carried out by the authors. An analysis of the survey results provides insights into the practices, problems, and limitations of the community utilizing the TEI for manuscript encoding. The results demonstrate the existence of a steep learning curve for the TEI, where many practitioners are self-taught and where learning-by-doing dominates; there exists a long gap between the first encounter with the TEI and its actual use in projects. Survey results highlight the need for user-friendly, bespoke tools facilitating the processing, analysis, and publishing of TEI-encoded texts. Feedback on the Guidelines themselves reveals aspects that do not fully meet the needs of those encoding manuscript material. To better address these needs, a strengthening of the Special Interest Groups is proposed.
Keywords : community, manuscript encoding, Special Interest Groups, Survey