Digitization of Tocharian Manuscripts from the Berlin Turfan Collection
Methods of linking textual and graphic data

Jost GIPPERT, Frankfurt a/M

The Berlin Turfan Collection 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 project I wish to talk about here has been undertaken as part of the "TITUS" project[3] since 1996, in a joint effort by the Berlin-Brandenburgische Akademie der Wissenschaften / Staatsbibliothek Berlin, the Institut für Vergleichende Sprachwissenschaft of Frankfurt University, and the Tamai foundation. It aims to establish and provide several kinds of electronic data, viz. digitized images of the manuscripts as well as textual data comprising a descriptive catalogue and a comprehensive 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 at the URL `http://titus.uni-frankfurt.de/texte/tocharic/index.html`); its aims and methods were discussed at length 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" (in: Manuscripta Orientalia 4/1, 1998, 49-57).

It is 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 documenta-

[1] Scholarly tradition nowadays assumes that at least two separate variants of Tocharian existed, namely A- (or East-) and B- (or West-) Tocharian respectively. Whether these have to be treated as separate languages or, rather, as 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 also be digitized so that they can be linked with the WWW edition of the Berlin manuscripts.

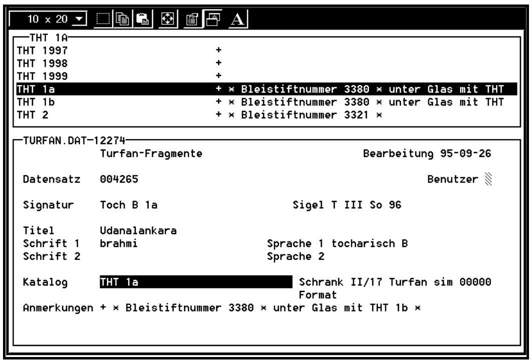
[3] "Thesaurus Indogermanischer Text- und Sprachmaterialien" (Thesaurus of Indo-European Text and Language Materials); cf. the URL `http://titus.uni-frankfurt.de/texte/texte.htm`.

[4] Parts of the present paper were also read on the Annual Meeting of the Pacific Neighborhood Conference which took place in Taipei, 14-20 May, 1998.

tion, 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 capacity to link various kinds of data. In this context, several types of both textual and graphic data to be processed and linked have to be considered. The textual data in question comprise, as was noted above, a descriptive catalogue giving information 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 careful reading of each manuscript item; the interpretations published in former printed editions[5] should be collated throughout during this process and can be used as supporting evidence. By now, about one quarter of the Berlin manuscripts have been electronically transcribed and transliterated for the project by Chr. SCHAEFER and T. TAMAI.

The graphic 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. Indeed, this will even be preferable because it entails reduced file sizes for transfer via the net. However, a maximum resolution is required for preservation and analysis as well as for achieving an adequate printing quality (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 graphic data to be considered is digitized images of older (black and white) photographs of manuscripts. These can be helpful in two respects. On



```
THT 1A
THT 1997                +
THT 1998                +
THT 1999                +
THT 1a                  + × Bleistiftnummer 3380 × unter Glas mit THT
THT 1b                  + × Bleistiftnummer 3380 × unter Glas mit THT
THT 2                   + × Bleistiftnummer 3321 ×

TURFAN.DAT-12274
            Turfan-Fragmente                      Bearbeitung 95-09-26

Datensatz   004265                                         Benutzer

Signatur    Toch B 1a                   Sigel T III So 96

Titel       Udanalankara
Schrift 1   brahmi                  Sprache 1 tocharisch B
Schrift 2                           Sprache 2

Katalog     THT 1a                   Schrank II/17 Turfan sim 00000
                                     Format
Anmerkungen + × Bleistiftnummer 3380 × unter Glas mit THT 1b ×
```

Figure 1: Database entry (catalogue)

[5] About one quarter 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).

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

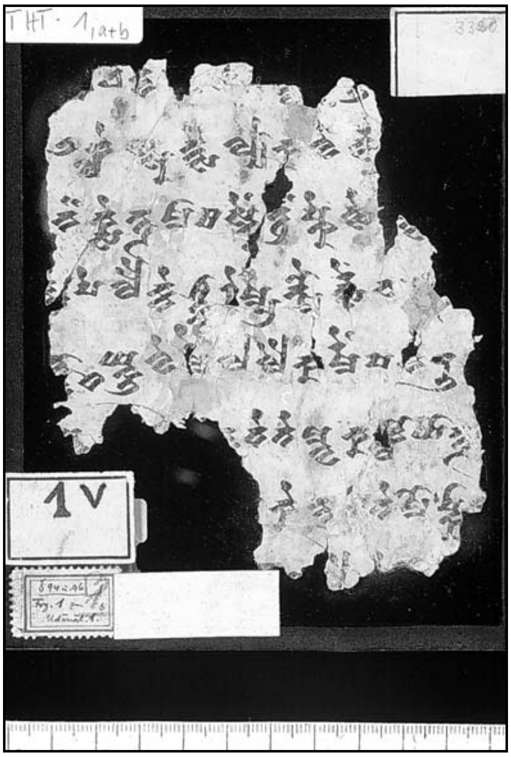


Figure 2: Digitized image of THT 1ABr

the one hand, they may be used as additional reference materials documenting the previous state of manuscripts that have suffered damage as a result of 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 very large gaps owing to its evacuation from Berlin for safety reasons during World War II, after which parts of it were never returned; cf. Box 1 and Box 2 where the items in question (111 of 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 twenties), it remains the least desirable option.

It goes without saying that the types of data listed above have to be ranked differently according to the tasks envisaged. For the task of preservation, the establishing and processing of graphic and catalogue data is preferable to the treatment of textual contents, the content data being valuable only as mirrors of the graphic data, and the security of storage has priority over the accessibility and handling of the digitized images. For a plain documentation, graphic and textual data can be assumed to be more or less equal while accessibility and easy handling are preferable to measures of storage. With a view to (philological or palaeographical) analysis, catalogue data will be less important than content and graphic data; the latter 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 notwith-

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)

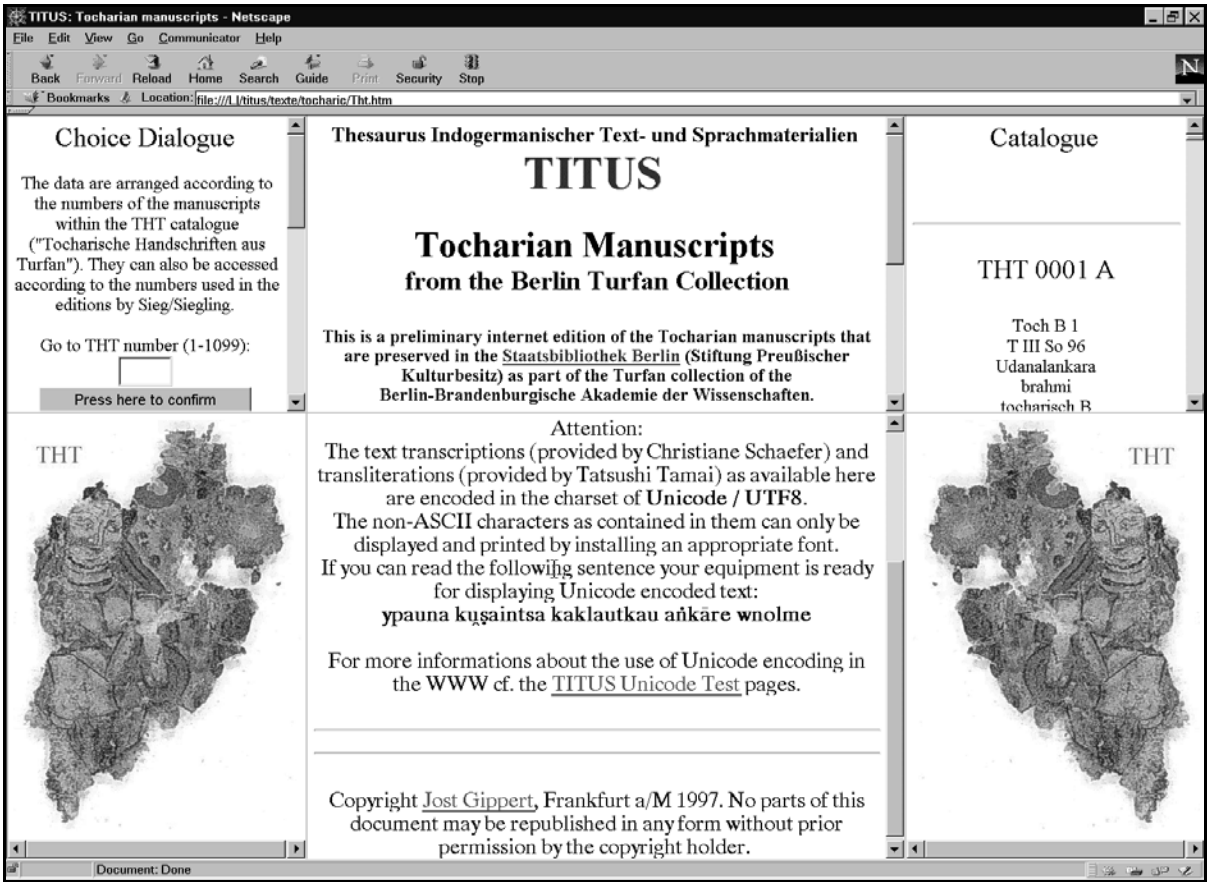
634, 635, 636, 637, 638, 639, 640, 641, 642, 643, 644, 645, 646, 647, 648, 649, 650, 651, 652, 653, 654, 655, 656, 657, 658, 695, 696, 697, 698, 699, 700, 701, 702, 704, 705, 707, 708, 710, 711, 712, 712, 722, 723, 724, 727, 729, 730, 731, 732, 747, 748, 749, 750, 751, 752, 753, 777, 778, 779, 780, 781, 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)

standing, WWW technology can be applied with a view to fulfilling all three tasks, providing a means of arranging and displaying the individual data in a common environment adaptable to users' preferences. The procedure 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 in the WWW. As an example, Figure 3 shows the starting page of the Tocharica project (interim URL: `http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/tht.htm`) which has been divided into six frames: A main frame giving information 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 had been selected via the dialogue frame and both its graphic and its textual representations had been 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 too can be retrieved in an extra window as Figure 5 demonstrates.

[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.

[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.


TITUS: Tocharian manuscripts - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: file:///L|/titus/texte/tocharic/Tht.htm
THT
THT
Document: Done

Choice Dialogue

The data are arranged according to the numbers of the manuscripts within the THT catalogue ("Tocharische Handschriften aus Turfan"). They can also be accessed according to the numbers used in the editions by Sieg/Siegling.

Go to THT number (1-1099): □

Press here to confirm

Thesaurus Indogermanischer Text- und Sprachmaterialien

TITUS

Tocharian Manuscripts
from the Berlin Turfan Collection

This is a preliminary internet edition of the Tocharian manuscripts that are preserved in the Staatsbibliothek Berlin (Stiftung Preußischer Kulturbesitz) as part of the Turfan collection of the Berlin-Brandenburgische Akademie der Wissenschaften.

Attention:
The text transcriptions (provided by Christiane Schaefer) and transliterations (provided by Tatsushi Tamai) as available here are encoded in the charset of **Unicode / UTF8**.
The non-ASCII characters as contained in them can only be displayed and printed by installing an appropriate font.
If you can read the following sentence your equipment is ready for displaying Unicode encoded text:
ypauna ku̯ṣaintsa kaklautkau aṅkāre wnolme

For more informations about the use of Unicode encoding in the **WWW** cf. the TITUS Unicode Test pages.

Copyright Jost Gippert, Frankfurt a/M 1997. No parts of this document may be republished in any form without prior permission by the copyright holder.

Catalogue

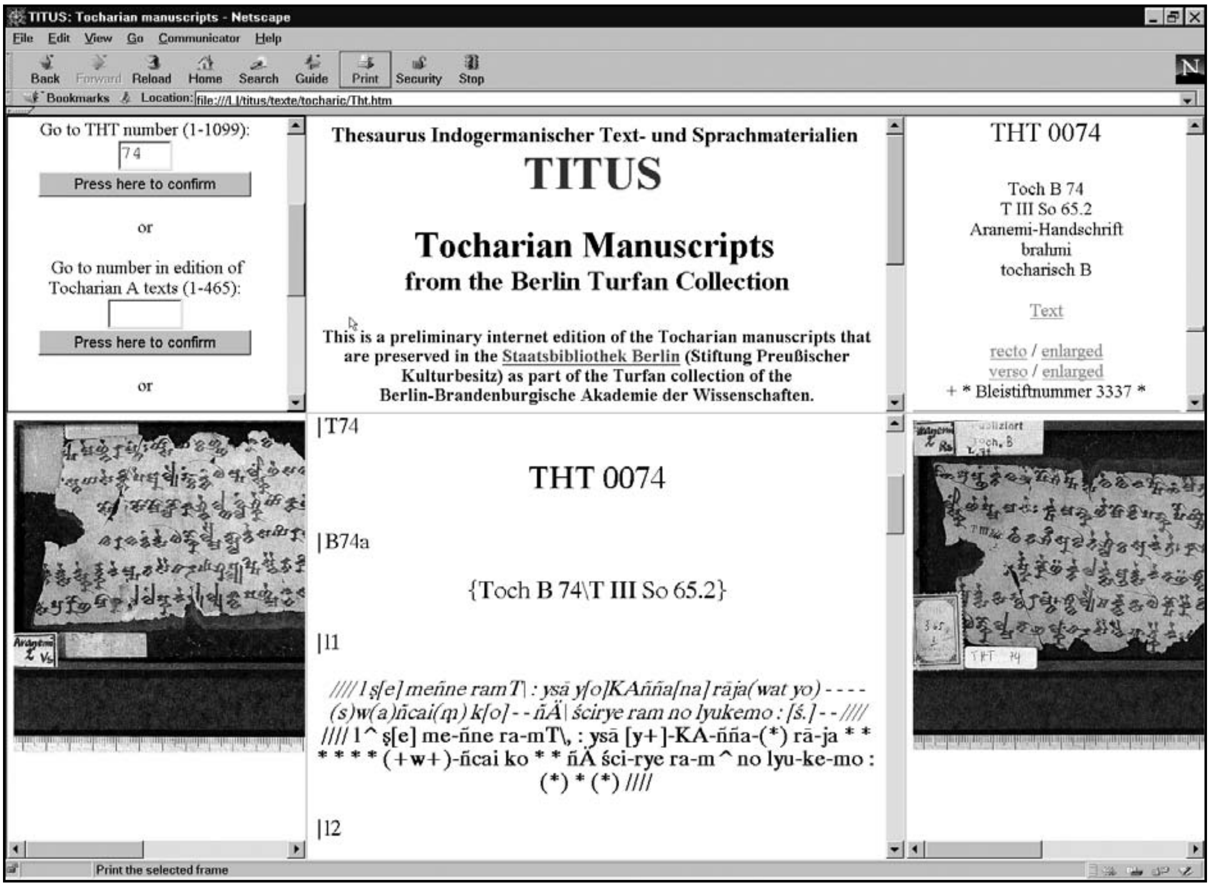
THT 0001 A

Toch B 1
T III So 96
Udanalankara
brahmi
tocharisch B

Figure 3: WWW start page of hypertext documentation

Several aspects of data preservation and and processing have to be considered when a WWW edition of this type is the goal. 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 large 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 graphic images. For high resolution images that are meant to be integrated into a WWW publication, this disadvantage can be ignored because the loss of information brought about by a compression rate of less than 15 percent will hardly be perceptible when the image is displayed on a computer screen or printed out on a standard laser printer. Anyway, the capacity of server disks may remain a decisive factor in selecting a resolution and the time necessary for data transfer of large files should also be taken into account. That said, 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 those with 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


TITUS: Tocharian manuscripts - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location:
Print the selected frame

file:///L|/titus/texte/tocharic/Tht.htm

Go to THT number (1-1099):
74
Press here to confirm

or

Go to number in edition of Tocharian A texts (1-465):

Press here to confirm

or

Thesaurus Indogermanischer Text- und Sprachmaterialien

TITUS

Tocharian Manuscripts
from the Berlin Turfan Collection

This is a preliminary internet edition of the Tocharian manuscripts that are preserved in the Staatsbibliothek Berlin (Stiftung Preußischer Kulturbesitz) as part of the Turfan collection of the Berlin-Brandenburgische Akademie der Wissenschaften.

THT 0074

Toch B 74
T III So 65.2
Aranemi-Handschrift
brahmi
tocharisch B

Text

recto / enlarged
verso / enlarged
+ * Bleistiftnummer 3337 *

|T74

THT 0074

|B74a

{Toch B 74\T III So 65.2}

|l1

*//// l ṣ[e] meñne ramT\ : ysā y[o]KAñña[na] rāja(wat yo) - - - -
(s)w(a)ñcai(ṃ) k[o] - - ñÄ\ ściryе ram no lyukemo : [ś.] - - ////*
//// l^ ṣ[e] me-ñne ra-mT\, : ysā [y+]-KA-ñña-(*) rā-ja * *
* * * * (+w+)-ñcai ko * * ñÄ ści-rye ra-m^ no lyu-ke-mo :
(*) * (*) ////

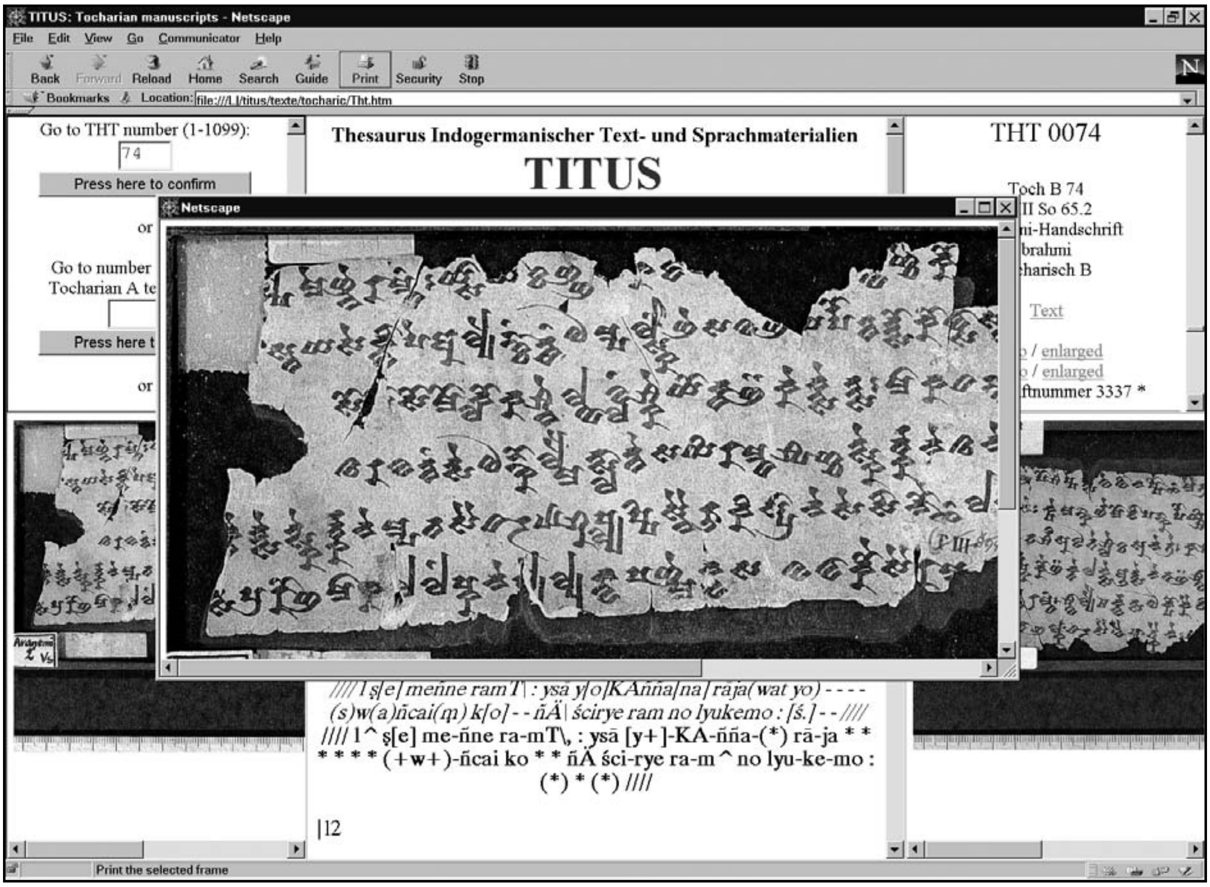
|l2

Figure 4: Selection and output of THT 74rv

number of files readily retrievable consists involves 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 information about the manuscript items or parts thereof 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].

Regarding the data contained in the catalogue, the first task is to convert them 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 must bear in mind, however, that the resulting HTML structure may not be as open to retrieval of information as the underlying database itself, e.g. if it is to be used for a search of manuscript provenances, for an arrangement according to

[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.


TITUS: Tocharian manuscripts - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: file:///L|/titus/texte/tocharic/Tht.htm
Press here to confirm
Press here t
Netscape
Print the selected frame

Go to THT number (1-1099):

74

or

Go to number
Tocharian A te

or

Thesaurus Indogermanischer Text- und Sprachmaterialien

TITUS

THT 0074

Toch B 74
II So 65.2
ni-Handschrift
brahmi
charisch B

Text

o / enlarged
o / enlarged
ftnummer 3337 *

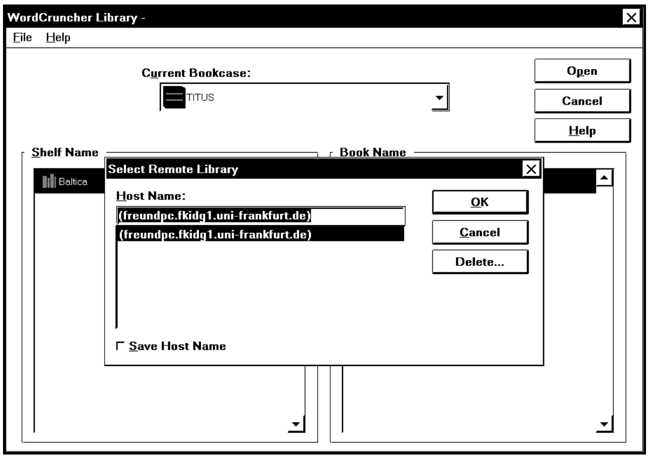
////1 ṣ[e] meñne ramT\ : ysā y[o]KAñña[na] rāja(wat yo) - - - -
(s)w(a)ñcai(ṃ) k[o] - - ñÄ\ ściryе ram no lyukemo : [ś.] - - ////
////1^ ṣ[e] me-ñne ra-mT\, : ysā [y+]-KA-ñña-(*) rā-ja * *
* * * * (+w+)-ñcai ko * * ñÄ ści-rye ra-m^ no lyu-ke-mo :
(*) * (*) ////

|12

Figure 5: Switching between frames and extra windows

the size of manuscripts, or the like. This is a problem of browser technology rather than the HTML language. Today’s net browsers 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 for, 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 find not only 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).

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 rather to facilitate (linguistic or philological) analysis, the search function of browsers will hardly be 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, indeed, it hardly seems possible to find an "intelligent" search engine for linguistic units in an HTML



Current Bookcase: TITUS

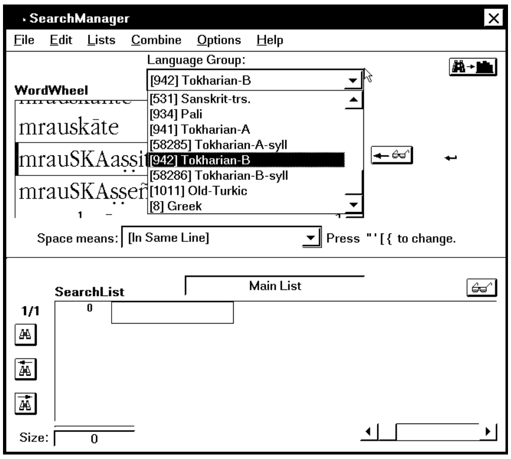
Shelf Name: Baltica

Host Name: (freundpc.fkidg1.uni-frankfurt.de)

(freundpc.fkidg1.uni-frankfurt.de)

Figure 6: WordCruncher remote library access

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 local machines and from remote servers (cp. Figure 6). Given that linkage of textual and graphic data is also possible in the preindexed texts (cf. Figure 8 and Figure 9), the WordCruncher system seems to be a worthwhile alternative to using HTML structures[11]. It has one shortcoming, however, in that it is still restricted to usage on MS-Windows based computers[12].



Language Group: [942] Tokharian-B

[531] Sanskrit-trs.
[934] Pali
[941] Tokharian-A
[58285] Tokharian-A-syll
[942] Tokharian-B
[58286] Tokharian-B-syll
[1011] Old-Turkic
[8] Greek

mrauskāte
mrauSKAassi
mrauSKAssef

Space means: [In Same Line] Press "'[{ to change.

0

Size: 0

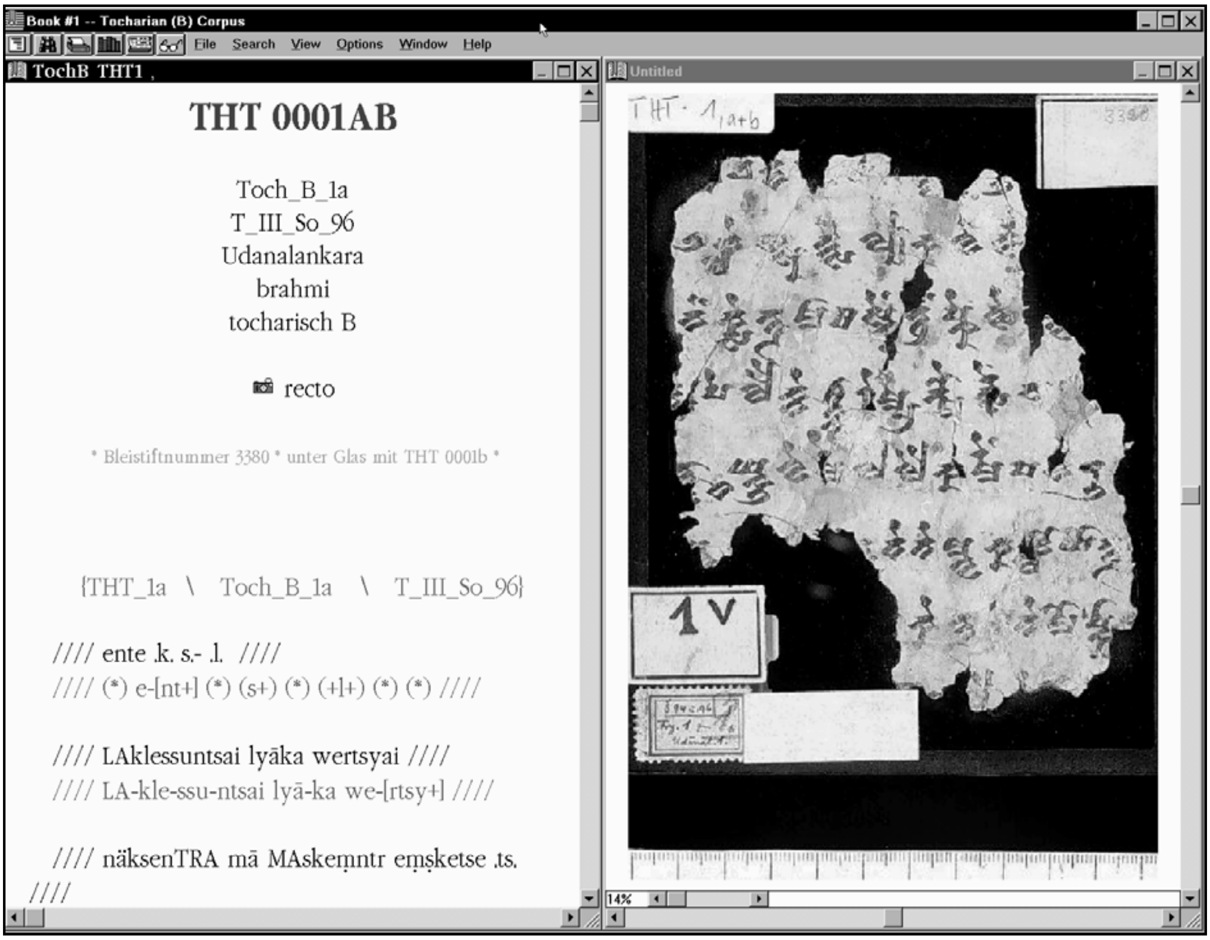
Figure 7: Word wheel choice

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, no encoding standard whatever is available, 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 is exacerbated when different scripts are to be mixed within a text document to be displayed and/or analyzed. A solution which is only just developing is the so-called "Unicode" standard which in its present

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

[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.



THT 0001AB

Toch_B_1a
T_III_So_96
Udanalankara
brahmi
tocharisch B

recto

* Bleistiftnummer 3380 * unter Glas mit THT 0001b *

{THT_1a \ Toch_B_1a \ T_III_So_96}

//// ente .k. s.- .l. ////
//// (*) e-[nt+] (*) (s+) (*) (+l+) (*) (*) ////

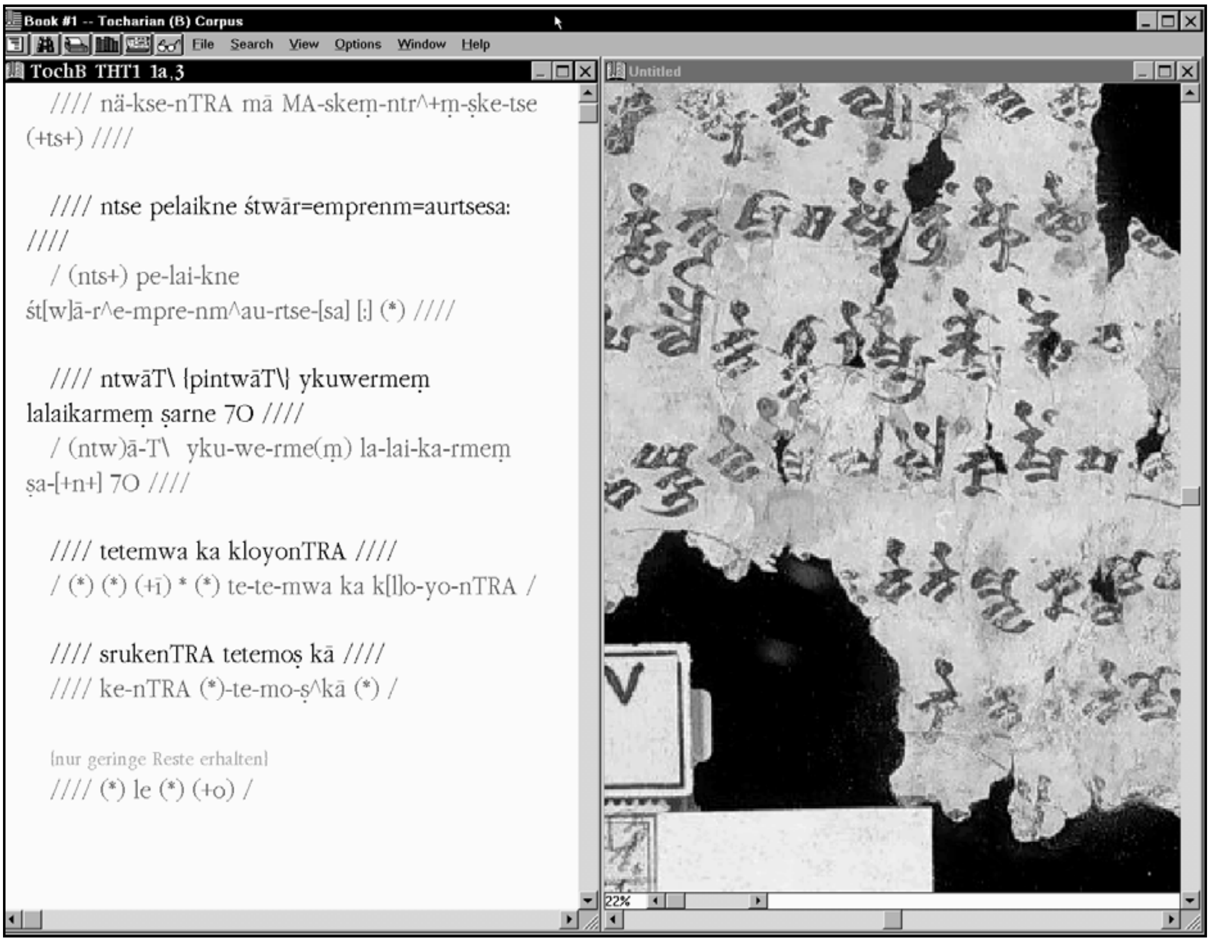
//// LAklessuntsai lyāka wertsyai ////
//// LA-kle-ssu-ntsai lyā-ka we-[rtsy+] ////

//// näksenTRA mā MAskeṃntr eṃṣketse .ts.
////

Figure 8: Representation of textual and graphic data in WordCruncher

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.

[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 information and tools.



//// nä-kse-nTRA mā MA-skeṃ-ntr^+ṃ-ṣke-tse (+ts+) ////

//// ntse pelaikne śtwār=emprenm=aurtsesa: ////
/ (nts+) pe-lai-kne śt[w]ā-r^e-mpre-nm^au-rtse-[sa] [:] (*) ////

//// ntwāT\ {pintwāT\} ykuwermeṃ lalaikarmeṃ ṣarne 7O ////
/ (ntw)ā-T\ yku-we-rme(ṃ) la-lai-ka-rmeṃ ṣa-[+n+] 7O ////

//// tetemwa ka kloyonTRA ////
/ (*) (*) (+ī) * (*) te-te-mwa ka k[l]o-yo-nTRA /

//// srukenTRA tetemoṣ kā ////
//// ke-nTRA (*)-te-mo-ṣ^kā (*) /

{nur geringe Reste erhalten}
//// (*) le (*) (+o) /

Figure 9: Representation of textual and graphic data in WordCruncher