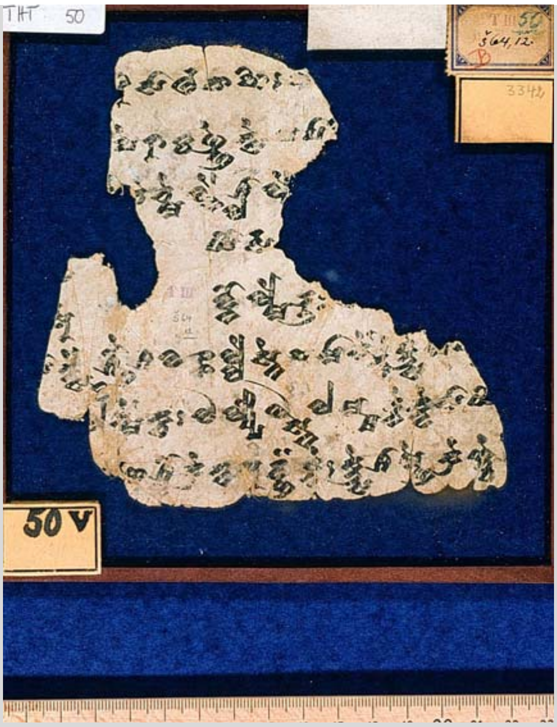


Plate 1

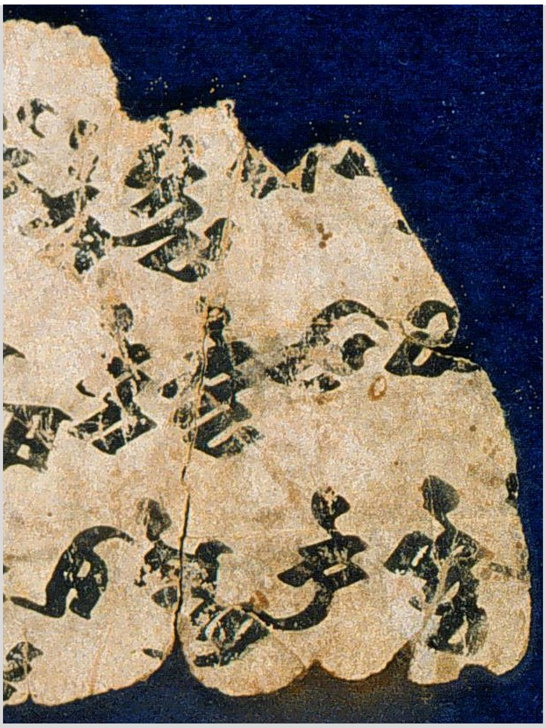


Plate 2

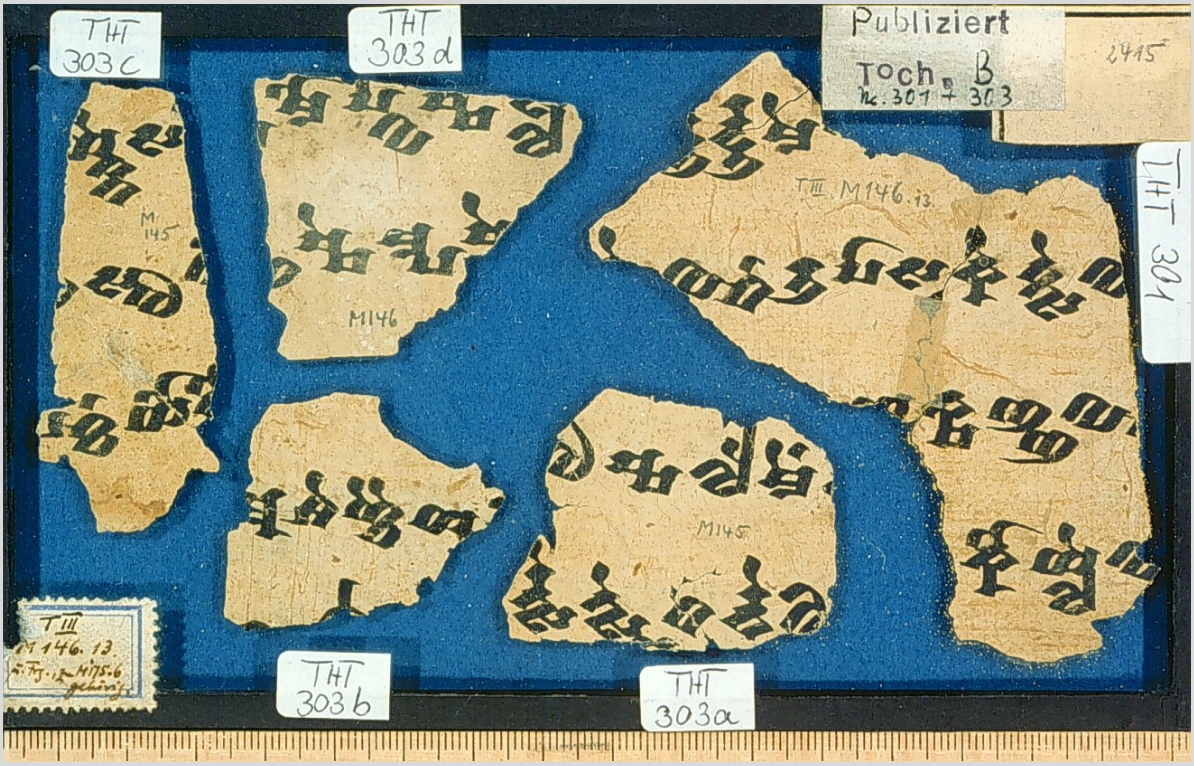


Plate 3

J. Gippert

DIGITIZATION OF TOCHARIAN MANUSCRIPTS FROM THE BERLIN TURFAN COLLECTION

The starting point of the project to be reported about here [1] was the conference dedicated to the fulfilment of "100 Years of Tocharian Studies" which took place in Saarbrücken (Germany) on 13–15 October, 1995 [2]. Within a panel session, the participants of the conference discussed the necessity of digitizing the Tocharian manuscripts that are preserved in several European museums with a view to two major aims. One of these consists in preserving the data the manuscripts contain for eternity. This is an aim of high priority, at least as far as the Berlin collection is concerned, because here, many manuscripts, albeit preserved in glass frames, have suffered great damages during World War II (when the collection had to be evacuated, and there are hardly any means of protecting them from further erosion.

The second aim consists in making the contents of the manuscripts more easily accessible to the scholarly world. This is a high priority aim as well, given that studies concerning the spread of Buddhist thought along the Silk Road are facing steadily increasing interest these days, digitization of original documents playing an important role [3].

As a result of the Saarbrücken discussions, the digitization of Tocharian manuscripts as preserved in the Stiftung Preußischer Kulturbesitz in Berlin (ca. 4,300 items) has meanwhile begun (since autumn 1996). At present, work is proceeding in a joint effort by Berlin-Brandenburgische Akademie der Wissenschaften and Berlin Staatsbibliothek, Institut für Vergleichende Sprachwissenschaft of Frankfurt University, and Tamai Foundation. Means and procedures as developed for the running project will be briefly demonstrated here using the Berlin Tocharian manuscript THT fol. 50r and others as examples [4].

The first task consists in photographing of the manuscripts. For the time being, this is being done using high-resolution colour slide films [5] because "classical" photographing still has several advantages as against using digital cameras. It still yields much better results with respect to orthochromaticity and resolution, and slides can be stored as multi-purpose reference copies of the documents. Not depending on the availability of digital equipments, they are easily at hand for copying, presentation, etc.

In the course of the developing project, several attempts have been made as to finding the most suitable background for the photographing. It turned out that using a bright-coloured paper (white or grey) must be preferred as against any dark-coloured (dark blue or black) background. The reason is that the Berlin manuscripts are stored in glass frames throughout and must not be taken out because this would lead to damages in many cases. When photographed through a glass frame, however, the manuscripts, lit from above, cause some shading so that writing elements on their edges may become hardly distinguishable from a dark background, especially where edges are damaged. Cf. *Plates 1* and *2* (see p. 49) where this effect is demonstrated with ms. THT fol. 50r, on a dark blue background; also *Plate 3* (see p. 49) and *Plate 4* (see p. 52) where two pictures of the fragments THT 301 and 303 are contrasted with different backgrounds, and *Plate 5* (see p. 52) where a greyish background showing the effect of shading is used with ms. THT fol. 508r.

When photographing the manuscripts in their frames, some further problems have been encountered. One of them consists in the labels that are usually fixed on the frames and which may sometimes cause a loss of readability, covering parts of the manuscripts, as in THT fol. 508r (cf. *Plate 5* on p. 52). A similar problem may arise when the glass frame is broken (cf., for example, *fig. 1*, manuscript THT fol. 252v, on a dark blue background). In these latter cases, a restoration of the frame may be inevitable. In every case, a ruler should be added to the item being photographed in order for its original measurements to remain calculable.

The digitization of the colour slides thus produced requires a special high-resolution colour slide scanner with a scanning resolution of at least 2,500 dpi (slide adapters that can be fixed to flat bed scanners do not yield a sufficient resolution) [6]. For the purposes of the present project, scanning is being done in at least two steps.

The first step consists in a total scan of the picture, comprising the manuscript within the complete glass frame and the ruler for measuring. Doing this at a medium resolution of 1,000 to 1,300 dpi, this yields digital images that fill

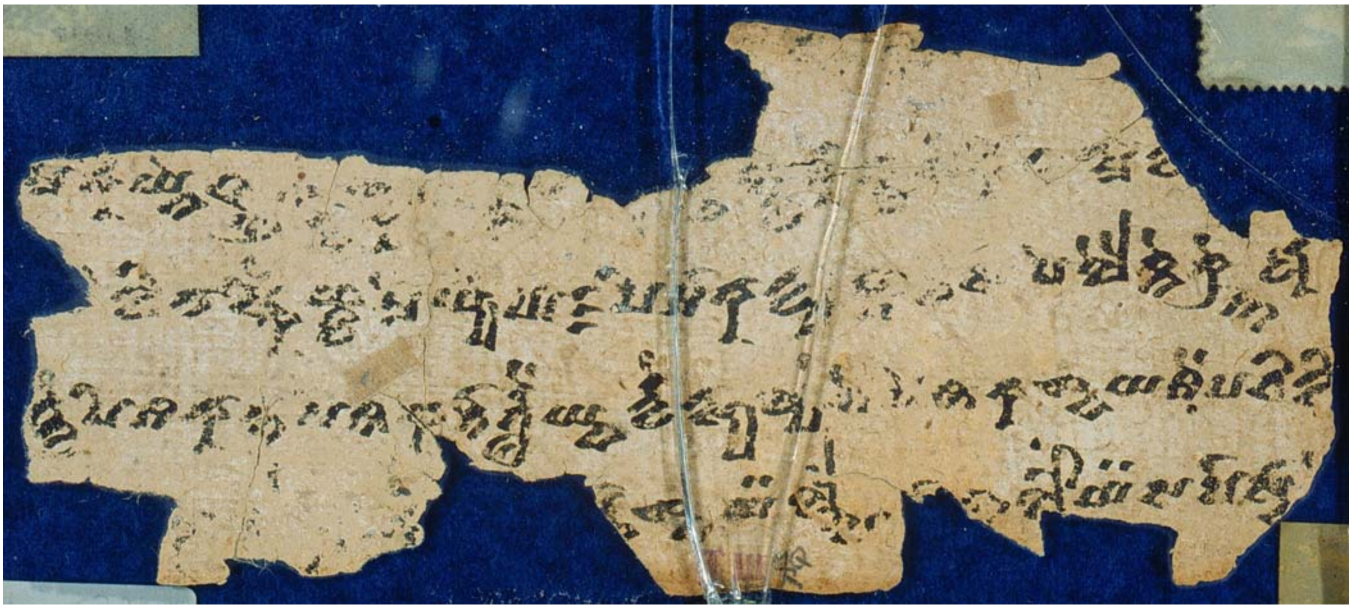


Fig. 1

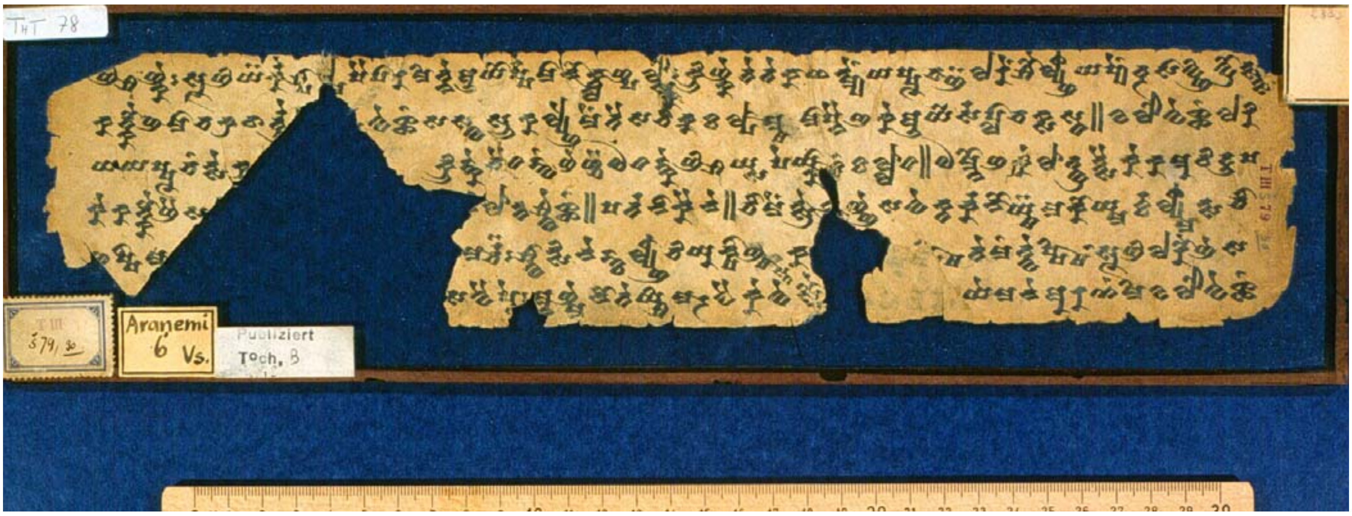


Fig. 2

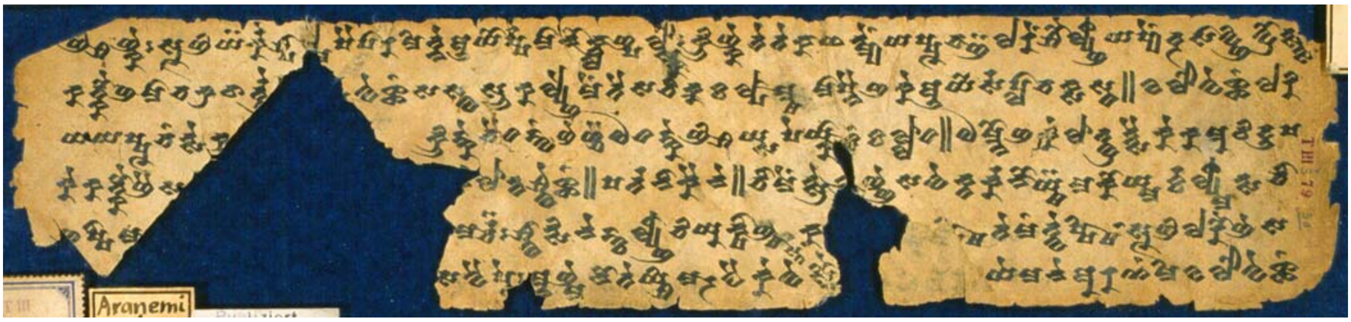


Fig. 3

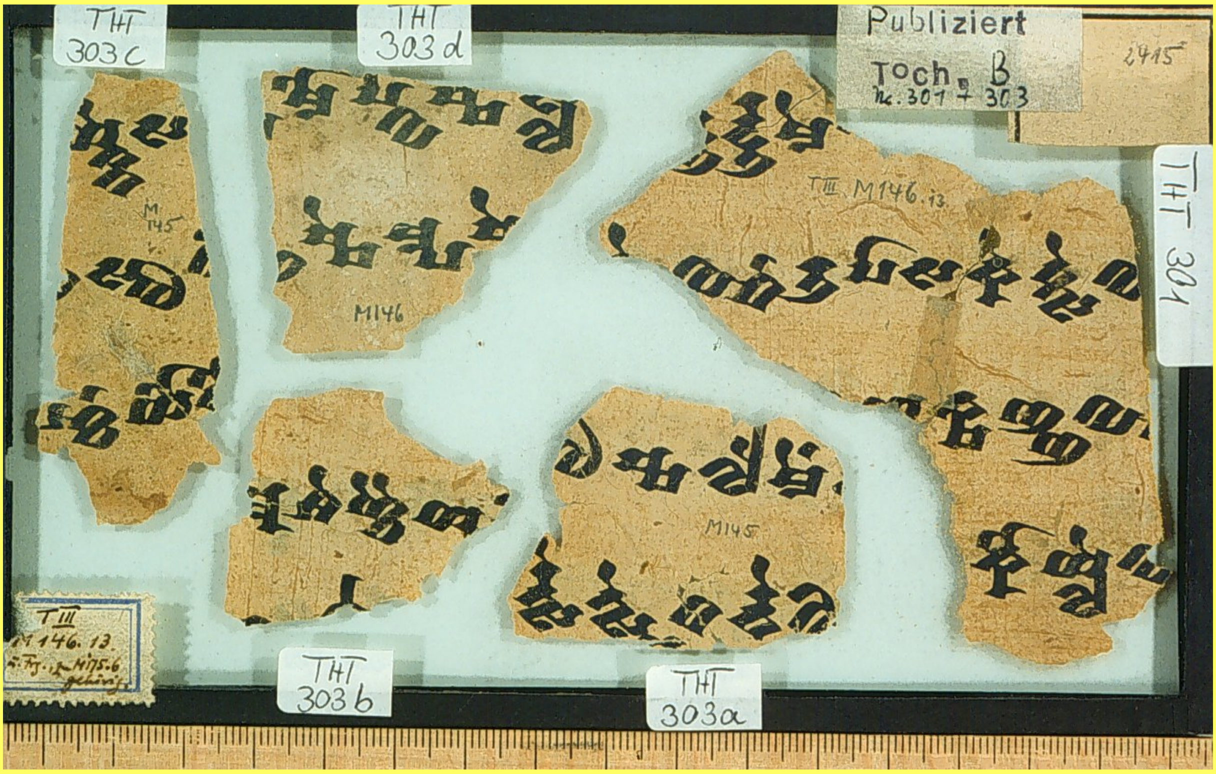


Plate 4

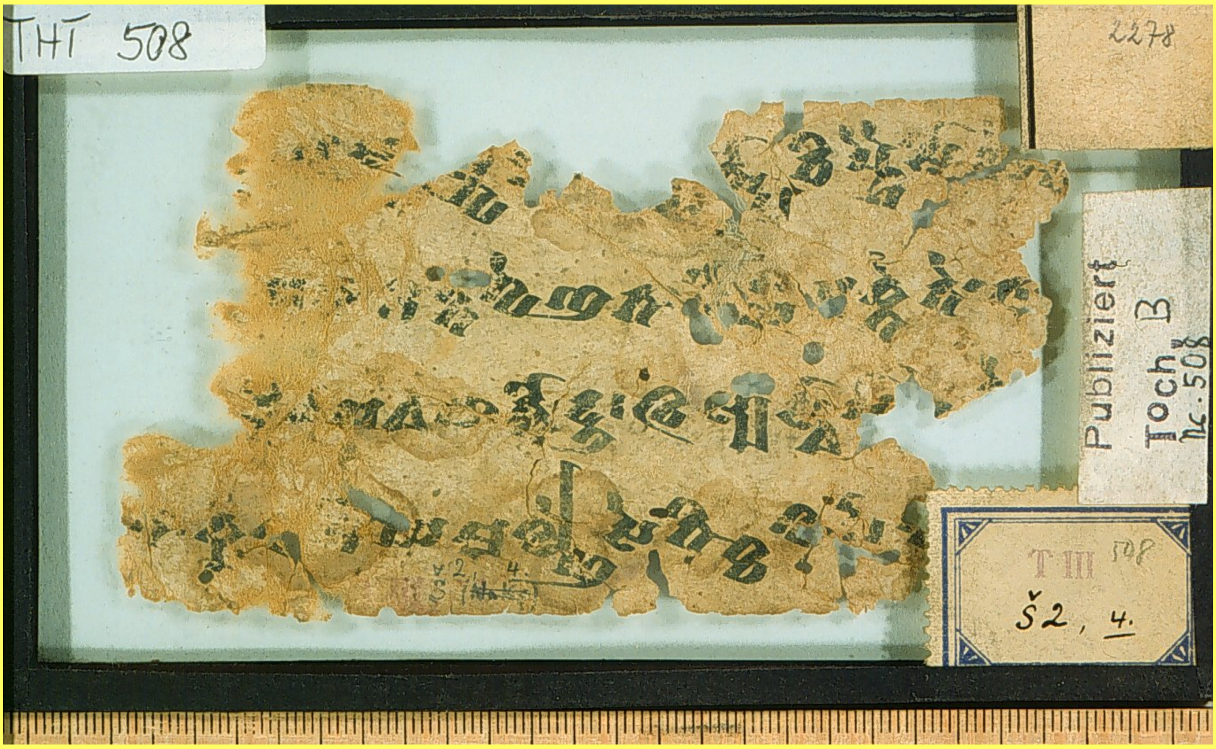


Plate 5

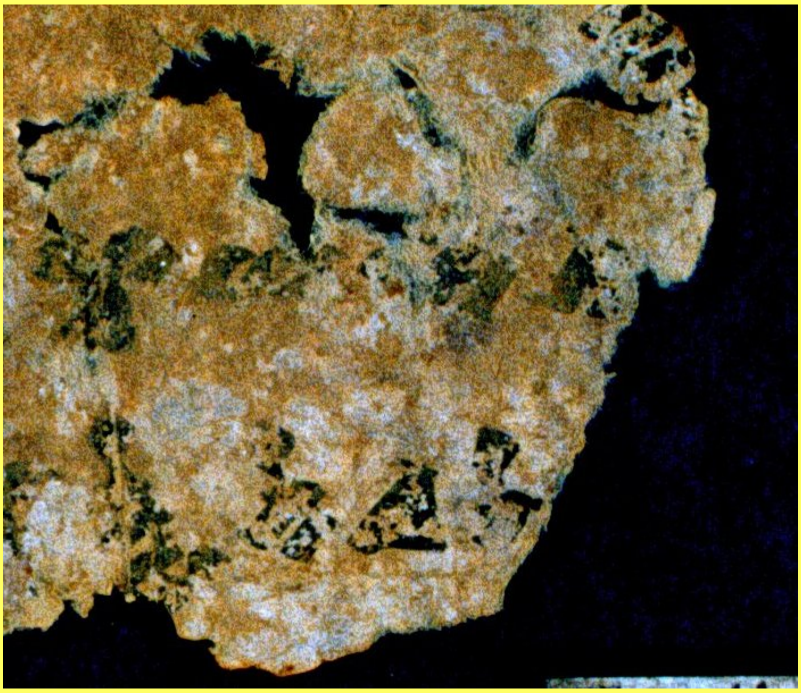


Plate 6

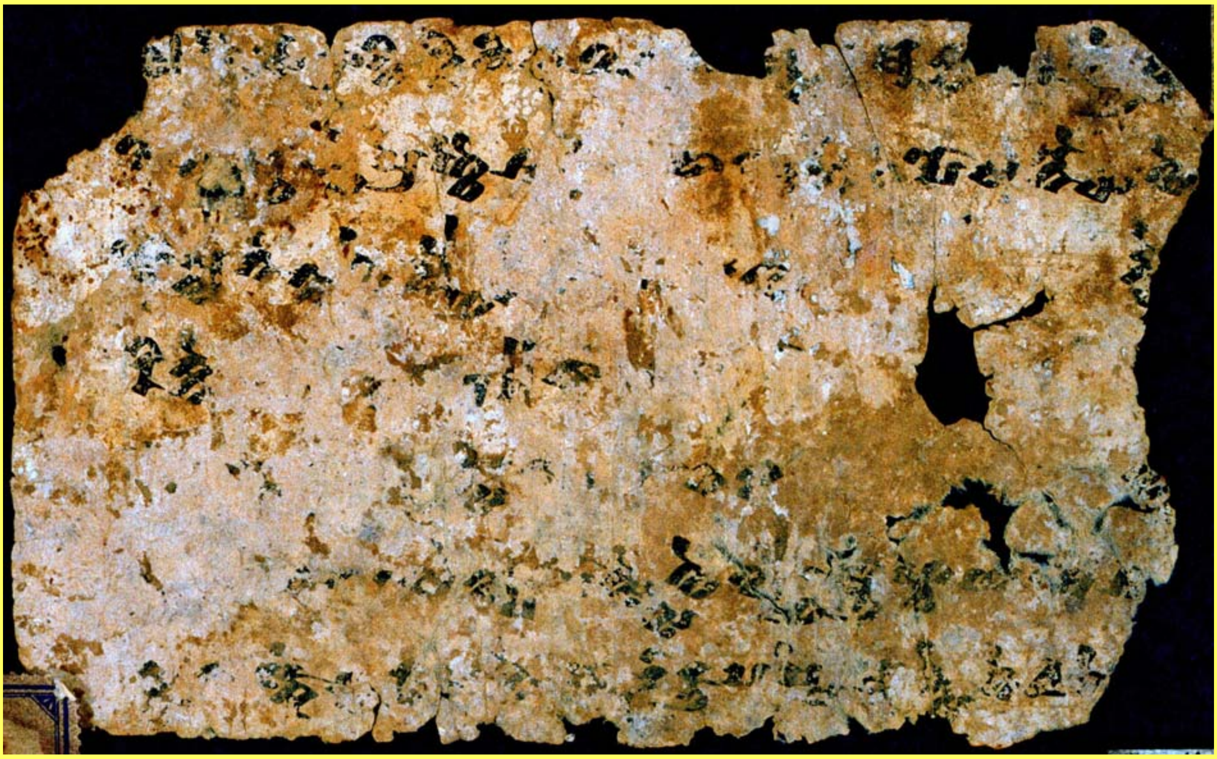


Plate 7

a normal (high resolution) computer screen (at an average size of 1200 by 800 pixels). The images thus produced will normally suffice for a reading of the manuscript contents.

Another scanning procedure, with a much higher resolution, is necessary with a view to preserving a maximum information for "eternal" storage as well as for high resolution printing. The calculation of what resolution is necessary depends on several factors. Starting from the size of the colour slides (36 × 24 mm), a resolution of 2,400 dpi would theoretically yield a printed picture of double size (72 × 48 mm) if it could be printed with the same density (i.e., without any loss of information) with a high quality printer using 1,200 dpi, and in a picture of 144 × 72 mm if printed with 600 dpi. The actual results to be achieved with today's laser printers are quite different from that, however, because here, normally, three or four pixels are assembled together to represent grey tones which results in the typical dot rastering.

The most important factor when calculating the required scanning resolution is the actual size of the original document. If a manuscript of 360 mm length, comprised in a colour slide of 36 mm, were to be printed in its actual size at a resolution of 1,200 dpi without any rastering, a resolution of (1,200 dpi × 10 =) 12,000 dpi would be required. Scanned files of such a resolution could hardly be handled and stored, however, considering that they would extend to more than 300 MB each (if representing a complete colour slide). This is why for the present project, a maximum scanning resolution of 2,700 dpi was accepted as sufficient for the majority of pictures. Only in rare cases where the size of the original document exceeds 40 cm in length or 25 cm in height, a higher resolution would be preferable (cp. THT 78, *figs. 2* and *3*, as an example); as this cannot be achieved with the present equipment, a practical solution consists in separate photographing of parts of the document.

Even with a maximum resolution of 2,700 dpi, file sizes may extend up to 26 MB if complete slides are scanned with full colour density and the resulting images are saved in plain BMP or PCX format. In order to decrease the necessary storage capacity (a standard CD-ROM with 650 MB could only comprise some 25 image files of this size), only the actual manuscripts are scanned in this way, omitting the surrounding parts of the glass frame, labels and rulers. The file size can further be sharply reduced (to about 15%) by applying data compression methods such as the ones provided by JPG formats. In this case, a minor loss of information has to be accepted which will hardly be recognizable on the screen or a printout though. With a view to eternal preservation of the manuscript contents, the development of loss-free compression methods remains a desideratum. There is one other feature, however, that should always be kept in mind when choosing file formats: Given the fast development of operating systems and peripheral equipments, we should care for the data to be stored in a format that is probable to remain convertible into future formats for a sufficient period of time. The same holds true, of course, for the storage media we now use.

Another way of decreasing the amount of storage capacity would consist in choosing a lower number of colours to be represented, i.e., 256 instead of 16 million colours, or in storing the files in greyscale format. While this would hardly have any influence on a later greyscale printout, it would mean a considerable loss of information comparing the digitized image with the colour slide it is based on. With a view to data preservation, reducing the number of colours will always mean a deterioration of quality so that it is not recommendable.

On the other hand, the digitized images can be enhanced in many ways. This begins with the choice of settings for the scanning procedure. An up-to-date scanning software will allow for a large scale of settings to be applied to each scan. Whenever written manuscripts are the object of digitizing, a maximum of contrast between script elements and their (paper or parchment) background will be required. This can be be achieved by choosing a maximum of "sharpening" of contours, maybe in connection with a smoothing of contourless elements of the scanned area ("sharpen detail, smooth noise"). A sharpening function can further be applied after scanning, i.e., when handling the digitized image in a photo editing software. The results acquired in this way will be visible on the screen as well as in a printout; as they tend towards a deformation of the original visual appearance of the object, they should not be exaggerated with a view to data preservation, however.

Many of the Turfan manuscripts require a special treatment because their contents are hardly readable due to damages or fainting of the ink. Such treatment may consist in additional contrast enhancing, colour splitting and the like, to be applied during or after scanning. Although this may indeed enhance the readability (cp. *Plates 6* and *7* on p. 53, both part of THT, fol. 170r), it leads to an even stronger deformation of the represented object and should therefore not be applied to an image which is meant to document the actual state of the manuscript. It should also be noted that a digitized image can never contain more information than the colour slide it relies upon; in special cases, it may be preferable to look for enhanced methods of photographing. Unfortunately, the reading of the badly preserved manuscripts of the Turfan collection cannot be supported by using ultraviolet or infrared films [7].

Manuscript digitizing has another aspect which is rather interpretative than practical: The data the written manuscripts contain consist of textual elements that require a special treatment if they are to be analyzed with digital methods. The scanning of a photographed document yields a digital representation which is structured in just the same way as, e.g., a digitized image of a person's portrait or a still life: it consists of nothing but dots ("pixels") with a certain colour and brightness information. Although it is possible to a certain extent to analyze such an image with respect to the elements (i.e., groups of pixels) it contains, thus filtering letter shapes from a background (this is what a so-called OCR, "Optical Character Recognition", software does), this will not normally work with old manuscripts where letters are broken, damaged, or otherwise hardly distinguishable even for a human reader. This is why entering the texts that the Turfan manuscripts contain must still be done manually, all the more since the reading of a given text passage or element requires scholarly experiences and skills.

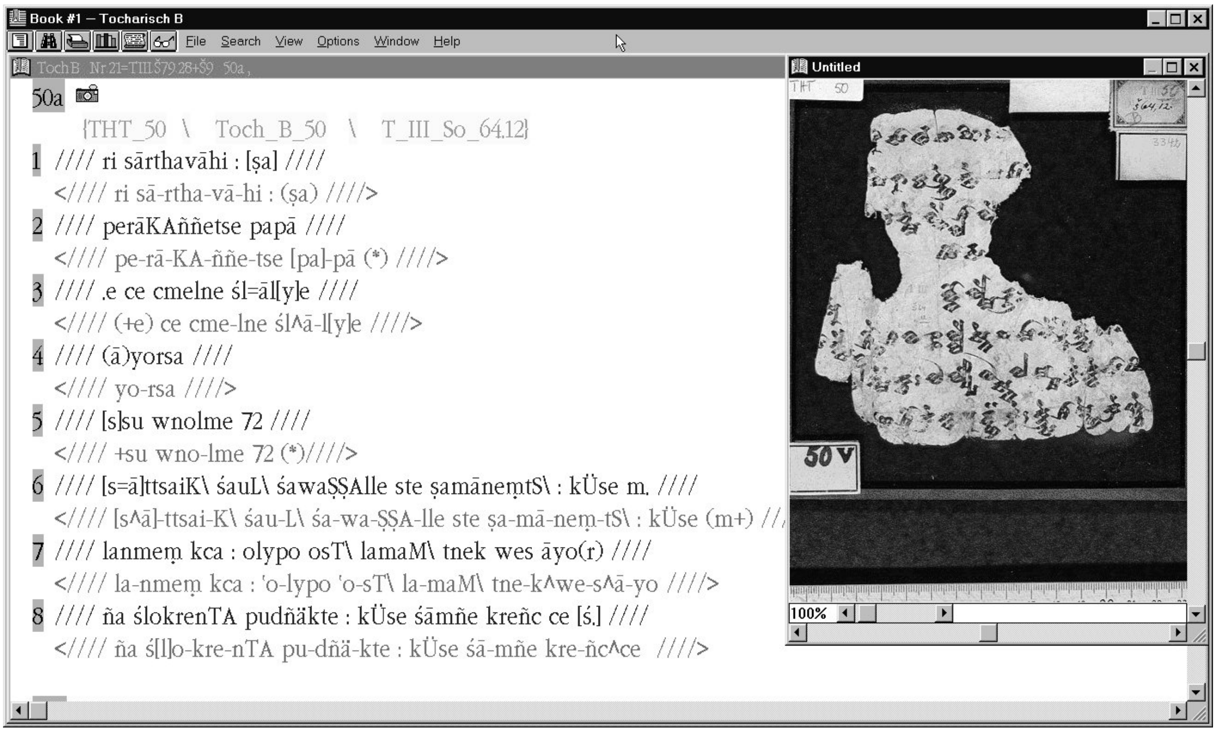
For the entering of plain text, the question of which format to choose applies as well. This is a question of data structure rather than a question of representation, given that handling of transliterational or transcriptional and even original scripts becomes more and more easy on nowadays' computers. The data structure depends on what scholarly analysis the material has to be prepared for. Normally, the digitization of manuscript contents will form part of an

Table 1

```
|c50a HI{icon}{0050rt.bmp}1
      Tvod16 {THT_50 \ Toch_B_50 \ T_III_So_64.12} Tn16
|p1   Tcotb16 //// ri sārthavāhi : [ṣa] //// Tn16
      Tcotbx16 <//// ri sā-rtha-vā-hi : (ṣa) ////> Tn16
|p2   Tcotb16 //// perāKAññetse papā //// Tn16
      Tcotbx16 <//// pe-rā-KA-ññe-tse [pa]-pā (*) ////> Tn16
|p3   Tcotb16 //// .e ce cmelne śl=āl[y]e //// Tn16
      Tcotbx16 <//// (+e) ce cme-lne śl^ā-l[y]e ////> Tn16
|p4   Tcotb16 //// (ā)yorsa //// Tn16
      Tcotbx16 <//// yo-rsa ////> Tn16
|p5   Tcotb16 //// [s]su wnolme 72 //// Tn16
      Tcotbx16 <//// +su wno-lme 72 (*)////> Tn16
|p6   Tcotb16 //// [s=ā]ttsaiK\ śauL\ śawaṢṢAlle ste ṣamāneṃtS\ : kÜse m. //// Tn16
      Tcotbx16 <//// [s^ā]-ttsai-K\ śau-L\ śa-wa-ṢṢA-lle ste ṣa-mā-neṃ-tS\ : kÜse (m+) ////> Tn16
|p7   Tcotb16 //// lanmeṃ kca : olypo osT\ lamaM\ tnek wes āyo(r) //// Tn16
      Tcotbx16 <//// la-nmeṃ kca : `o-lypo `o-sT\ la-maM\ tne-k^we-s^ā-yo ////> Tn16
|p8   Tcotb16 //// ña ślokrenTA pudñākte : kÜse śāmñe kreñc ce [ś.] //// Tn16
      Tcotbx16 <//// ña ś[l]o-kre-nTA pu-dñā-kte : kÜse śā-mñe kre-ñc^ce ////> Tn16
```

editorial process, aiming at either a critical or a diplomatic edition of a text. Having a digitized text at hand, however, several other aims may be envisaged that go far beyond preparing a normal printed edition. For the present project, one such aim consists in establishing the relationships between the Tocharian texts and other branches of the Buddhist tradition (Sanskrit, Pali, Chinese, Tibetan, etc.). This presupposes the existence of digitized text corpora that can be aligned cross-linguistically. As a common basis for such an investigation, the "Wordcruncher" text retrieval program as developed by Brigham Young University [8] has proved to be well suited [9].

Another aim concerns investigations into the palaeography of the Brahmi script as used by the Tocharians. For this purpose, it is necessary to prepare the texts in a way that allows for an indexation of separate elements, i.e. *akṣaras*. Using the Wordcruncher retrieval system, this can be done as indicated in *Table 1* where THT fol. 50r is taken as an example again. Here, a "broader" transcription and a one-to-one-transliteration of each manuscript line are arranged interlinearily, the former one being indexable as to word forms, the latter one as to *akṣaras* (which are separated either by spaces or by hyphens in the transliteration). As can be seen in the sample, several additional signs have



```
50a
    {THT_50 \ Toch_B_50 \ T_III_So_64.12}
1 //// ri sārthavāhi : [ṣa] ////
  <//// ri sā-rtha-vā-hi : (ṣa) ////>
2 //// perāKAññetse papā ////
  <//// pe-rā-KA-ññe-tse [pa]-pā (*) ////>
3 //// .e ce cmelne śl=āl[y]e ////
  <//// (+e) ce cme-lne śl^ā-l[y]e ////>
4 //// (ā)yorsa ////
  <//// yo-rsa ////>
5 //// [s]su wnolme 72 ////
  <//// +su wno-lme 72 (*)////>
6 //// [s=ā]ttsaiK\ śauL\ śawaṢṢAlle ste ṣamāneṃtS\ : kÜse m. ////
  <//// [s^ā]-ttsai-K\ śau-L\ śa-wa-ṢṢA-lle ste ṣa-mā-neṃ-tS\ : kÜse (m+) //
7 //// lanmeṃ kca : olypo osT\ lamaM\ tnek wes āyo(r) ////
  <//// la-nmeṃ kca : 'o-lypo 'o-sT\ la-maM\ tne-k^we-s^ā-yo ////>
8 //// ña ślokrenTA pudñākte : kÜse śāmñe kreñc ce [ś.] ////
  <//// ña ś[l]o-kre-nTA pu-dñā-kte : kÜse śā-mñe kre-ñc^ce  ////>
```

Fig. 4

been chosen for a marking of missing elements, word boundaries within *akṣaras*, and the like; and as against the traditional transcription, the so-called “Fremdzeichen” are transcribed as majuscules (e.g., KA instead of *ka*). This, however, is for the practical purposes of the given task only; a conversion into other transcriptional systems remains possible as long as there is a unique relationship between each graphical item and its encoding. For the screen output of the Wordcruncher text (Windows version) cf. *fig. 4* where the digitized image of THT fol. 50r is linked to the rendering of the text.

In the course of the present project, the contents of the manuscripts are being typed in successively as the digitization of the slides proceeds, and by now (February 1998), about one fourth of the Berlin materials has been prepared [10]. In the near future, we hope to be able to make the results available to the scholarly world in at least two different ways: Both on CD-ROMs (which are already being used for the storage of data) and via an internet server. For the latter purpose, the Tocharian data have been embedded in a larger project which I have been running for about ten years now. This is the project of a “Thesaurus of Indo-European Textual and Linguistic Materials” (“TITUS”) the basic part of which is a steadily increasing text data base that aims at preparing all textual materials as relevant for Indo-European studies (Old Indic, Old and Middle Iranian, Anatolian, Old Germanic, Old Celtic, Italic, etc.) in electronic form for linguistic and literary analyses [11], the Tocharian texts forming one of its outstanding parts. The TITUS text data base is now being prepared for an immediate internet retrieval (on the basis of a “Wordcruncher server”). It is to be hoped that other collections of Tocharian manuscripts will be made available in a similar way soon.

Notes

1. A first short account of the project has been published in *Tocharian and Indo-European Studies*, VII (1997), pp. 265f. It is also available on the WWW page http://titus.uni-frankfurt.de/texte/tocharic. A preliminary version of the present report was read on the 35th ICANAS at Budapest, July 9th, 1997.

2. The conference proceedings have been published in *Tocharian and Indo-European Studies*, VII (1997).

3. Several projects concerning Buddhist documents have been initiated in recent times; for a survey cf. the WWW site of the Buddhist Text Encoding Initiative, http://www.iijnet.or.jp:80/iriz/irizhtml/ebti/ebtie.htm.

4. In the Berlin collection, manuscripts pages had originally been marked according to German usage where *V* stands for “Vorderseite”, i.e., recto, and *R* for “Rückseite”, i.e., verso. Comparing the usual abbreviations for the Latin equivalents, viz. *r* for recto and *v* for verso, this may bring about some confusion whenever the older designations are used.

5. Best results were obtained with a Kodak Ektachrome Professional film (EPY 5018).

6. Within the present project, a Polaroid SprintScan 32 Plus with a maximum resolution of 2700 dpi is being used.

7. This is due to the ink they are written with as was stated by the director of the Oriental department of the Berlin Staatsbibliothek, H.O. Feistel.

8. Cf. the WWW page http://www.wordcruncher.com for details.

9. For a first-hand description cf. my contribution to the Saarbrücken conference, in *Tocharian and Indo-European Studies*, VII (1997), 17–34. A detailed account of the requirements of a multi-lingual text retrieval will be published in the proceedings of the 2nd Conference on Language, Logic, and Computation, Tbilisi 1998 (with examples taken from the Christian traditions of the Near East).

10. The texts as available via former printed editions were first entered in a raw format by P. Olivier; the B-Tocharian texts were then corrected by Chr. Schaefer. The present arrangement is being prepared by T. Tamai in cooperation with K. T. Schmidt and myself.

11. For details, cp. the WWW pages http://titus.uni-frankfurt.de/texte/texte.htm and http://titus.uni-frankfurt.de/texte/titusldv.htm.

Illustrations

Plate 1. Tocharian manuscript THT 50r from the Stiftung Preußischer Kulturbesitz in Berlin, photographed on a dark blue background (see p. 49)

Plate 2. An enhanced fragment of the same manuscript folio, shown on a dark blue background, see p. 49.

Plate 3. Tocharian fragments THT 301 and 303 from the Stiftung Preußischer Kulturbesitz in Berlin,, contrasted with light blue background (see p. 49).

Plate 4. The same fragments, contrasted with greyish background (see p. 52).

Plate 5. Tocharian manuscript THT 508r, shown with the labels covering part of the text (see p. 52).

Plate 6. Tocharian manuscript THT 170r (fragment, enhanced), see p. 53.

Plate 7. The same manuscript, with contrast enhancing and colour splitting applied (see p. 53).

Fig 1. Tocharian manuscript THT 252v from the Stiftung Preußischer Kulturbesitz in Berlin, photographed with glass frame broken.

Fig. 2. Tocharian manuscript THT 78, shown with a resolution of 2,700 dpi.

Fig. 3. Tocharian manuscript THT 78, shown with a resolution higher than 2,700 dpi.