Achtung!
Dies ist eine Internet-Sonderausgabe des Aufsatzes
„Electronic Resources Development and SEER.
The Preparation and dissemination of electronic resources
pertaining to Eastern studies“
von Jost Gippert (1998).
Sie sollte nicht zitiert werden. Zitate sind der Originalausgabe in
Proceedings [of the] Annual Meeting of the Pacific Neighborhood Consortium, May 15-18, 1998, Taipei,
Taipei: Academia Sinica 1998, 361-370
zu entnehmen.

Attention!
This is a special internet edition of the article
“Electronic Resources Development and SEER.
The Preparation and dissemination of electronic resources
pertaining to Eastern studies”
by Jost Gippert (1998).
It should not be quoted as such. For quotations, please refer to the original edition printed in
Proceedings [of the] Annual Meeting of the Pacific Neighborhood Consortium, May 15-18, 1998, Taipei,
Taipei: Academia Sinica 1998, 361-370.

Electronic Resources Development and SEER

The Preparation and dissemination of electronic resources pertaining to Eastern studies

Jost GIPPERT, Frankfurt a/M (Germany)

During the 35th International Congress of Asian and African Studies (Budapest, 7-12 July 1997), an informal group of scholars who are engaged in the preparation and dissemination of electronic resources pertaining to Eastern studies, established itself with a view to a coordination of efforts. A main task of the "SEER" group ("Scholars Engaged in Electronic Resources") will consist in adapting the means as provided by the World Wide Web to make electronic resources of all kinds readily available for scholarly and scientific purposes. The types of resources to be treated in this respect include written sources (manuscripts, texts) in various languages (ancient and modern) as well as images (e.g., pictures of architectural sites) and databases (e.g., historical or statistical data) of all kinds.

In the present paper, requirements and possible solutions of this task are discussed with respect to one major type of resources, viz. manuscripts containing textual data. A proper example to show what problems have to be solved in this respect, is the project of a digitization of the Tocharian manuscripts preserved in the Berlin Turfan Collection. This represents the largest collection of sources written in the extinct Tocharian language[1] which was spoken in the Hsinkiang area in the first millennium A.D.[2] The digitization has been undertaken as part of the "TITUS" project[3] since 1996 in a joined effort by the Berlin-Brandenburgische Akademie der Wissenschaften, the Staatsbibliothek Berlin, the Institut für Vergleichende Sprachwissenschaft of Frankfurt University, and the Tamai foundation. It aims at establishing and providing several kinds of electronic data, viz. digitized images of the manuscripts as well as textual data comprising

[1] Scholarly tradition nowadays assumes at least two separate variants of Tocharian to have existed, named A- (or East-) and B- (or West-) Tocharian, resp. Whether these have to be treated as separate languages or, rather, dialects of one language, need not be discussed here.

[2] Other collections of Tocharian manuscripts exist in the Bibliothèque Nationale (Paris), the British Library (London), the St. Petersburg Academy of Sciences, and in several museums of the Hsinkiang area and other parts of China (?). It is to be hoped that in due time, the manuscripts contained in these collections will as well be digitized so that they can be linked with the WWW edition of the Berlin manuscripts.

[3] "Thesaurus Indogermanischer Text- und Sprachmaterialien" (Thesaurus of Indo-European Text and Language Materials); cf. the URL `http://titus.uni-frankfurt.de/texte/texte.htm`.

a descriptive catalogue and a thorough transcription of the texts contained in the manuscripts. The project was first outlined in a short notice in Tocharian and Indo-European Studies 7, 1997, 265-266 ("Digitization of Tocharian Manuscripts"; also available under the URL `http://titus.uni-frankfurt.de/texte/tocharic/index.html`); its aims and methods were discussed at large in "Tocharisch mit dem Computer: Ziele und Verfahren" (in: Tocharian and Indo-European Studies 7, 1997, 17-34) and "Digitization of Tocharian Manuscripts from the Berlin Turfan Collection" (to appear in: Manuscripta Orientalia, St. Petersburg, 1998).

It has to be noted that a major task of the project consists in the preservation of the manuscripts[4]. With respect to this task, digitization, i.e., electronic storage of the manuscript contents as graphic images, is an auxiliary method only. On the basis of the World Wide Web technology, however, the digital data produced in this connection can easily be adapted to both the task of documentation, i.e., making the data accessible to the public, and the task of analysis, i.e. investigating the manuscripts with a view to philological, linguistic, and palaeographic questions.

The main feature of WWW technology that has proved helpful in this respect is its capability of linking various kinds of data. In the given example, several types of both textual and graphical data to be processed and linked have to be considered. The textual data in question comprise, as was noted above, a descriptive catalogue which informs about the provenance of each manuscript, its size, actual state, signature in the collection, etc.; cf. Figure 1 showing a screen shot of the database (in DataPerfect format) as established by Dr. Schmieder-Jappe in the Berlin Staatsbibliothek. Another type of textual data to be treated is the linguistic contents of the manuscripts proper, i.e. electronic adaptations of the texts as contained in them. This requires a thorough reading of each manuscript item; the interpretations as published in former printed editions[5] should be collated throughout during this process, and they can be used as support-

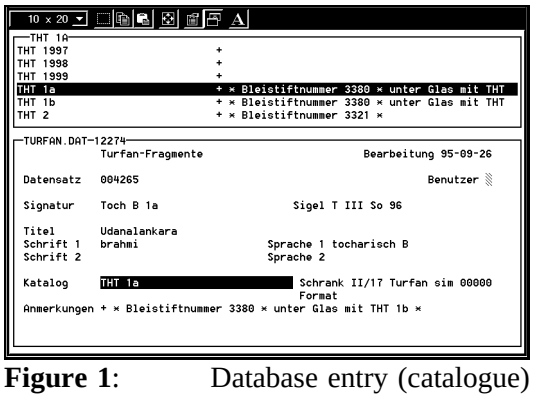

Figure 1: Database entry (catalogue)

[4] Parts of the present paper were also read on a conference dedicated to the Preservation of Dunhuang and Central Asian collections which took place in Berlin, 13-15 May, 1998.

[5] About one fourth of the Berlin manuscripts have been edited by E. SIEG and W. SIEGLING (East-Tocharian: "Tocharische Sprachreste", I. Band: Die Texte, A. Transcription; B. Tafeln, Berlin und Leipzig 1921; West-Tocharian: "Tocharische Sprachreste, Sprache B", Heft 1-2, Göttingen 1949 / 1953; the latter edition was reprinted with improvements by W. THOMAS, Göttingen 1987).

ing evidence. By now, about one fourth of the Berlin manuscripts has been electronically transcribed and transliterated for the project by Chr. SCHAEFER and T. TAMAI.

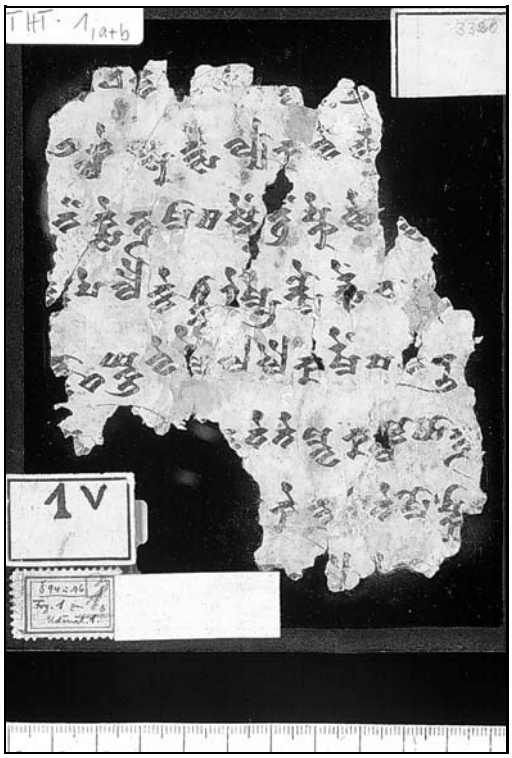

Figure 2: Digitized image of THT 1ABr

The graphical data to be processed mostly consist of digitized images of the original manuscripts. With a view to the tasks of preservation, documentation, and analysis, different quality standards have to be envisaged: For a WWW documentation which is based on screen representation, a lower resolution will be sufficient in most cases; it will also be preferable because it implies reduced file sizes to be transferred via the net. For preservation and analysis as well as for achieving a sufficient printing quality, however, a maximum of resolution should be attained (cf. Figure 2 showing the Berlin manuscript THT 1ABr digitized on the basis of a colour slide photograph with a resolution of ca. 2500 dpi)[6].

Another type of graphical data to be considered is digitized images of older (black and white) manuscript photographs. These can be helpful in two respects: On the one hand, they may be used as additional reference materials documenting a previous state of manuscripts that have suffered damage by external (environmental) factors or improper handling. In some cases, they may even be the only reference material available, viz. whenever the original manuscript has got lost. Unfortunately, the Berlin collection in its present state has large gaps indeed which resulted from the fact that it was evacuated from Berlin for safety reasons during World War II and parts of it have never returned; cf. Box 1 and Box 2 where the items in question (111 of

3, 5, 6, 8, 11, 14, 15, 17, 21, 22, 22, 23, 24, 25, 27, 28, 29, 30, 31, 33, 41, 42, 44, 44, 45, 45, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 66, 67, 93, 106, 108, 111, 112, 113, 114, 115, 116, 131, 147, 175, 176, 177, 194, 197, 199, 216, 230, 231, 236, 261, 263, 265, 267, 268, 269, 272, 304, 305, 308, 312, 313, 323, 332, 332, 342, 343, 345, 360, 373, 375, 383, 396, 397, 400, 405, 408, 421, 421, 433, 435, 492, 494, 497, 499, 514, 515, 516, 520, 548, 572, 572, 573, 591, 593, 598, 602, 602, 602, 602, 607

Box 1: Missing manuscripts (Toch. B: THT 1-633)

[6] For details to be considered when digitizing manuscripts of the kind discussed here, cf. my article in "Manuscripta Orientalia".

634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 695, 696, 697, 698, 699, 700, 701, 702, 704, 705, 707, 708, 710, 711, 712, 712, 722, 723, 724, 727, 729, 730, 731, 732, 747, 748, 749, 750, 751, 752, 753, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 813, 825, 836, 837, 845, 846, 847, 847, 848, 850, 851, 853, 854, 855, 859, 860, 861, 862, 863, 864, 869, 876, 877, 884, 885, 886, 887, 888, 889, 890, 891, 892, 893, 894, 895, 896, 897, 898, 899, 901, 903, 911, 912, 913, 921, 922, 923, 924, 925, 928, 930, 932, 933, 934, 934, 935, 936, 936, 937, 940, 946, 947, 951, 952, 953, 954, 966, 967, 968, 969, 970, 971, 974, 975, 976, 993, 995, 1025, 1028, 1031, 1033, 1037, 1063, 1064, 1065, 1066

Box 2: Missing manuscripts (Toch. A: THT 634-1100)

633 edited B-Tocharian mss. and 153 of 467 edited A-Tocharian mss.) are listed.
For the same reason, it may even be necessary to digitize printed images of manuscripts that have been published in facsimile editions. As this procedure has great disadvantages esp. when a rough raster was used in the printing (this is true for the existing facsimile editions of Tocharian manuscripts prepared in the 20ies), it remains the least preferable option.

It goes without saying that the types of data as listed above have to be ranked in divergent orders according to the respective tasks. For the task of preservation, establishing and processing of graphical and cataloguing data is superior to the treatment of textual contents, the content data being valuable as mirrors of the graphical data only, and the security of storage is superior to the accessibility and handling of the digitized images. For a plain documentation, graphical and textual data can be assumed to be equivalent while accessibility and easy handling are superior to storage. With a view to (philological or palaeographical) analysis, catalogue data will be inferior in comparison with content and graphical data; the latter ones can be regarded as supporting each other, and the structure of the content data requires special treatment depending on the actual purpose of investigation.
These divergences notwithstanding, WWW technology can be applied with a view to a fulfilment of all three tasks, providing a means of arranging and displaying the individual data in a joined environment adaptable to users' preferences. The procedure to be envisaged here is not plain linkage (e.g., from a catalogue entry to the digitized image of a manuscript) but using separate frames as the basic output structure for the different data types to be linked. The capability of defining frames to be arranged within a browser window has only recently been added to the HTML standard ("Hypertext markup language") used for WWW publishing, and an up-to-date version of graphically based net browsers will be required to make use of it[7]; nevertheless its advantages should not be neglected when attempting to publish hierarchically linked data structures

[7] Frames of the type used here are supported by Netscape Navigator / Communicator versions 3.0 and higher and Microsoft Internet Explorer version 4.0 and higher.

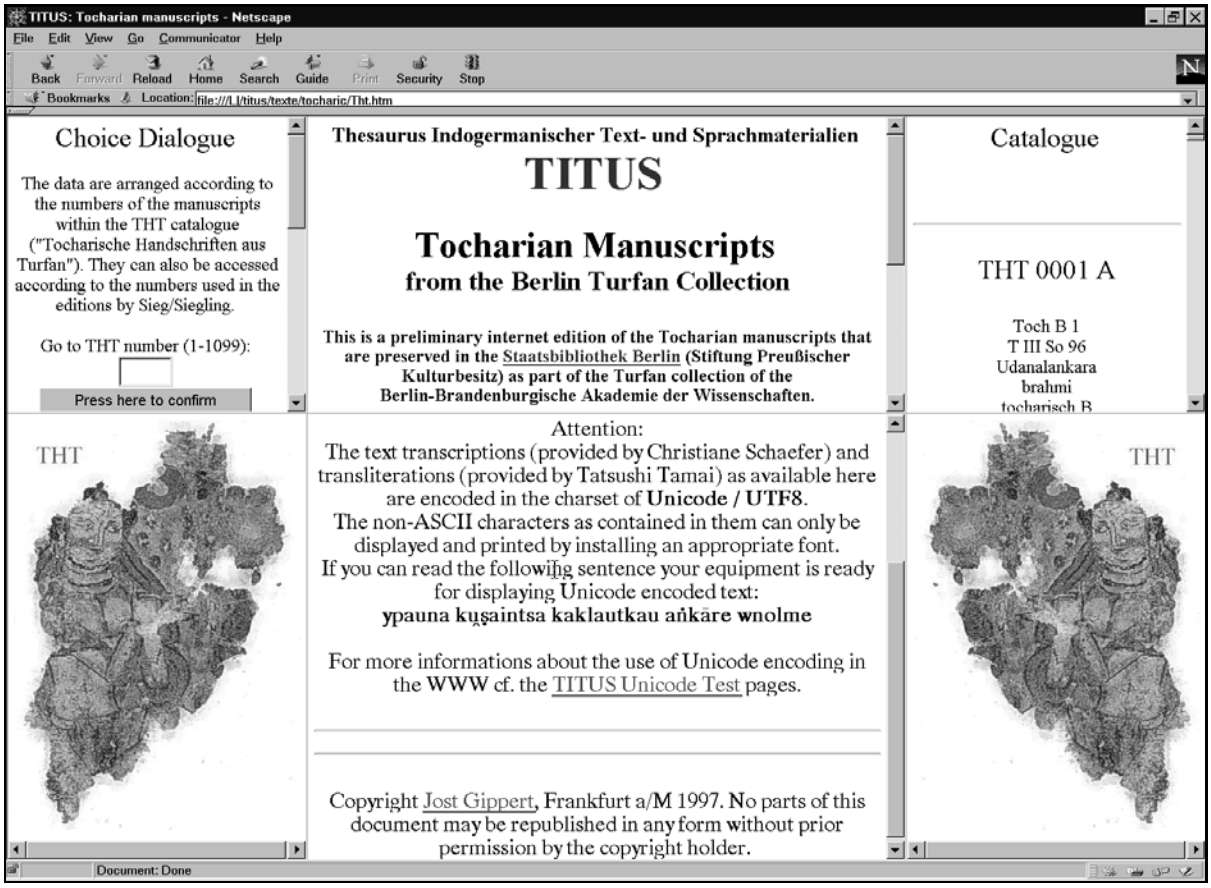

Figure 3: WWW start page of hypertext documentation

in the WWW. As an example, Figure 3 shows the starting page of the Tocharica project (future URL: `http://titus.uni-frankfurt.de/texte/tocharic/tht.htm`) which has been divided into six frames: A main frame informing about the project and its aims (upper half, middle), a dialogue frame for the selection of a catalogue item to be displayed (upper left corner), a frame showing the corresponding catalogue entry (upper right corner), two frames for displaying the two sides of a manuscript leaf (lower right and left corners), and one frame showing the textual contents (lower half, middle). The actual arrangement can be seen in Figure 4 which illustrates the output of the respective data of manuscript item nr. THT 74rv[8] after this was selected via the dialogue frame and both its graphical and its textual representations were loaded using the links as present in the catalogue frame. The images to be displayed in frames are low resolution graphics only; whenever a high resolution variant of the images is available, this can additionally be retrieved in an extra window as Figure 5 demonstrates.

[8] The numbers of the THT ("Tocharische Handschriften aus Turfan") catalogue agree with the numbers used in the printed editions in the following way: Nrs. 1 to 633 ≈ nrs. 1 to 633 of edited B-Tocharian texts, nrs. 634 to 1099 ≈ nrs. 1 to 465 of edited A-Tocharian texts. Nrs. 1100 to 4072 have not been published before.

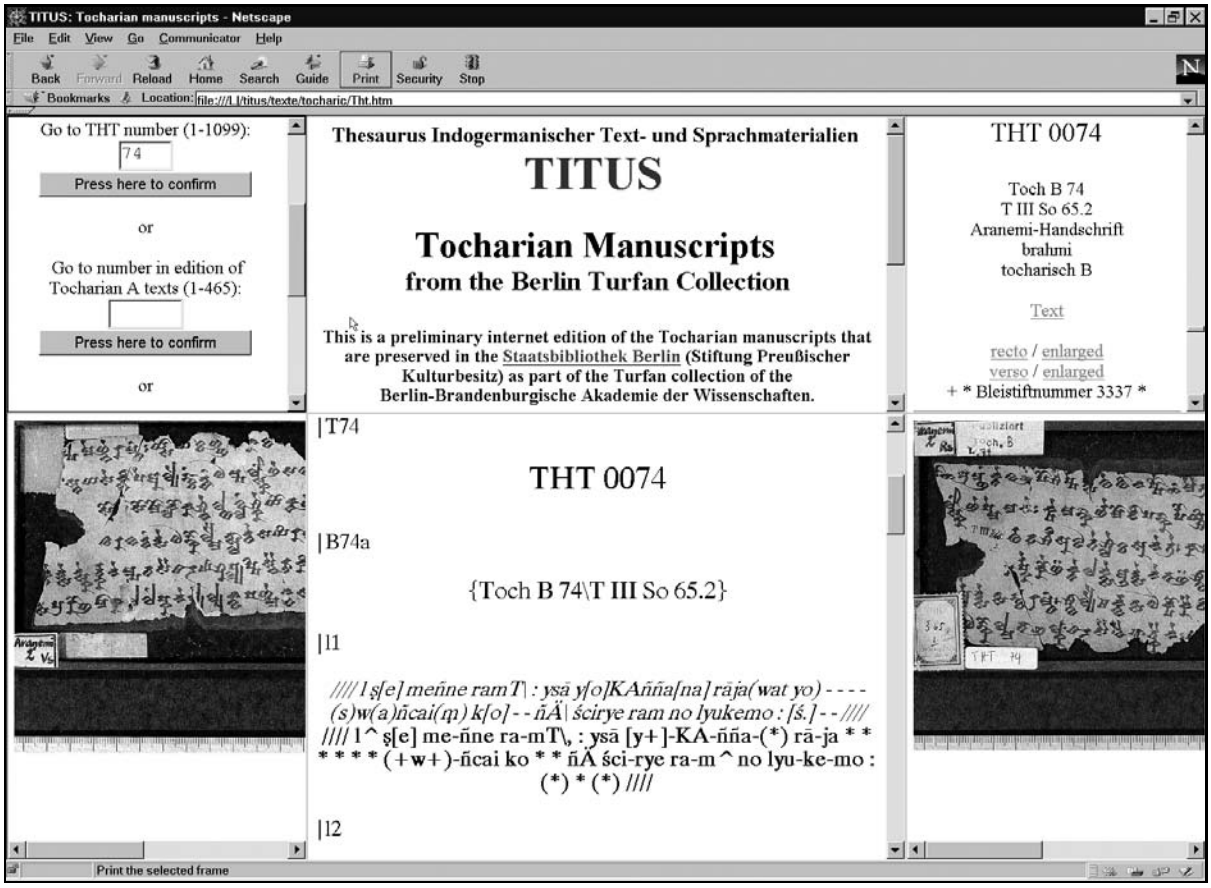

Figure 4: Selection and output of THT 74rv

Several aspects of data preservation and and processing have to be considered when a WWW edition of this type is aimed at. Starting from the requirements of preservation, we first have to note the problem that a huge amount of disk space is necessary for the high resolution images that have to be produced for this purpose; a colour slide of 24 × 36 mm, e.g., will yield a graphics file of 26 MB when digitized with a resolution of 2700 dpi. Although the necessary space can be reduced to a high extent by applying image compression methods such as the one provided by the widely used JPG-format, this has the disadvantage of also reducing the quality of graphical images. For high resolution images that are meant to be integrated in a WWW publication, this disadvantage can be neglected because the loss of information a compression of less than 15 percent brings about, will hardly be visible when the image is displayed on a computer screen or printed out on a standard laser printer. In any way, the capacity of server disks may remain a decisive factor for the resolution to be chosen, and the time necessary for data transfer of large files should also be taken into account. But even if the resolution to be made available via the net must be lower than the one used for preservational storage, the necessary files can normally be "derived" from the ones having a higher resolution.

Another problem that has to be considered is the number of data files to be stored and handled via their names. The best solution for keeping a large

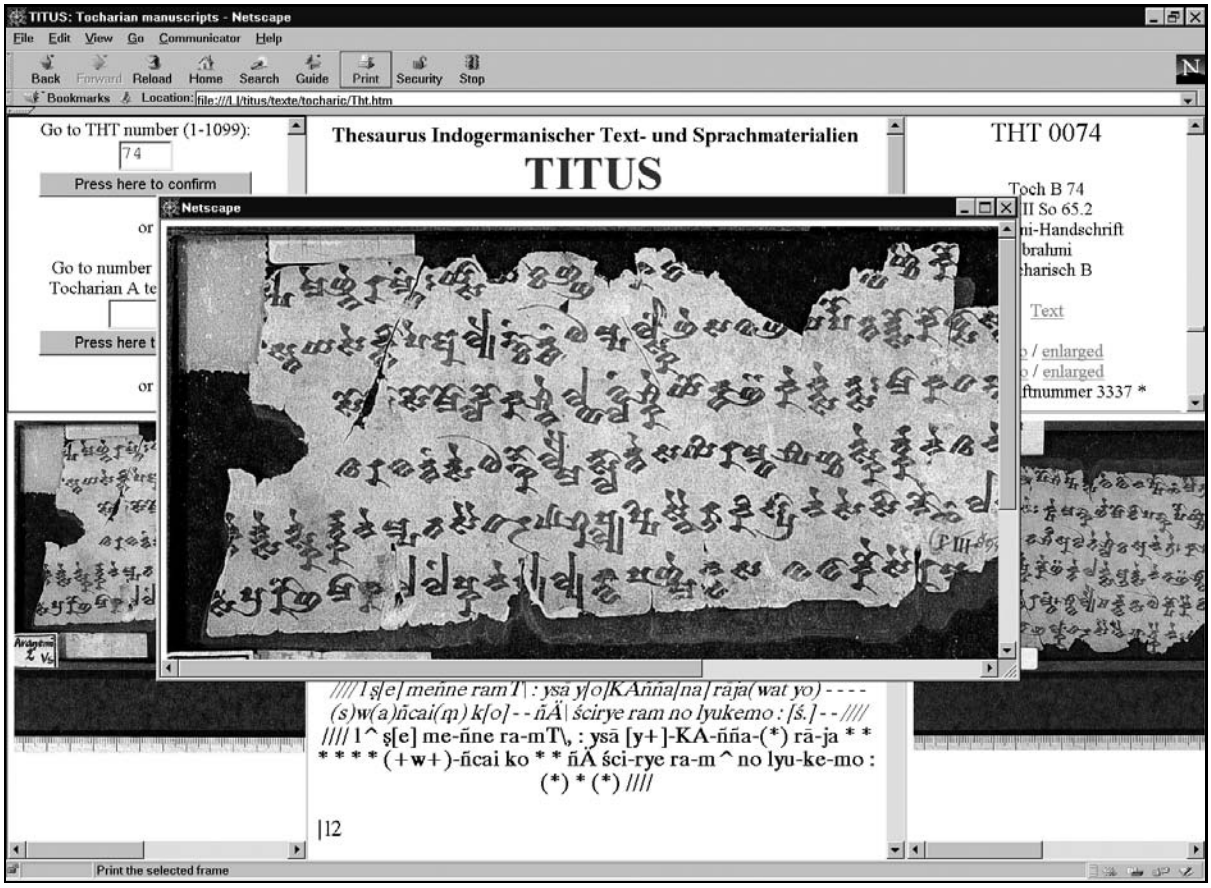

Figure 5: Switching between frames and extra windows

number of files readily retrievable consists in applying a transparent, "predictable" file naming system right from the beginning. For the Berlin project, an eight-byte system reflecting the catalogue number as well as informations about the manuscript items or parts has proved sufficient; e.g., a name like "0001ABRT.JPG" denotes the photo showing the *total* view ("T") of the *recto* ("R") of both *fragments A and B* ("AB") of *THT number 1* ("0001")[9].

For the data contained in the catalogue, a first task consists in their conversion into the HTML format as required for all kinds of textual materials to be disseminated via the WWW. Whenever the conversion can start from a structured database as in the given example, the conversion proper will be a minor problem with respect to data arrangement and visual representation. We have to bear in mind, however, that the resulting HTML structure may not be as open for a retrieval of informations as the underlying database itself may be, e.g. if it is to be used for a search of manuscript provenances, for an arrangement

[9] A minor problem in this connection consisted in the fact that in the (older) German catalogue, "R" was used for "Rückseite" which is *verso* while "V" meant "Vorderseite" which is *recto*; moreover, SIEG/SIEGLING used "a" and "b" for *recto* and *verso*, resp., in their editions while "A", "B" etc. are now being used for individual parts or fragments of manuscripts in the THT catalogue. It goes without saying that a file naming system of the type discussed here must be consistent in the conventions it uses.

according to the size of manuscripts, or the like. This is a problem of browsers rather than the HTML language: Today's net browser handle HTML files as plain text files; and the search function they provide is applicable to a so-called sequential search of text elements (normally words) only, not to more "sophisticated" searches regarding, e.g., elements contained in certain "fields" of "records" in a database. Even the simple selection function as shown in Figure 4 requires an additional, non-genuine extra program (to be written as a "java" or "javascript" applet) to work successfully in the given environment. If a plain search for the number "74" were applied to the HTML catalogue, we would not only find the item searched for (THT 0074), but also any other items containing the digits "7" and "4" together (e.g., THT 1740 or TochA 74).

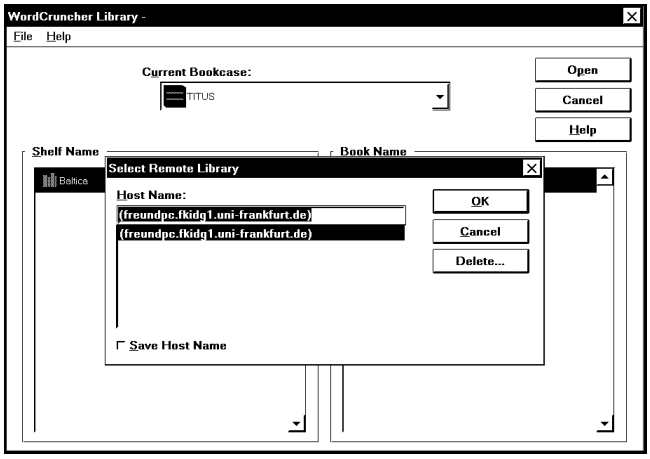

Figure 6: Wordcruncher remote library access

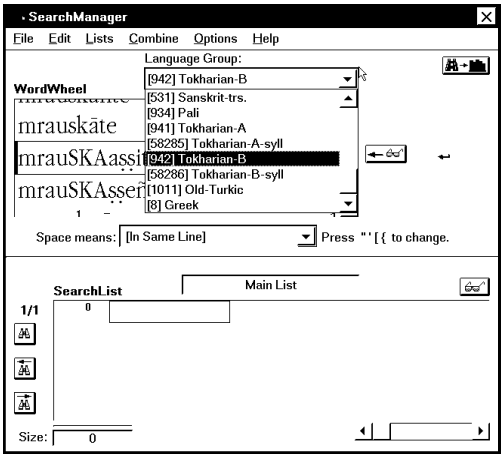

Figure 7: Word wheel choice

The same problem, i.e., the lack of a satisfactory search engine applicable to a WWW/HTML environment, has a bearing on the text data proper as well. Whenever a WWW publication is not meant to display text contents for easy reading only but (linguistic or philological) analysis is intended, the search function of browsers will be hardly sufficient. If, e.g., the B-Tocharian word *ram* "like" as appearing in THT 74 is to be searched for in a browser, other words containing the same sequence of characters such as *ramer* "fast" or *kramartse* "heavy" will also be found. At present, it seems indeed hardly possible to find an "intelligent" search engine for linguistic units in an HTML environment. There is an independent solution for the WWW, however, which is not based on HTML. This is the Wordcruncher text retrieval system as developed by Brigham Young University[10]. Its main feature is a preindexation of the texts to be analyzed; after preindexation, the text elements can immediately be accessed and displayed in their textual environment via a search engine (the so-called word wheel, cf. Figure 7) which runs both on a local machine and from a remote server (cp. Figure 6). Given that linkage of textual and graphical

[10] For details cf. the URL `http://www.wordcruncher.com` where a free version of the search programm, the so-called "viewer", is downloadable from.

data is also possible in the preindexed texts (cf. Figure 8 and Figure 9), the Wordcruncher system seems to be a considerable alternative to using HTML structures[11]. It has a shortcoming, however, in that it is still restricted to usage on MS-Windows based computers[12].

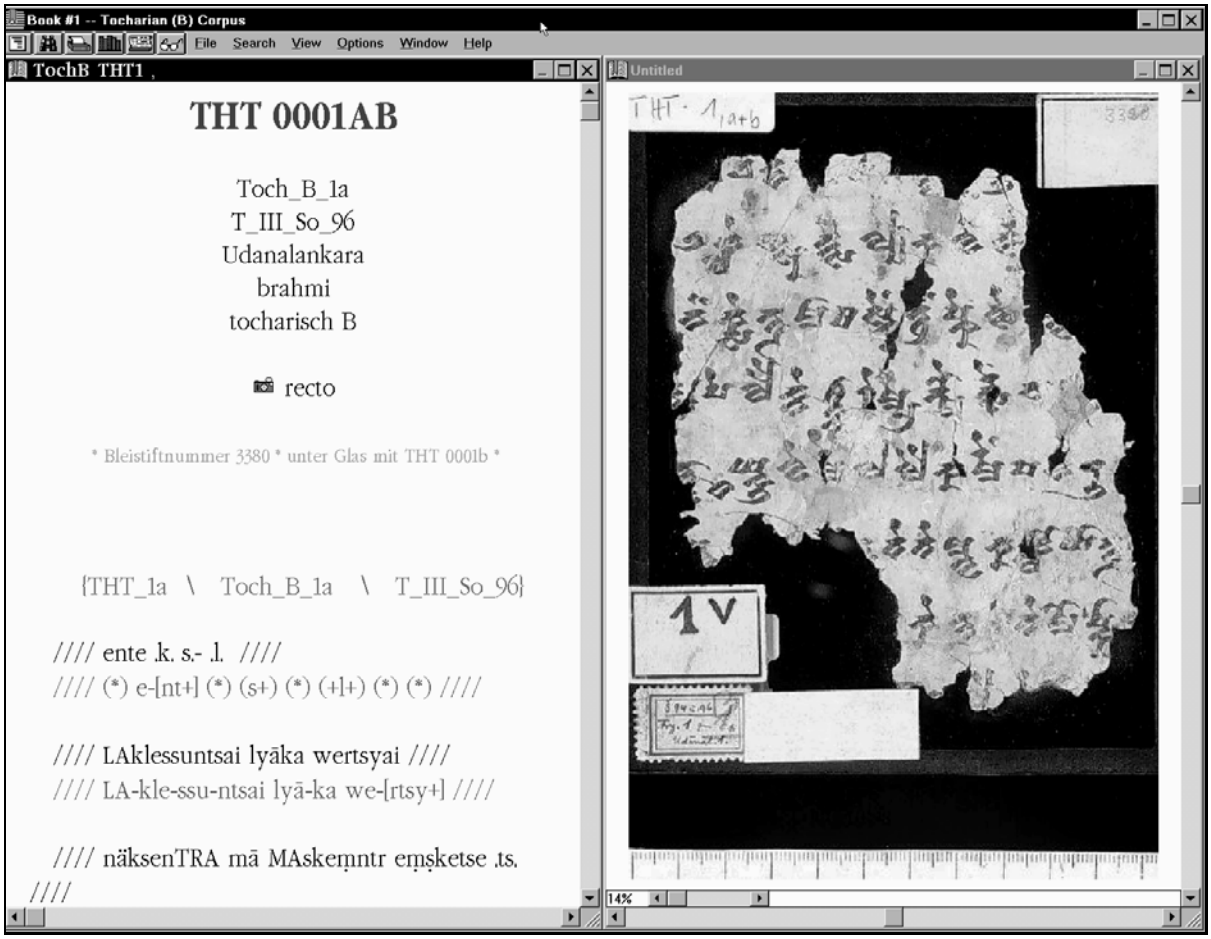

Figure 8: Representation of textual and graphical data in Wordcruncher

Another problem that has to be solved with respect to a WWW publication of textual data, is the encoding of both original scripts and transcriptional or transliterational systems as used in the example discussed here. For the Brāhmī script that was adapted by the Tocharians from a Northern Indic model, there is no encoding standard available at all, and even the symbols necessary for a transcriptional rendering are not covered completely by any one of the existing encoding patterns that have been accepted for HTML. The problem becomes even larger when different scripts are to be mixed within a text document to be

[11] For a thorough examination of the applicability of the Wordcruncher retrieval system, cf. my forthcoming article "Multilingual text retrieval: Requirements and solutions" (to appear in *Studia Iranica, Mesopotamica et Anatolica* 3, 1998; a short version will also appear in the proceedings of the 2nd Symposium on Language, Logic, and Computation, Tbilisi 1998).

[12] The Wordcruncher server module can only be installed on Windows-NT servers (4.0), the viewer program runs on Windows 3.1.1 and higher versions.

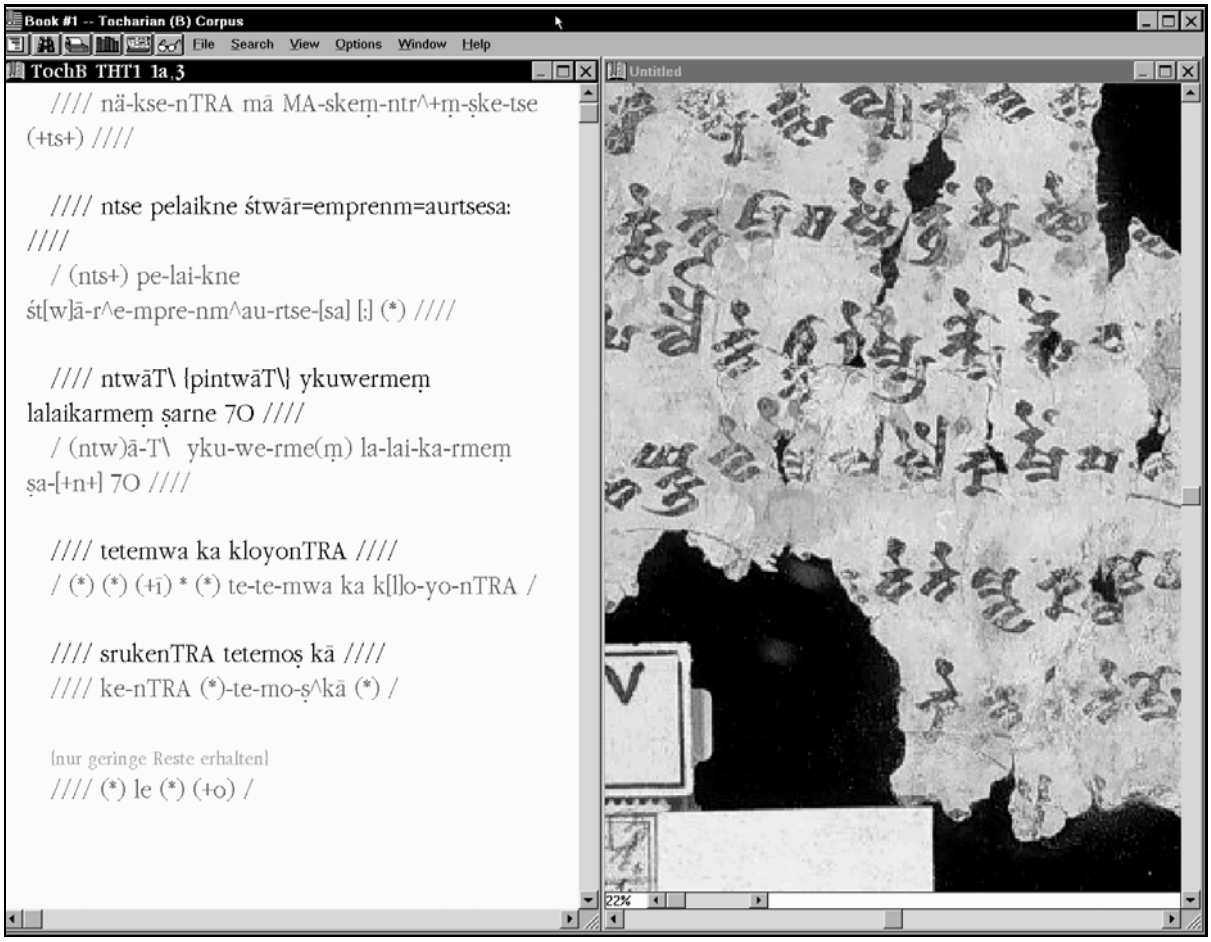

Figure 9: Representation of textual and graphical data in Wordcruncher

displayed and/or analyzed. A solution which is only just developing is the so-called "Unicode" standard which in its present stage (version 2.1) comprises nearly all script systems used in today's national alphabets (including some 40.000 Chinese characters) as well as an increasing set of ancient script systems and transcriptional elements. It is to be hoped that further additions will rapidly be adopted to this standard so that in the near future, even the Tocharian type of Brāhmī script will be encodable[13]. Restricting the textual representation to a transcriptional and/or transliterative Unicode rendering as in the example discussed here, can only be an interim solution. It will be one of the tasks of the "SEER" group to collect opinions and proposals as to the requirements of scripts and script elements to be encoded in the WWW environment and to thus support its further improvement as a basis of data exchange.[14]

[13] The further extension of the Unicode standard is one of the aims of the TITUS project; cf. the URL `http://titus.uni-frankfurt.de/unicode/unitest.htm` which provides informations and tools.

[14] As a first step, a homepage where informations about actual developments can be retrieved from, was established on the TITUS server; cf. the URL `http:/titus.uni-frankfurt.de/seer/index.htm`.