tranScriptorium

D2.2: Data Maintenance (M24)

Including

D2.1. Data Collection and Ground Truth Annotation (M9)

Günter Mühlberger, UIBK

Distribution: Public

---

tranScriptorium
ICT Project 600707
Deliverable 2.2 (January, 9th 2015)

Deliverable 2.1 (December, 8th 2013)

January 9th, 2014

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development

Project ref no.	ICT-600707
Project acronym	**tranScriptorium**
Project full title	**tranScriptorium**
Instrument	STREP
Thematic Priority	ICT-2011.8.2 ICT for access to cultural resources
Start date / duration	01 January 2013 / 36 Months

Distribution	Public
Contractual date of delivery	D2.2: December, 31, 2015
Actual date of delivery	D2.2: January 9th, 2015
Date of last update	
Deliverable number	D 2.2
Deliverable title	D2.2: Data Maintenance
Type	Report/Collected Data/Tool Development
Status & version	D2.2: Version 1
Number of pages	23
Contributing WP(s)	2
WP / Task responsible	Günter Mühlberger
Other contributors	INL, UCL, UPVLC
Internal reviewer	Joan Andreu Sánchez
Author(s)	Günter Mühlberger (editor), Katrien Depuydt (INL), Tim Causer (UCL), Joan Andreu Sánchez (UPVLC), Basilis Gatos (NCSR)
EC project officer	José María del Águila Gómez
Keywords	Reference data, Ground Truth, Handwritten Text Recognition

The partners in **tranScriptorium** are:

Universitat Politècnica de València - UPVLC (Spain)
University of Innsbruck - UIBK (Austria)
National Center for Scientific Research "Demokritos" - NCSR (Greece)
University College London - UCL (UK)
Institute for Dutch Lexicology - INL (Netherlands)
University London Computer Centre - ULCC (UK)

For copies of reports, updates on project activities and other **tranScriptorium** related information, contact:

The **tranScriptorium** Project Co-ordinator
Joan Andreu Sánchez,
Universitat Politècnica de València
Camí de Vera s/n. 46022 València, Spain
jandreu@dsic.upv.es
Phone (34) 96 387 7358 - (34) 699 348 523

Copies of reports and other material can also be accessed via the project's homepage: http://www.transcriptorium.eu/

Table of Contents

Executive Summary for D2.2 4
1. D2.2: Data Maintenance 4
1.1. Transcription & Recognition Platform (TRP) as Production System for GT 4
1.2. Standardized Workflow for GT Production 5
1.3. Lessons learned 7
2. Actual Status of GT Production 7
2.1. UCL – Bentham Collection 8
2.2. INL 9
2.3. UPVLC 10
2.4. UIBK 11
Executive Summary for D2.1. 14
1. Introduction 15
1.1. Data Collections and Ground Truth 15
1.2. tS Overall architecture for GT 21
2. Data collections for tS 26
2.1. Introduction and overview 26
2.2. Bentham 29
2.3. UPVLC 29
2.4. INL 32
2.5. UIBK 33
3. GT production 35
3.1. Manual GT Production: ALETHEIA 35
3.2. Semi-automated Ground Truth Production 36
4. Further aspects 38
2.5. Accessibility of GT 38
2.6. Competitions 38
2.7. More ground truth 38
3. References 40
4. Appendix 41
4.1. Model agreement for access to GT 41
Appendix: Basic concepts in tS 42

Executive Summary for D2.2

In the second year of Work package 2 "Data Collection and Ground Truth Annotation" we followed the plan set out in Task 2.1. and described in Deliverable 2.1. Data Collection and Ground Truth Annotation.

Two main activities were set:

(1) The infrastructure to produce ground truth was put into place and integrated into the Transcription and Recognition Platform developed within the project. Now a user is able to produce GT in a highly flexible and standardized way with the tools developed in the project. With the integration of the HTR engine into the platform we will follow this path and provide a comprehensive system which enables users to transcribe a given handwritten document in a way that it can be used within the Digital Humanities domain, as well as for training HTR engines.
(2) As set out in the overall work plan the actual production of Ground Truth was continued. A list of already produced GT will be provided in the last section of this report.

1. D2.2. Data Maintenance

1.1. Transcription & Recognition Platform (TRP) as Production System for GT

Within *Work Package 5 Integration* and *Work Package 6 User Interfaces* UIBK is currently developing a Transcription & Recognition Platform which integrates the tools developed in the project, mainly Document Image Analysis (DIA) tools and the Handwritten Text Recognition (HTR) engine.

Though it was not planned from the very beginning it became obvious during the course of the project that it makes sense to design this platform in a way that it will also meet the requirements for GT set out in Deliverable 2.1. In this way it was possible to avoid double effort and also to use the resources dedicated for GT production to test and evaluate the transcription GUI.

In this way the TRP Server is also the main repository for all GT documents in the project. This repository can be accessed via the Transcribus client and according to our work plan also via a dedicated web-site.

TRP comes as a Client/Server application. The server covers the following tasks:

- User and role management
- Business logic (database)
- File management (file server)
- Server based services, such as DIA, HTR, PDF Generation, TEI generation, METS generation, etc.
- REST services for communicating with clients

The Graphical User Interface (GUI) is a JAVA application, called Transcribus and comes with the following main features:

- Several viewing modes (to support professional users with several screens)

- Local mode for processing documents without server support
- User login
- Upload of documents (directories of images) to the TRP server
- Upload of documents and related PAGE files to the TRP server
- Display of thumbnails for each document
- Display of documents available for the public or for a specific user
- Display of jobs carried out on a document (e.g. line processing)
- Display of individual versions of a document
- Display of the structure of a page image
    - According to the level of segmentation the blocks, lines, words and the transcription is displayed in a separate window (easier navigation)
- Editing the segmentation of a page image
    - Editing metadata
    - Adding, removing and editing regions
    - Adding baselines with automated generation of line and text regions
    - Applying structural metadata (e.g. margins)
- Editing of the transcription
    - Applying format information (italic, superscript, etc.)
    - Applying user tags to the transcription (e.g. named entities)
    - Adding special characters with a virtual keyboard
- Trigger automated processes
    - Block segmentation
    - Line segmentation
    - Baseline segmentation (only applicable where line region is already available)
- Export of documents
    - As TEI with region or line based coordinates
    - As PDF with image in the foreground and text in the background
    - As complete package with images and PAGE files (for HTR training)

For completing the ground truth infrastructure we will in year 3 include the "Editorial Declaration" which means that the user is able to describe the features of the Ground Truth as well of its transcription. E.g. some documents contain correct line regions, whereas others were produced with correct baselines only. This can then be recorded in the "Editorial Declaration".

1.2. Standardized Workflow for GT Production

In the following we describe briefly the standard workflow for GT production within tranScriptorium as it can now be carried out.

1. Selection of material
   Page images need to be scanned with at least 300 ppi. Lower resolution can be technically processed, but will lead to less accurate results.
   Users are completely free to decide which documents they want to upload and process.
2. Upload of documents to the server
   A registered user will upload a document to the server via Transcribus. An FTP based connection enables the upload of large collections of page image files.
3. DIA tools

A registered user is able to run the automated block and line segmentation (hosted in the server application) on the pages foreseen for processing, typically the whole document.

4. Correction of segmentation
   In order to simplify the correction process for the segmentation tools this process was strongly simplified. Users now need to just correct the baseline of a line and not the complete line region. A special feature supports them in adding baselines.
5. Transcription
   Once the segmentation is corrected users will transcribe the text, usually the first 50 pages of a document. If the text is already available than it can be copied directly into the text processing window (as long as the number of lines corresponds with the number of lines within the page image).
6. Editorial Declaration
   As already mentioned this feature will come in 2015 but already now users are able to use the "Description" field to record the most important decisions within their transcription work.
7. Export
   The user as well as the research groups in the project are able to access the GT documents and to download the documents in a standard format (PAGE) which can be used as input for training the HTR engine.

Once the training is completed and the corresponding HTR profile is submitted to the TRP server hosted at UIBK the production of GT can be continued with support of the HTR engine.

8. HTR Processing
   The remaining pages of a document or additional documents are processed on the basis of the 50 pages transcribed. Both training, as well as recognition is currently carried out offline, i.e. not accessible to the user via the Transcribus tool. The results are then automatically displayed in Transcribus as a new version.
9. Correction with HTR support (not implemented yet)
   A specific user interface will support the user to complete the transcription of his document with support of the HTR engine.
10. Finally the correct transcript can be exported in several formats, as they were already set out in Deliverable D2.1., namely TEI, PAGE, and METS.

The following screenshot shows the main working environment for GT production.

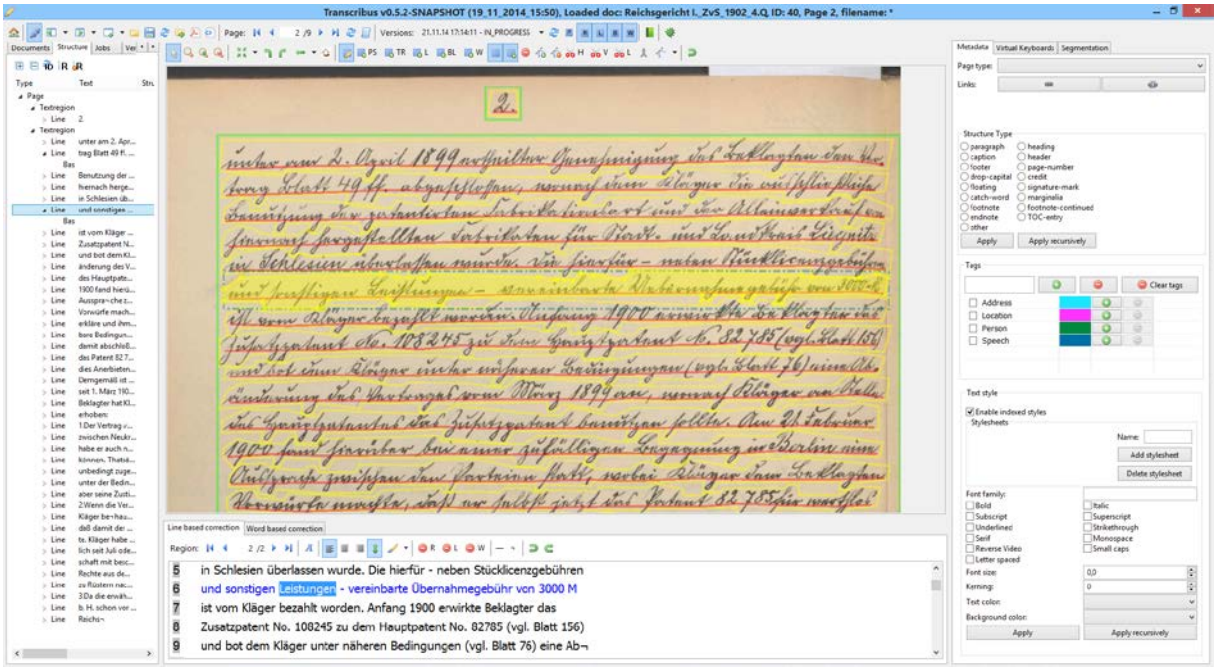

Figure 1 Display of Segmentation and Transcription

1.3. Lessons learned

There are mainly two lesssons we have learned in year 2 of the project with respect to the production of GT, but also in transcribing text in general:

- First, the production of correct line regions is very time consuming (takes up to 20-30‘) per page and it is therefore not to expect that many users will accept this hurdle. We therefore had to find an alternative and are now working with corrected baselines. This improves the speed of manual correction significantly and does on the other hand not affect the quality of the HTR recognition process.

- Second, it will be necessary to provide guidelines for generating GT („How to“) and lead the user directly to this information in order to standardize the GT production as much as possible. The main advantage of our approach not to separate between GT and transcription process is that all these features are of eminent importance for a user friendly GUI as well.

2. Actual Status of GT Production

All documents foreseen for GT can be accessed directly at the TRP System and can be viewed with the Transcribus tool.

The following table provides a short overview on the current status (December 2014). Please be aware that this is the snapshot of a running system, so that slight differences may appear with the figures from January or February 2015.

2.1. UCL – Bentham Collection

ID	Title	No. pages	Comments
106	Batch1-Final	433	- All pages with text regions, lines, and words indicated - All pages correctly transcribed - Text on word and line level for all pages
107	Batch2-Final	363	- All pages with text regions and lines indicated - All pages correctly transcribed - Text on word and line level for all pages
108	Batch3	47	- All pages with text regions, lines, and baselines indicated - Manually corrected? - Pages are untranscribed. - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform
109	Batch3-Uncorrected	53	- All pages with text regions, lines, and baselines indicated. - Baselines corrected manually by UCL - Pages are untranscribed - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform
110	BenthamNov2014	100	- All pages with text and line regions, and baselines, indicated - Baselines corrected manually by UCL - Pages are untranscribed - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform

111	BenthamDec2014	100	- All pages with text and line regions, and baselines, indicated - Baselines corrected manually by UCL - Pages are untranscribed - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform
131	BenthamJan2015	100	- All pages with text and line regions, and baselines, indicated - Baselines corrected manually by UCL - Pages are untranscribed - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform
151	BenthamFeb2015	100	- All pages with text and line regions, and baselines, indicated - Baselines corrected manually by UCL - Pages are untranscribed - To be transcribed by volunteers using the *TSX* crowdsourced transcription platform

2.2. INL

ID	Title	No. of pages	Comments
6	*"Hattem"* Ms. Hattem, SM C5; date: c. 1450; full manuscript size: 286 folio's or 572 pages	40	- A selection of 40 pages has been made to be used in the ground truthing process - The ms. is completely digitized and corrected. - The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are transcribed with both literal form and expansion. - Line segmentation has performed for the 40 page selection. Line polygons have been fully corrected manually.

31	*"Leiden MS pages selected for GT"* Ms. Leiden UB, BPL 3094; date: 1475-1500; full manuscript size: 159 folio's or 318 pages	96 pages, 48 images	- A selection of 96 pages has been made to be used in the ground truthing process - the selection is completely digitized and corrected, and available in Page format - The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are transcribed with both literal form and expansion. - The line segmentation has been performed automatically and serious errors have been corrected manually. - The baselines have been fully corrected manually.
37	*"Meermanno GT Selection"* Ms. The Hague, Museum Meermanno, 10 C 17; date: c. 1470; full manuscript size: 208 folio's or 416 pages	50 pages (100 columns)	- A selection of 50 pages (100 columns) has been made to be used in the ground truthing process - The selection is completely digitized, and available in Page format - The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are expanded but marked. - Line segmentation is automatical but of good quality. No full manual correction has been performed, but serious errors (merge/split/missing lines) have been corrected manually.

2.3. UPVLC

ID	Title	No. pages	Comments

101	Esposalles	173	- Manually corrected block and baseline segmentation - Correct text, abbreviations extended - No line regions
TBC	Plantas	1000	- Manually corrected block and baseline segmentation - Correct text - No line regions

2.4. UIBK

Bozen Ratsprotokolle 1620, HS 37a ID: 62

ID	Title	No. pages	Comments
84	Ratsprotokolle Bozen HS37a	508	- Pages 1–200 manually corrected blocks and baselines - Pages 1-50 correct transcripton - Pages 200 – 508 will be transcribed with support of the HTR engine in 2015

Zwettl Ratsprotokolle Vol. 2.1 ID: xxx

ID	Title	No. pages	Comments
30	Zwettl_GT	112	- Line regions, word segmentation and baselines manually corrected up to p. 83 - pages 84-112 will be corrected in the next weeks - Text on line and word level for pages 1-112

Reichsgericht

ID	Title	No. pages	Comments
40	Reichsgericht I._ZvS_1902_4.Q	9	Block regions and baselines manually corrected Text transcribed
41	Reichsgericht I._ZvS_1901_1.Q	9	Block regions and baselines manually corrected Text transcribed
43	Reichsgericht I._ZvS_1901_4.Q	9	Block regions and baselines manually corrected Text transcribed
45	Reichsgericht I._ZvS_1904_2.+3.Q_I_149_1904	6	Block regions and baselines manually corrected Text transcribed
47	Reichsgericht I._ZvS_1908_1.Q	11	Block regions and baselines manually corrected Text transcribed
48	Reichsgericht I._ZvS_1906_2.+3.Q	8	Block regions and baselines manually corrected Text transcribed
49	Reichsgericht I._ZvS_1905_4.Q	11	Block regions and baselines manually corrected Text transcribed
50	Reichsgericht I._ZvS_1905_2.+3..Q	10	Block regions and baselines manually corrected Text transcribed
51	Reichsgericht I._ZvS_1905_1.Q	16	Block regions and baselines manually corrected Text transcribed
52	Reichsgericht I._ZvS_1903_4.Q	16	Block regions and baselines manually corrected Text transcribed
57	Reichsgericht I._ZvS_1904_4.Q_I_542_1903	7	Block regions and baselines manually corrected Text transcribed
58	Reichsgericht I._ZvS_1901_2.+3.Q	6	Block regions and baselines manually corrected

			Text transcribed
59	Reichsgericht I._ZvS_1904_4.Q	12	Block regions and baselines manually corrected Text transcribed

Executive Summary for D2.1.

Work package 2 “Data Collection and Ground Truth Annotation” has two main objectives:

- to collect handwritten documents that can be used as test cases within the project and
- to transfer them into reliable “ground truth” for training and evaluating all technical deliverables in the project.

The first 9 months of the project were mainly dedicated to realize two tasks:

- Firstly to get a common understanding of “what is ground truth”, which formats have to be obeyed, what level of accuracy is required and which tools and procedures shall be applied? This is referred to as setting up the “ground truth infrastructure”. This is the focus of the current deliverable 2.1.
- Secondly and in order to lose no time a decision was made at the kick-off meeting to start ground truth production as early as possible in the project and to tackle the most prominent use case in the project: The Bentham collection.

In both tasks significant progress has been made and the following document is an expression of a very fruitful and intensive examination of this complex issue from several sides, including also several viewpoints and arguments.

As a matter of fact ground truth production must not be seen as being finished with the first phase of the project. This has several reasons, we mention here just two:

- The more ground truth will be produced in the project, the better. Not only for the success of this project, but for the success of HTR research in general it is of utmost importance that HTR research groups get “real world” and “highly representative” reference data for their future work.
- Within the project itself computer assisted transcription will be carried out via two scenarios, the Crowd Sourcing scenario (WP6, UCL) and the Content Provider scenario (WP6, UIBK). Obviously in using these interfaces ground truth in the sense of correct text aligned with page images will be produced anyway.

In work package 2 two other deliverables will be produced, one in month 24 and the final one in month 36. In these deliverables the progress on the ground truth production itself will be described on collection and document level.

1. Introduction

1.1. Data Collections and Ground Truth

1.1.1. Need for data

Though reasonable progress has been made in recent years in the development of software capable of recognising historical handwriting it is common sense among the research community that the provision of more training and evaluation data would be one of the most important prerequisites for a real breakthrough in Handwritten Text Recognition (HTR). Among several other statements this is also emphasized by (Plötz & Fink, 2009)in their research survey about the application of Hidden Markov Models (HMM) in HTR.

Plötz & Fink identify several main challenges, among them the lack of training and evaluation data:

> *The most prominent one, which can be considered a universal problem of any area of statistical pattern recognition, is the problem of limited data. Though some notable data collection efforts exist and some quite substantial datasets have also been made publicly available already, these sample sets are still far too small – and probably will be for the foreseeable future – for training a statistical recognizer that might be able to showclose to human performance in automatically reading handwritten script (p294).*

Though the concept itself is not mentioned in this consideration it is obvious that Plötz & Fink are requesting larger sets of "ground truth" data.

The concept of "ground truth" originates from cartographical tasks but it is also applicable to image processing. Ground truth means in general the linking of data to an image or more precisely to a certain area on the image. In our case ground truth data is understood that the expected output of HTR – correct text – is aligned with the image of a scanned manuscript page with the actual handwriting on the page.

The following figure gives a schematic overview:

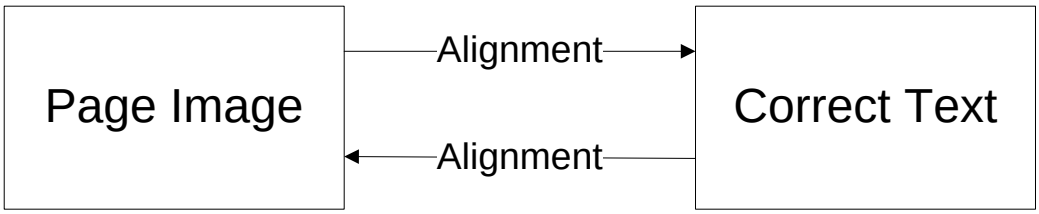

Fig. 1 Ground Truth Basic Schema for HTR

All three components provide specific challenges for the production of ground truth data which we will now discuss in more detail.

1.1.2. Page Images

Historical documents are a challenging task for digitisation. In many cases the distinction between background and actual content (handwritten characters or graphics) is a hard task on its own. Shining through or even bleed through of the ink used on the other side of the leave are general phenomena. When manuscripts are scanned most often the whole page is scanned and in contrast to book scanning it is rather unusual to crop pages in order to get "smarter" images. This leads to typical borders within the image.

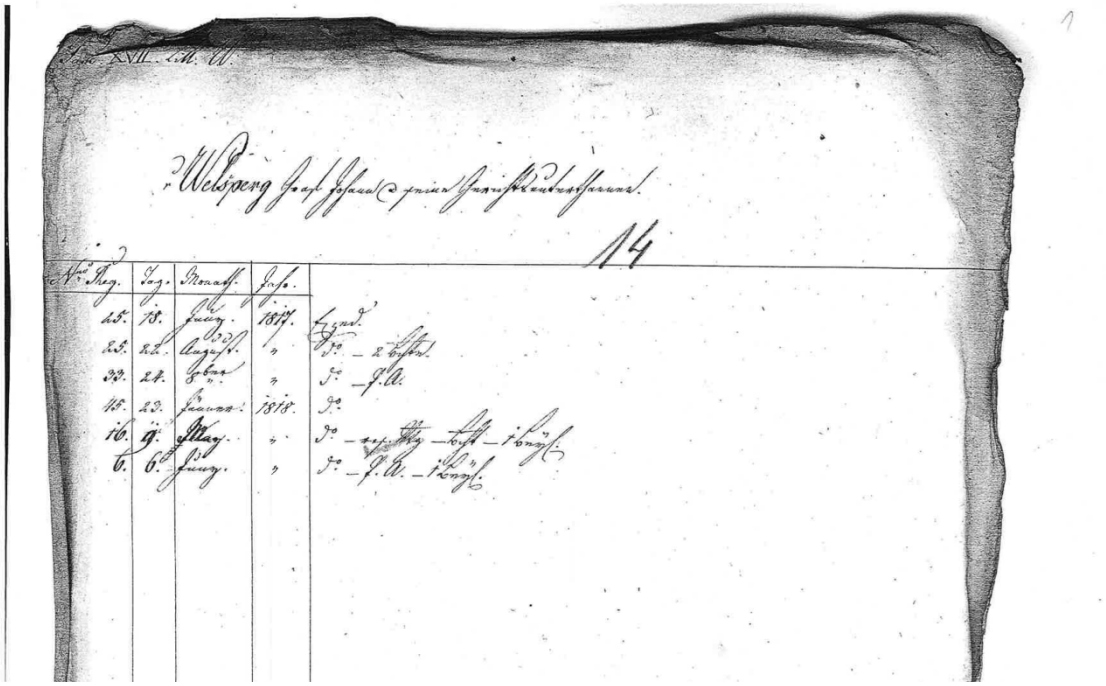

Fig. 2 Example of a typical page image from historical manuscripts

Moreover historical documents were often bound though they were– also in contrast to books - never foreseen for any binding. Therefore the text often runs to the very end of a page image and is sometimes even hidden within the binding. If such pages are scanned we can also observe some typical warping at the end of a line.

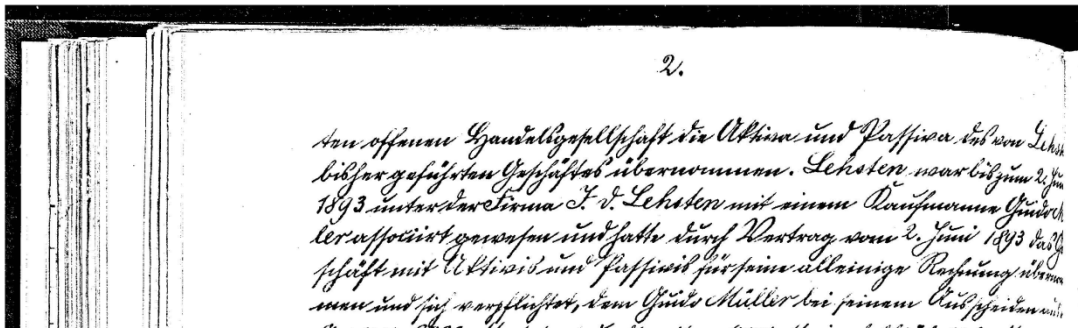

Fig. 3 Page warping and missing text

Also a common effect is that the page itself was not aligned with the edge of the whole document so that the whole page is skewed. To make this even worse all the described effects

may change within a document or a collection of similar documents from page to page so that predictions cannot be made easily. Last but not least manuscript pages are often arbitrarily warped due to the effects of humidity (Rahnemoonfar & Antonacopoulos, 2011).

As a consequence the task of image pre-processing, border detection, discrimination of background and actual content and similar steps in the workflow are a real challenge which need to deal with a large number of artefacts that will not appear in modern documents.

The consequences for ground truth production are obvious: If we speak about "the page image" as basic input for HTR we cannot expect or guarantee a certain quality. Even when scanning the same document it will make a big difference how the scanning was actually carried out, e.g. by using a V-Scanner where the manuscript needs to be opened by 120° or by applying some image post-processing and enhancement steps, such as cropping, deskewing or dewarping. Such image enhancement operations are among the standard repertoire of software included in scanning devices and are often applied by service providers.

From mass digitisation projects in OCR processing we do know that the image quality is of eminent importance for the final recognition results. It is to expect that this observation will also be true for HTR and this means that it would be desirable to have objective measures at hand to define "noise" and "scanning quality". We are not covering this aspect in our ground truth production but we do think that it is important to at least mention this aspect.

To sum up: Though the "page image" is often regarded as being a fixed input which needs not to be questioned, it will have a strong impact on the final recognition results. For real world applications carefully scanned page images are an important prerequisite which cannot always be taken as guaranteed.

1.1.3. Correct text

If we have a look at the literature on Handwritten Text Recognition we see in most cases that the "transcription" or the "correct text" are also seen as being rather simple concepts. This may be true for modern text, where handwriting does not differ from actual printing and where the modern writing system covers all aspects one may observe in manuscripts. But for historical text this assumption is naive and the difficulties to generate a "correct text" must not be underestimated.

There is on the one hand the challenge to deal with characters which are no longer part of our character set, such as an "y" with diacritics ("Ÿ") or the long "s"(„ſ").There are also many characters in historical texts which were never part of a regular set but were introduced to indicate abbreviations. In order to cope with this situation a dedicated group called Medieval Unicode Font Initiative (MUFI) has been set up by experts and provides information about specific character sets in Unicode.

Though we can observe in historical science a clear trend for "diplomatic transcriptions" many available transcriptions of historical documents rely on a rather strongly normalized text. For instance capital letters were introduced according to the modern writing style (capital letters for named entities, such as locations or persons), or common abbreviations were extended in order to ease the readability of the text. Also common historical practises such as to make no discrimination between "u" and "v" were tacitly normalized in the transcribed text. And even in modern diplomatic transcriptions we will find very rarely the long "s" which was common to most European languages until the 19th century.

From the point of view of producing GT the situation that the writing system used in the source document differs significantly from the writing system in the target document leads obviously to some drawbacks for HTR processing and HTR quality. The practical consequences also need to be seen: The production of correct text that is totally in line with the writing system of the source may be for some historical documents extremely challenging and will need the involvement of experts in palaeography. This is especially true for medieval manuscripts but also in early modern texts the writing system will need high level expert knowledge.

1.1.4. Alignment

Given that we have obtained a "good quality" page image and a text that provides also a good representation of the source document we still need to cope with the task of putting these two entities into a defined relationship. This is known as "alignment".

In the Text Encoding Initiative (TEI) which provides the most important and most widespread standardized transcription system the alignment is done usually on the level of pages: In this case the (correct) text is mainly organised in paragraphs and linked to the page image via a defined element (digital facsimile). In rare cases also a more detailed linking to areas on the page image will be done, but one will not find too many examples for this. Instead in many cases also a line break is included in the text which is – though there is no specific linking to an area on the page – an additional "alignment" information from a practical point of view.

It is obvious that for HTR tasks such a "broad" alignment is not sufficient. If we look at the HTR community the alignment of text with page images is described in detail by (Fischer et al., 2010; Fischer, Frinken, Fornés, & Bunke, 2011) for their IAM HistDB. Starting point for their ground truth production are some hundred pages of several medieval manuscripts. After having identified the text areas on the page image, line areas are detected and finally words are segmented. These areas are bounding polygons according to the granularity level (areas, line areas, words). The corresponding text is than matched with these areas.

The following table is taken from Fischer et al, 2010 and provides an overview on the alignment and how the final ground truth is available:

Ground Truth Item	Description	Format
Document Images	300 dpi, colored	TIFF
Transcription	Plain text, Unicode	TXT
Text areas	Bounding polygon	SVG
Text line areas	Bounding polygon	SVG
Text line images	Normalized, binary	TIFF
Word images	Normalized, binary	TIFF

Fig. 4 GT format overview (Fischer et al, 2010)

Though the alignment looks at this example rather simple the actual processing of real world data leads to much more complex situations. The manuscript taken by Fischer et al, 2010 is very near to a printed book – characters are seldom overlapping, lines are straight and the reading order is simple. Such a manuscript may be representative for medieval writing, but

for the mass of handwritten documents as they can be found in archives all over Europe it is by far not.

E.g. the following page image represents a typical administrative paper, as it can be found nearly in every town or larger communality: It is the first page of a protocol (minutes) of a town's council from Austria, Zwettl from the 16th century.

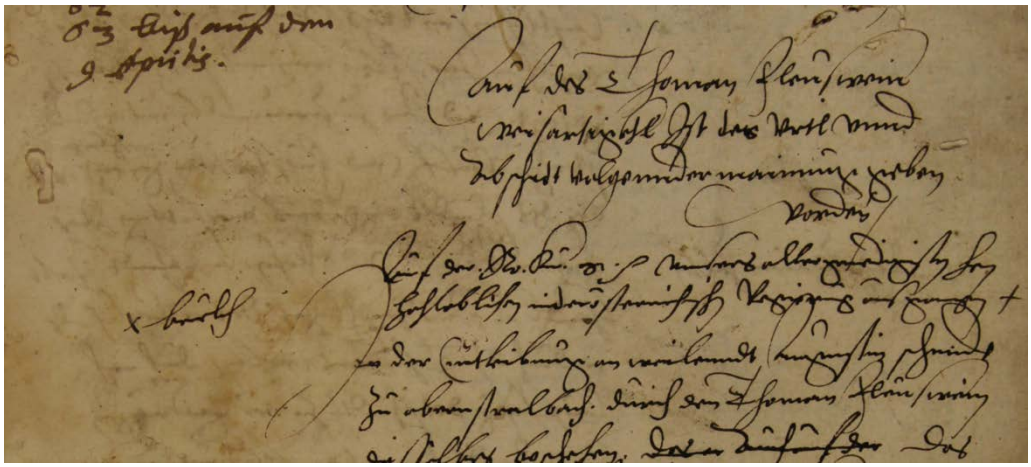

Fig. 5 Handwritten document from 16th century (Austria)

As we have seen the common way to align this source image with the concurrent text is to draw bounding polygons around the text of an area, a line or a word. But as we can see the characters touch other lines, even cross several lines, or text areas. If we want to align the text with the image this will lead to strongly overlapping bounding boxes so that the ground truth gets in some way ambiguous, since without understanding the content a decision if a specific pixel area belongs to one line or the other or both will be somehow arbitrary.

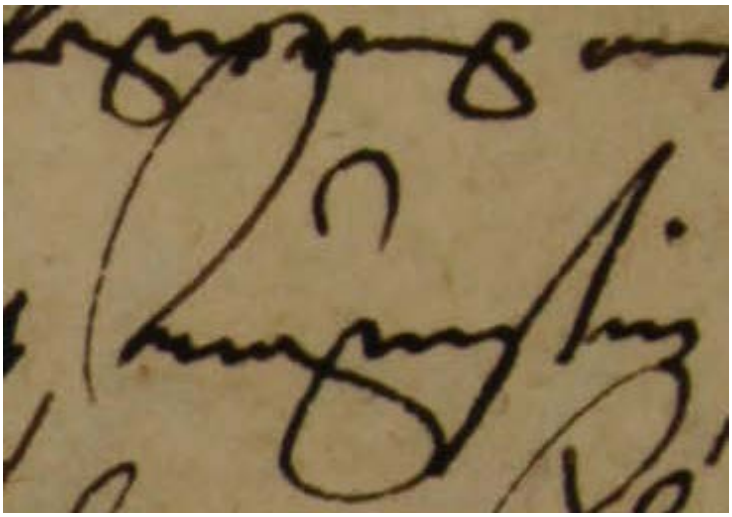

Fig. 6 Overlapping characters as a problem for alignment

1.1.5. Accuracy and data quality, or: What is ground truth?

"Ground truth" is usually understood as representing the "perfect" or "expected" result of a given criteria. Fischer et al, 2010:

> *For the ground truth, however, errors are not tolerable, because ground truth data should provide clean learning examples for the recognition systems and serve as an error-free reference to evaluate the correctness of the systems.*

Therefore in IAM HistDB the alignment of text with line areas and words is performed automatically, but corrected manually which will lead to error-free, or at least "near" error-free GT. This concept was also the common understanding in the tranScriptorium project at the start of the project.

But during the first two quarters of the project several discussions came up if it really makes sense to invest all the resources into a small portion of GT which is highly accurate and detailed, or to also accept ground truth with a lower ambition.

Just recently (Kondermann, 2013) argued very similar that

> *in many occasions ground truth generation is not necessary for performance analysis. As the most important constraints are the accuracy of the data, the cost of and the time needed for acquisition as well as the quantity of data to be recorded, it is often not feasible to e.g. focus on large quantities of highly accurate ground truth.*

Therefore he introduces three types of ground truth:

a) Reference data in large quantities

   The main criteria to still call them "ground truth" is the claim that they represent in an adequate way the problem which shall be tackled. For instance the medieval document selected by Fischer et al. may be representative for medieval texts but it is not representative for manuscripts in general. Moreover "large quantities" would mean that thousands or even hundred thousands of images from at least hundreds or thousands of different writers and documents would be available. A random selection, carried out on the basis of a large state archive would probably be in this case the best way to generate really representative ground truth. The main advantage of such a large dataset would also be that a researcher could be sure that – in principle – he will be able to find for all research questions examples but also to get an impression on the frequency and distribution of a problem. Kondermann therefore emphasises that this type of ground truth can be a source of "inspiration" and guide the research process.

   If we apply this approach to tranScriptorium it is clear that mainly the datasets from Bozen (30.000 images) and the Reichsgericht (1000 images but representative for at least some ten-thousands of pages) fulfil this criteria.

b) Reference data with weak ground truth

   In this case the task will be to "*finding methods to annotate existing reference data with ground truth information while ignoring any hard accuracy constraints*". Obviously some compromises need to be accepted when using such data for evaluation purposes, but nevertheless it will be possible to use the results for improving parameter settings or to try out new methods. Applied to our project this could mean that transcriptions are aligned on page level with text blocks or line level by using the information provided in the text format (line breaks) and by an automated method of detecting baselines in the corresponding images.

c) Reference data with Ground Truth
This category represents the "traditional" approach of ground truth, where all data are manually assessed and corrected if necessary. Though even such a procedure will not produce "the truth" but only reflect one view on a complex situation it is clear that such high level data would be highly desirable. Nevertheless the main constraints are costs and resources needed to produce them. Within tranScriptorium some hundreds of pages with ground truth will be produced from a variety of documents and in several languages.

1.2. tS Overall architecture for GT

1.2.1. General considerations

As we have seen from the introduction it is a complex task to produce ground truth that fits for Handwritten Text Recognition which needs to meet a number of requirements. In the following we will present our approach which is a true deliverable of the first three quarters of the project. Several working papers, presentations, tools and workflows were generated and are now representing the state of affairs in this matter.

In other words: The first months were dedicated to "requirements engineering" which will now be described in more detail:

(1) tS will follow a "soft" approach to GT
GT will come with different quantities and with different granularity levels in terms of features and accuracy. This view on GT goes together with a pragmatic approach that we "take what we can get" from the communities which are already transcribing handwritten texts and are therefore experts in the field.

(2) Metadata on the alignment process
In order to cope with the complex situation but also since we have this "soft approach" we will apply as much metadata to the GT data as possible. This will not only comprise "well-known" metadata such as date or language of the document but include also metadata on the transcription and alignment process. As far as we can see this has not been attempted before and is one of the innovative aspects of our approach.

(3) Use of existing standards and resources
Instead of "reinventing the wheel" we will take up existing standards from the domains in which HTR and text recognition takes place, such as the TEI text format, or the PAGE format for describing layout features and aligning text to images. In this way we hope to ensure that our reference data will become "real" references and serve as input to many other research projects.

(4) Adaptive workflow
According to our "soft" approach for GT we also need a flexible and pragmatic solution for organising the GT workflow. Therefore purely automated solutions are coexisting with manual procedures. Besides specific tools developed in the project we will also rely on existing and well-known tools from the document analysis community.

The following sections will explain in more detail how these requirements are realized in the project.

1.2.2. Transcription and alignment rules as metadata

In the introduction we have explained that the transcription process requires a large number of detailed decisions, such as the way the long “s” is treated: Either as a specific character „ſ“or according to our modern writing system as a simple “s”. Instead of providing a prescriptive rule which would say “transcriptions must be diplomatic and therefore the character set of the source document must be represented correctly also in the character set of the transcription” we introduce a pragmatic approach: There are several options to handle this issue, but standardized information is required to describe the differences.

In other words: Metadata on our data set will help to provide a detailed description of the reference set. Moreover this metadata will be provided in a standardized way. Currently even in digital editions of transcribed texts the “editorial declaration” comes as a running text and with no standardisation at all. In contrast we have set up a feature list with which the inherent transcription rules and decisions can be made explicit. A simple example is shown in the table below.

Transcription feature	Option 1	Option 2	Option 3	Option 4
Capital letters are irregularly used in the source	Transcribed as in the source	Always normalized to capital letters according to a modern lexicon	Only capitalized in specific cases, such as beginning of a sentence, or person and geo-names	To be defined

Fig. 7 Metadata for the editorial declaration in TEI

All three options are allowed and other options can also be added if necessary. These options describe common ways to deal with this feature of the source document. It is our ambition to collect such editorial metadata and to provide a comprehensive table that can be used to describe most transcriptions as they were produced by historians or linguists. One of the tasks for ground truthing will than be to generate such an editorial declaration for each reference set – which can be done by experts easily and with a minimum of effort.

The same approach will be taken for making the alignment process explicit as well. In this case decisions are made such as: How is the overlapping of line regions handled? How are baselines defined? Or what is regarded as “content” of a page and what as “noise” (e.g. page numbers or stamps introduced by an archivist).

But there is one important difference to transcriptions: Whereas transcriptions stem usually from human beings and represent therefore a high quality (e.g. transcribed text will come with an accuracy rate of beyond 99,9 percent) the alignment process will often rely on automated or semi-automated procedures. For instance given that a TEI text is available with line breaks and the document has a clear structure (e.g. only one or two paragraphs per page) the automated alignment of the text on line level will produce good but not perfect results. And

in this way thousands or ten-thousands of page images may be processed with a moderate effort. Nevertheless this process will also produce errors. Here we will apply the same approach as above and provide a simple accurateness figure for this aligned feature. Again we are convinced that this meta-information will make our data much more transparent to other researchers and support the uptake of the reference set.

1.2.3. GT Format: PAGE

One of the main deliverables of the IMPACT project (Balk & Conteh, 2011) was the introduction of the PAGE (Page Analysis and Ground-truth Elements) format to the research community by the PRIMA research group at the University of Salford (Clausner, Pletschacher, & Antonacopoulos, 2011; Pletschacher & Antonacopoulos, 2010).

PAGE is an XML based format which is dedicated to provide a comprehensive and detailed description of ground truth for Image and Layout Analysis and Optical Character Recognition (OCR) by synthesizing and exceeding several other available formats. Since its introduction in 2010 PAGE was used in several projects, but also for e.g. organising competitions at the ICDAR conference 2011 and 2013 (Antonacopoulos, Clausner, Papadopoulos, & Pletschacher, 2013).

Taking this into account UIBK suggested to uptake the PAGE format for the tranScriptorium project. Main criteria were:

- The format is open and freely available.
- The format is well acknowledged by the document analysis community.
- The format covers nearly all features which are required for HTR processing.
- The PRIMA research group is part of the IMPACT Centre of Competence (http://www.impact-project.eu/news/coc/) and was therefore able to guarantee the further maintenance of the format at least for 2013-2015.
- UIBK was one of the project partners at IMPACT and responsible for organising the actual GT process (with support of service providers).
- There is a supporting tool, called ALETHEIA which was developed for the manual production of GT and which may also be used for (some) parts of the GT workflow in HTR processing.

At the kick-off meeting a decision was taken and the PRIMA research group was contacted for further support to adapt PAGE for some detailed feature requests. The main request was the introduction of a “baseline”. A baseline is defined as a (virtual) poly line where the characters are actually “sitting”.

Fig. 8 Concept of a baseline

Due to requirements raised by other research groups as well, the PRIMA group released a new version of the PAGE format in November 2013 (with a pre-release in August 2013). In parallel the ALETHEIA tool was also updated so that a full tool set is now available for the manual production of GT also for HTR purposes.

1.2.4. GT Format: TEI

The TEI (Text Encoding Initiative) document format dates back to the 1990s and can be regarded as a de-facto standard in the humanities for editing text. It is also used for transcriptions of historical handwritten documents from researchers all over the world.

Nevertheless there are some drawbacks of TEI:

- In some research communities, such as history or social sciences, TEI does not play such a dominant role as with the literary and linguistic community. Therefore many transcriptions which may be interesting for HTR are not available in TEI but as Word or PDF files.
- TEI is a "grown" format and rather complex. It provides in many cases more than one solution so that the usage and application of the elements is somehow arbitrary and therefore not as standardized as one may wish.
- The linking between the text and a page image is possible, also for areas on a page image, but there are no standard tools or other resources available to carry out this work easily.
- The same is true for editing TEI texts in general. Though some TEI editors were launched the most common way to produce a TEI file is still to work with an XML editor, mainly Oxygen.

The consequences are that TEI as input format for ground truth text does not fulfil strict constraints and requirements which are usually regarded to characterise a ground truth format. On the other hand, and taking into account what we said above about "soft ground truth" the decision was taken in the project to accept TEI as the main input format.

1.2.5. GT Package

We have now the main components of our envisaged ground truth infrastructure available.

These components are:

a) PAGE format
   An open, XML based format to link text to page images on the level of areas (regions), lines, words and even glyphs, with additional features such as the baseline. So the final output of high accuracy GT will come as PAGE files.
b) TEI format
   An open, XML de-facto standard for large amounts of (transcribed) text. TEI is mainly used as document format, page breaks and line breaks can be included. TEI will be used as GT format mainly for the "soft ground truth" of larger amounts of text.
c) Transcription and alignment rules on the basis of document features
   In order to describe the detailed decisions which are (or were) necessary to produce a transcription respectively to align text with the corresponding page image a metadata table on feature level will be produced.
d) A workflow which offers two principal ways to produce GT either semi-automated by applying several tools from UPVLC (cf. Part 2 of this deliverable) or to produce it completely manually with the ALETHEIA tool (cf. section 4).

The following figure provides an overview of the envisaged ground truth schema which is used within tranScriptorium.

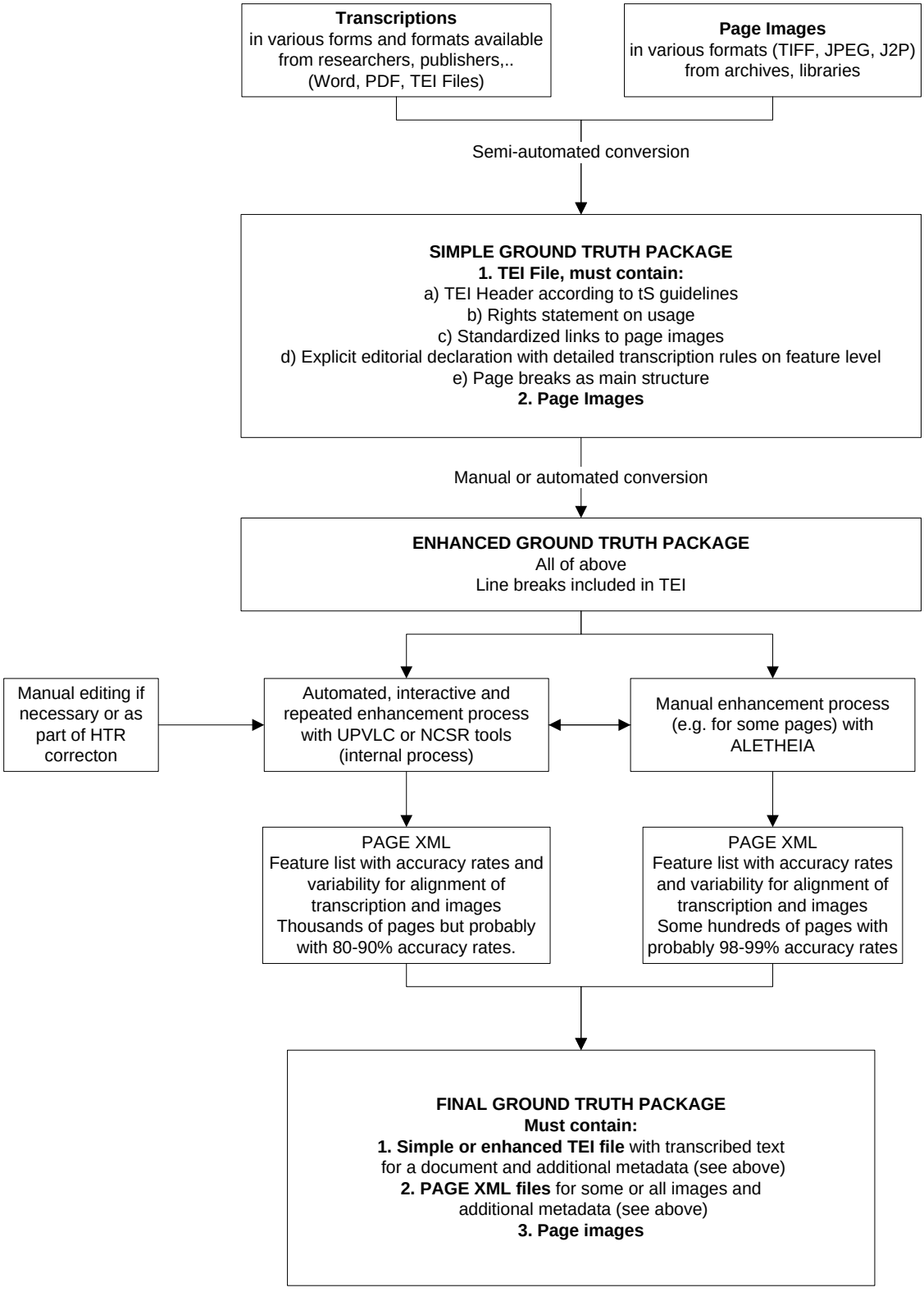

Fig. 9 Ground Truth overall schema

2. Data collections for tS

2.1. Introduction and overview

According to the partners of the project we agreed to provide data collections and ground truth in four languages: Dutch, English, German, and Spanish. The selection of the data was done as a collaborative effort, where the following partners are responsible for the above mentioned languages: INL, ULC, UIBK and UPVLC.

The main selection criteria were:

a) Significance of the document and/or collection

   Since tranScriptorium is dedicated to apply HTR to "real world documents" the selected documents should be representative, either for the type of document itself, or for the handwriting or for specific problems connected with transcribing documents in general. Therefore the documents range over more than 500 years, from late medieval texts to early 20th century, from botanical books to administrative papers and philosophical working papers and notes.

b) Availability of images and text

   In the best case, both text and images were already available at the beginning of the project. The task would than be reduced to aligning these two components. This was the case with the Bentham collection. In other cases it was necessary to also organize also the transcription of a document, or the other way round: Transcriptions were available of highly significant documents, but the images had to be produced.

c) Intellectual Property Rights (IPR)

   From the very beginning the project aimed at providing ground truth which can also be used outside of the project or after the project. Though it is obvious that there is no copyright on page images many archives and libraries still hold back the master files of their digitisation projects. Moreover transcriptions are an intellectual work and the rights are therefore with the author respectively with the publisher if the transcription was published as a book. Finally person rights have to be obeyed for example in the case of 20th century material.

The following table provides an overview of the collections and documents which were selected at a first run in the project.

Collection/document	Partner	Significance	Availability	IPRs
Bentham	UCL	Notes and working papers of the 18th century from several hands in English.	Images and thousands of pages of transcribed text available as high quality images together with TEI documents.	Free for research
Plantas	UPVLC	A famous document from the 17th century in old Spanish.	Approx. 1000 page images	Restricted access for research is possible
Esposalles	UPVLC	Two volumes of a large collection (291 vol.) of marriage registers	Some hundred page images are used for GT	Restricted access for research is possible.
Zwettl Ratsprotokolle	UIBK	A complete run of administrative documents from Zwettl (town in Austria) from 1470 to the late 19th century.	More than 5000 page images and transcribed text (Word format) available.	Free for non-commercial purposes
Bozen Ratsprotokolle	UIBK	A complete run of administrative documents from Bozen (former Austro-Hungarian Empire) from 1470 to 1670.	More than 30.000 (!) page images. No transcription available at the start of the project.	Free for non-commercial purposes.
Urteile des Reichsgerichts	UIBK	A representative sample from the High Court in Germany (Reichsgericht) from early 20th century. The complete collection would comprise some	Around 1000 page images with transcriptions (in Word format).	Free for research purposes (person names must be deleted).

		ten-thousands of pages			
Ms. Vienna, ONB 2818; **date: c. 1490;** **size: 318 folio's or 636 pages;**	INL	The complete ms. will be used in the ground truthing process.	the ms. is completely digitized and corrected The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are expanded but marked.	Free for research	
Ms. Hattem, SM C5; **date: c. 1450;** **size: 286 folio's or 572 pages**	INL	A selection of 40 pages has been made to be used in the ground truthing process	The ms. is completely digitized and corrected. The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are expanded but marked.	Free for research	
Ms. Leiden UB, BPL 3094; **date: 1475-1500;** **size: 159 folio's or 318 pages**	INL	A selection of 96 pages has been made to be used in the ground truthing process	the selection is completely digitized and corrected The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are expanded but marked.	Free for research	

Ms. The Hague, Museum Meermanno, 10 C 17; **date: c. 1470;** **size: 208 folio's or 416 pages**	INL	A selection of 50 pages (100 columns) has been made to be used in the ground truthing process the selection is completely digitized.	The transcription is diplomatic and distinguishes u and v. Long and short s are not distinguished. Abbreviations are expanded but marked.

2.2. Bentham

The new critical edition of the works and correspondence of Jeremy Bentham (1748–1832) is being prepared from the collection of papers held at University College London Library. In spite of his importance as jurist, philosopher, and social scientist, and leader of the Utilitarian reformers, the only previous edition of Bentham's works was a poorly edited and incomplete one brought out within a decade or so of his death.

The papers consist of drafts and notes for published and unpublished works, and cover many subjects including: Bentham's codification proposal, a plan to replace existing law with a codified system, an idea which manifested itself in Constitutional Code (London, 1830), a blueprint for representative democracy and an entirely open and fully accountable government, 1815-1832, to mention just a few of a large number of different subjects.

The collection consists of approximately 60,000 folios. Whilst some are in the hand of secretaries or copyists, the majority are in Bentham's own handwriting which can be very difficult to decipher. In addition, Bentham often went over a manuscript several times, crossing out words or phrases and adding new ones in the form of notes, interlineal additions and marginalia.

2.3. UPVLC

2.3.1 The PLANTAS Database

The manuscript piece entitled „Historia de las Plantas" („PLANTAS") was written by Bernardo de Cienfuegos, who was one of the most outstanding Spanish botanists in the 17th century. This manuscript of seven volumes is held at the Biblioteca Nacional de Espaa (BNE) and a digital reproduction of it can be found at „Biblioteca Digital Hispánica".

Fig. 10 Plantas Example pages

The reason for working with this manuscript is that it is written in (old-) Spanish, which is one of the languages that the tS project is committed to working with. On the other hand, the partner UPVLC established a contact with the BNE institution, which showed a lot of interested in collaborating with the project by providing the mentioned manuscript piece. In addition, the BNE offered support in the transcription process of the manuscript, task which is currently carried out by UPVLC members and people of BNE. This manuscript, written in old Spanish employing quill-pen, is a source of valuable information of great scientific interest which attempted to review the botanical knowledge of that time along the history. The manuscript itself is composed of seven volumes of different numbers of pages, from which the first one was considered for a complete transcription, while the others for being indexed and then employed by the key-word spotting (KWS) approaches developed during the project. This volume, which has over 1,000 pages containing more than 17,000 handwritten text lines, includes also drawings of plants in colour and in black and white. In addition, it employs a large number of Latin names, frequently used to identify the different plant species, and other non-Latin terms as: Greek, Arabic, Hebrew, Portuguese, Valencian, Catalan, French, German, English, Flemish, Polish, Bohemian and Hungarian.

2.3.2. The ESPOSALLES Database

Huge amounts of handwritten historical documents residing in public and private institutions have been made available to the general public through specialised web portals. Many of these historical documents contain very valuable information in the form of records of quotidian activities. One example of this type of handwritten documents are marriage license books. In many cases, it would be interesting to transcribe these document images, in order to provide new ways of indexing, consulting and querying them. These documents have been used for demographic studies and migration research. Recently, interesting competitions have been proposed in order to promote the research in similar collections. This type of document is characterized for being very regular through the time, and some of these collections seem as an accounting book. Recognizing the regularities that are present in these documents makes them very attractive for layout analysis. In addition, the transcription of these documents is very challenging because they have really difficult language model problems

since they usually contain many proper names. These characteristics were considered very interesting in tS and this was one of the reasons these types of documents were chosen for project.

Partner UPVLC has been researching in the transcription of this type of documents in recent years in the context of several Spanish projects[1]. This research has been carried out in close collaboration with other research groups[2]. These research groups have working with this type of documents collected for another EU project[5]. As a result of this research, several databases with their corresponding GT have been produced (Romero et al., 2013) and were available for tS from the beginning of the project. This was another important reason for selecting the databases in tS.

The ESPOSALLES collection used in tS is currently composed by two volumes belonging to the Marriage License Books collection (called Libres d'Esposalles), conserved at the Archives of the Cathedral of Barcelona. The full collection is composed of 291 books with information of approximately 600 000 marriage licenses given in 250 parishes between 1451 and 1905.

One of the volumes used in tS has 173 [11] and more than 5,400 lines. It is written just with one writer in old Catalan. The other volume has 593 pages, although just 200 pages have been initially used. These 200 pages have more than 5,800 lines and it is written in old Spanish. As previously mentioned, this collection is being used in another EU project[3] that is focused in demographic research and it was agreed with responsible of that project that additional books could be available for tS upon request.

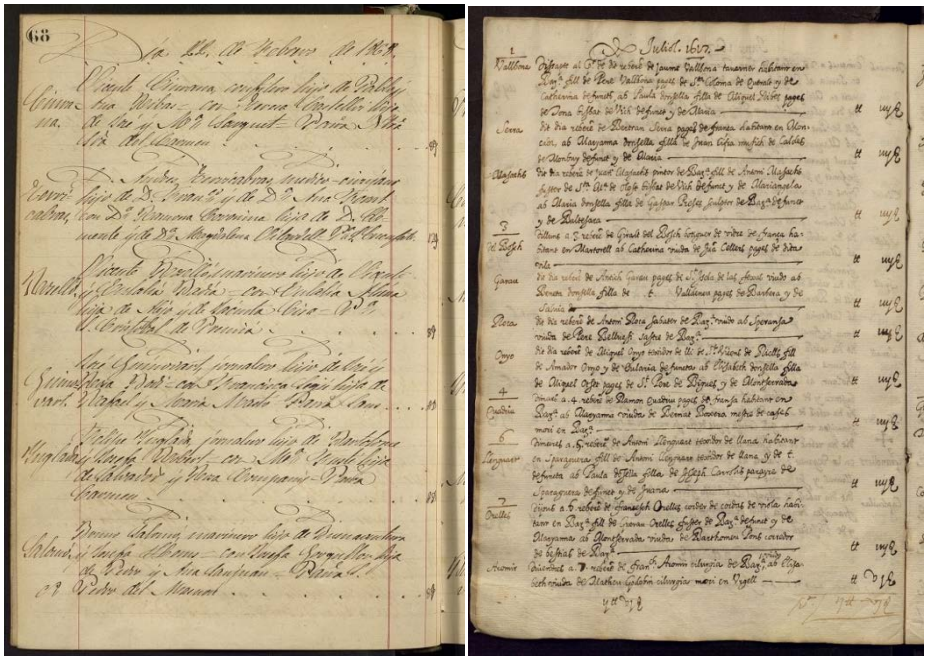

Fig. 11 Page images from Esposalles

[1]https://prhlt.iti.upv.es/page/projects/handwritten/iketdihc/mittral/index

[2] http://www.cvc.uab.es/?page_id=222

[3] http://dag.cvc.uab.es/5cofm/

2.4. INL

2.4.1. Selection

INL has decided on compiling a data collection of four late 15th-century manuscripts, all written in the cursive style and containing texts on the science and practice of medicine and surgery. This selection was made on the following lexicological grounds:

1) the manuscripts, all belonging to the same time period and to the same social and scientific domain, offer a well-delineated and consistent corpus of middle Dutch technical terminology; this uniformity undoubtedly is advantageous in the development of new HTR-technology;
2) this kind of texts only recently (i.e. since the eighties) became subject of wider linguistic research; therefore, the description of the vocabulary used in these fields is for a greater part absent in the vocabularies of historical Dutch; using these texts in tS gives us the opportunity to fill gaps in the stock of historical words kept at the INL.

2.4.2. Specific challenges

In all manuscripts the same problems had to be solved: how to deal with abbreviations, references (and their symbols), deleted, corrected or illegible letters and words; how to handle specific symbols and (to a lesser degree) illustrations; how to deal with the layout.

- For one ms. (ms. The Hague) the correction of the transcription isn't finished yet; for all other manuscripts (or for the selection taken from them), the transcription is available in a fully corrected form.
- for all complete manuscript transcriptions, TEI conversion is still pending
- for the selection taken from ms. Hattem the tei-conversion and the conversion of abbreviations and symbols into unicode/mufi-compliant unicode has been done; also the alignment + Aletheia-work has been done (although it still needs to be corrected); for the selections from ms. Vienna, Leiden and The Hague, these activities still remain to be done.
- for the Page XML encoding of the Hattem selection (2a), coordinates of line shapes are not completely corrected yet. This is a task to be completed with the help of the Aletheia tool.

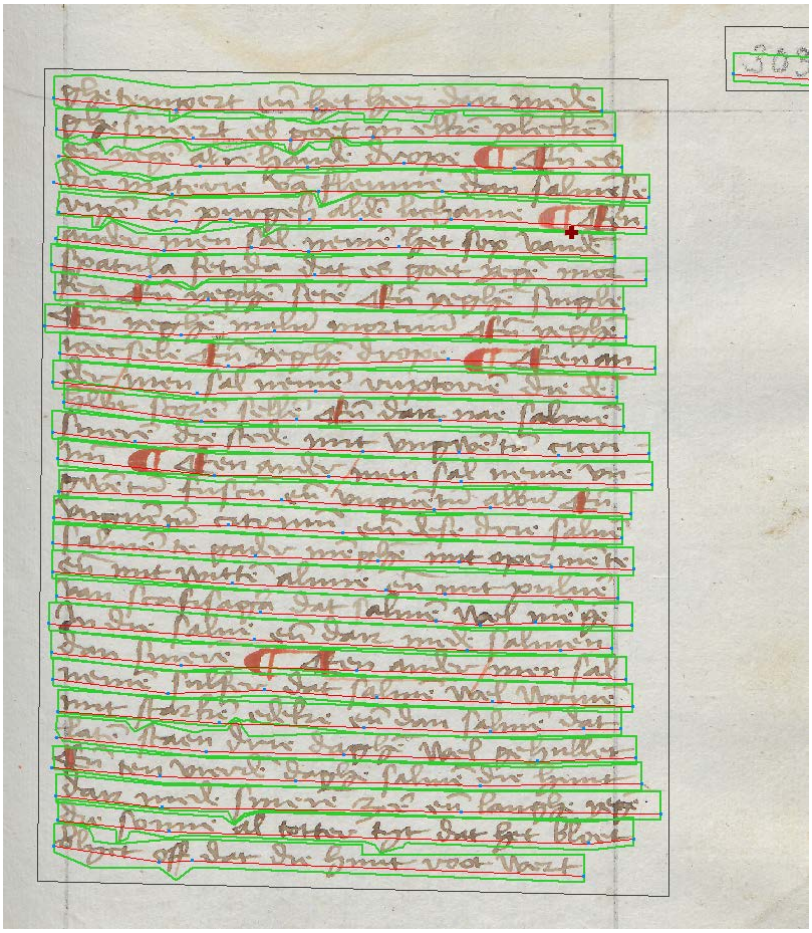

Fig. 12 Page image from Ms. Hattem

2.5. UIBK

2.5.1. Ratsprotokolle Zwettl

The Ratsprotokolle Zwettl cover the regular minutes of the council of Zwettl, a town in Lower Austria from the 14 July 1553 until 1880. They were digitised and transcribed in the last few years by a team of historians. Therefore more than 300 years are completely covered, comprising more than 5,800 page images. The transcription is available in 20 volumes, covering around 2,000 pages. The council protocols were created by a large number of writers and are highly representative not only for this specific genre ("Ratsprotokolle" are existing for nearly every town or larger communality) but also for administrative documents in general (Scheutz & Weigl, 2004).

Both page images (available as JPG from a camera) as well as transcriptions (available as Word file) are in a good quality so that we will be able to focus on the alignment process.

As a drawback of the source we have to mention that the hands are sometimes hard to read, and also the language is in the first centuries far away from any standard German and reflects a strong influence of the Lower Austria dialect.

One of the main advantages of the collection is that both transcription and images are freely available for non-commercial purposes, so they can be used in many ways and without further restrictions.

2.5.2. Ratsprotokolle Bozen

The same that has been said for Zwettl is also true for the Ratsprotokolle Bozen. The collection covers currently the period from 1469/70 until 1684 and comprise about 30,000 page images. The digitisation of the next period, from 1685 to 1800 is already in preparation and will be carried out in the next years. Finally about 70,000 page images for a period of more than 300 years will be available.

The collection is a “serial source” which means that it is ordered by year and date of a meeting. Also within the minutes a relatively stable structure is followed, e.g. to start with the number and date of the meeting and the name of all participants present. The policy of the archive is very open, so that again all images can be used for non-commercial purposes in any way.

The first part of the collection is already online and can be accessed via the BOhisto Portal. The following illustration gives an impression:

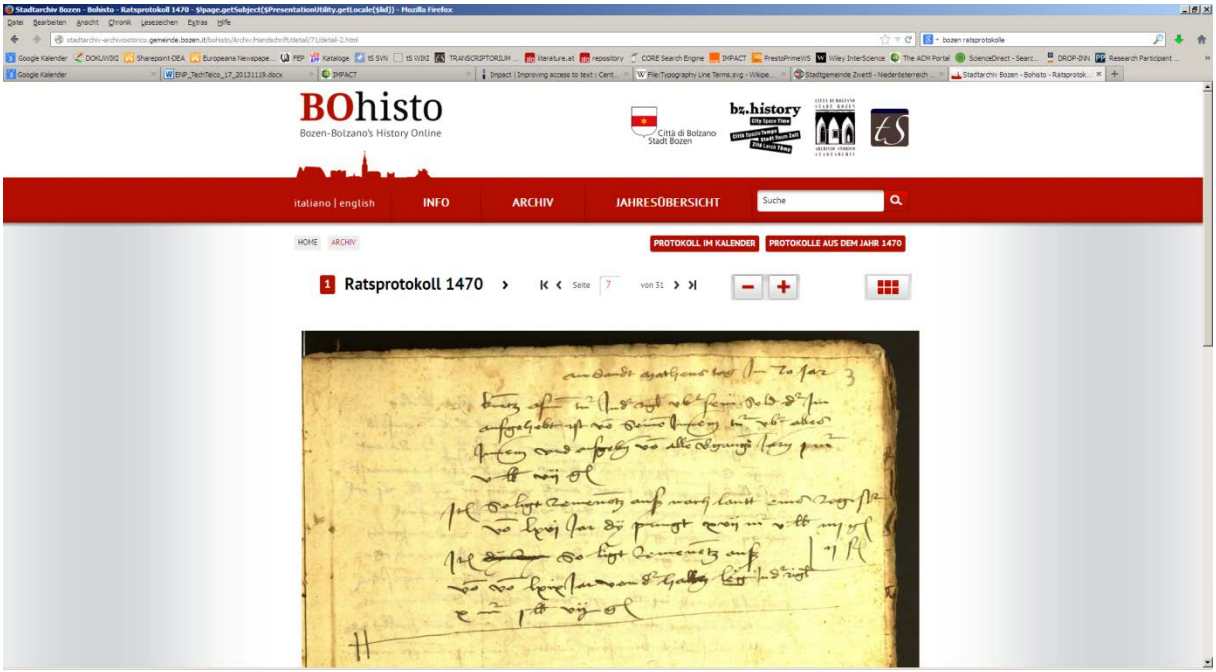

Fig. 13 Ratsprotokolle Bozen - Screenshot from BOhisto portal

In contrast to Zwettl there was at the beginning of the tranScriptorium project no complete transcription available, so that currently only 500 pages were already transcribed.

2.5.3. Reichsgericht

Due to a co-operation with the Eberhard-Karls-University-Tübingen, resp. the Institute for Legal History, we are also able to deal with files from the "Reichsgericht", which was between 1879 and 1945 the highest court in Germany. The files cover a broad range of writers and writing styles. Transcriptions are available for selected court decisions from the years 1900 to 1910 and cover about 1,000 pages.

What makes this collection so interesting are several factors:

- The selected decisions are representative for a large amount of court decisions. Not only for the highest court, but also for other courts in Germany.
- The papers have a regular structure and – what is even more important – are a clean copy of the original papers. They were written by professional writers who were instructed to write a standard German Sütterlin script.

- The language of the court decisions is modern or only slightly different from modern spelling, so that specific dictionaries are available.

3. GT production

3.1. Manual GT Production: ALETHEIA

One of the objectives of tranScriptorium is to create an awareness among archivists and historians that HTR will become a key technology in the next few years. In order to involve these groups it is important that non-technical people are able to contribute to the creation of (high-level) ground truth. With the adaptation of the ALETHEIA tool this prerequisite is met in the project.

As the PAGE format also ALETHEIA was developed within the IMPACT project by the PRIMA research group. The main objectives were:

- To create a standard ground truth tool that is able to directly export PAGE xml files
- To provide a tool which can be used by non-technical users in a simple and effective way.
- To allow that images can be locally loaded, binarised, and enriched according to the requested quality of the ground truth
- To support domain experts in creating ground truth so that computer science researchers get real world material

One of the main advantages of the tool is that it can be used so easily. In contrast to a highly specialised and optimized workflow which requires the involvement of computer specialists, a domain expert is able to create valuable ground truth after half an hour of training and completely independently from any background infrastructure by using ALETHEIA.

Another advantage was that the tool was already available at the beginning of the project in a mature version with which more than 50.000 ground truth pages were processed in the IMPACT project.

As with the PAGE format several requests were made in order to be able to produce manually ground truth files for HTR purposes. For instance, the baseline is now covered with a specific feature and can easily be drawn within a line region.

The following screenshot gives an impression of the ALETHEIA interface:

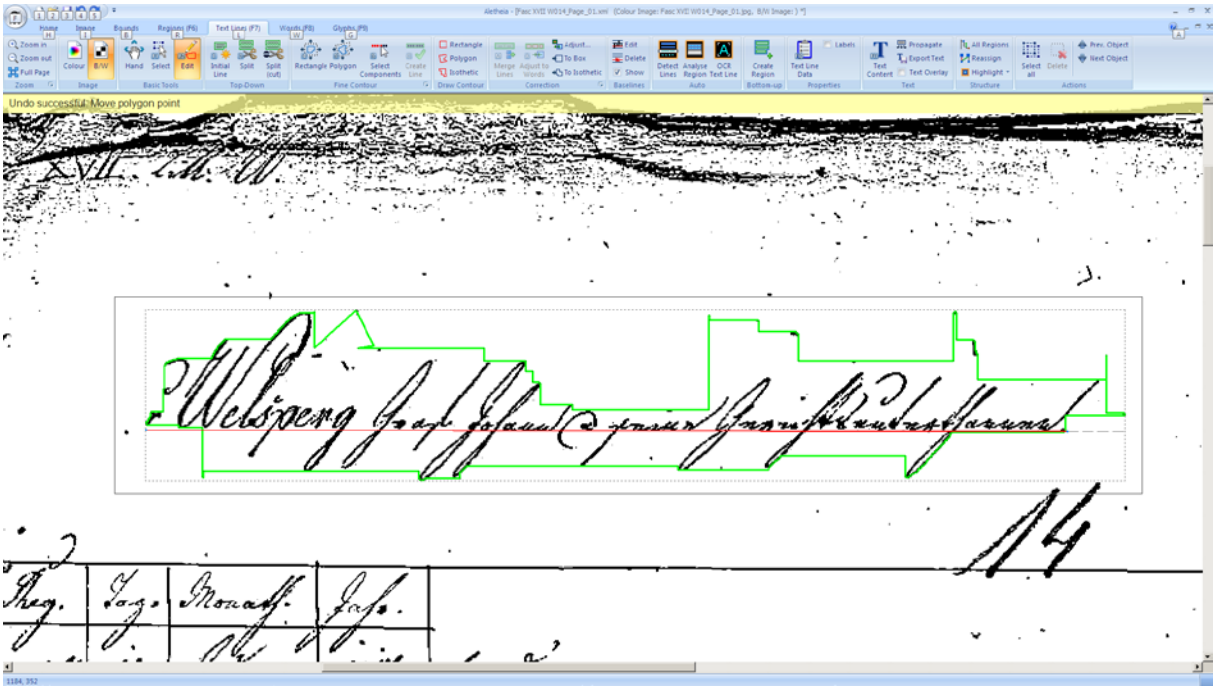

Fig. 14 ALETHEIA screenshot

3.2. Semi-automated Ground Truth Production

Since the production of ground truth requires a good deal of manual work it is obvious to automate this process as much as possible. Therefore several tools were developed by UPVLC and NCSR for overcoming these issues from the very beginning in the project and not delaying the actual production of ground truth within the project:

4. GT Tool PAGE: this tool can be used for generating and correcting basic ground truth for HTR and the information is registered in the latest version of the PAGE format (including baseline information). Fig. 12 shows a screenshot of this tool. This tool can also be used for incorporating the transcript of each line. Since the lines used for HTR could be registered with a polygon, the tool can be also used for generating and correcting this information. The tool is accompanied with a user manual that describes its facilities.
5. esposallesToPAGE: this tool takes all the existing ground truth existing for the Esposalles database and generates a PAGE file.
6. generateCollectionXML: this tool takes ground truth in several formats and generates a PAGE file.
7. PAGE2010To2013, PAGE2013To2010: these tools can be used to transform PAGE files into the two PAGE versions
8. Word Transcript Mapping: this tool takes as input the transcription in text line level as well as the corresponding text line segmentation ground truth (PAGE format) and generates a new PAGE file which includes the word level segmentation as well as the alignment with the text. An efficient transcript mapping technique to ease the construction of document image word segmentation ground truth that includes text-image alignment is used. It is based on a gap classification technique constrained by the number of words. Moreover, the tool enables the user to perform a few actions to finalize the ground truth regions such as editing, inserting or deleting segmentation regions. Only a small number of segmentation results needs correction since the proposed automatic transcript mapping technique has been proved efficient and time saving. Fig. 16 shows a screenshot of this tool.

These tools have been developed with the Qt libraries. Qt is a cross-platform applicationand UI framework for developers using C++. It is able to handle images in many different formats. It is compatible with many operating systems and platforms.

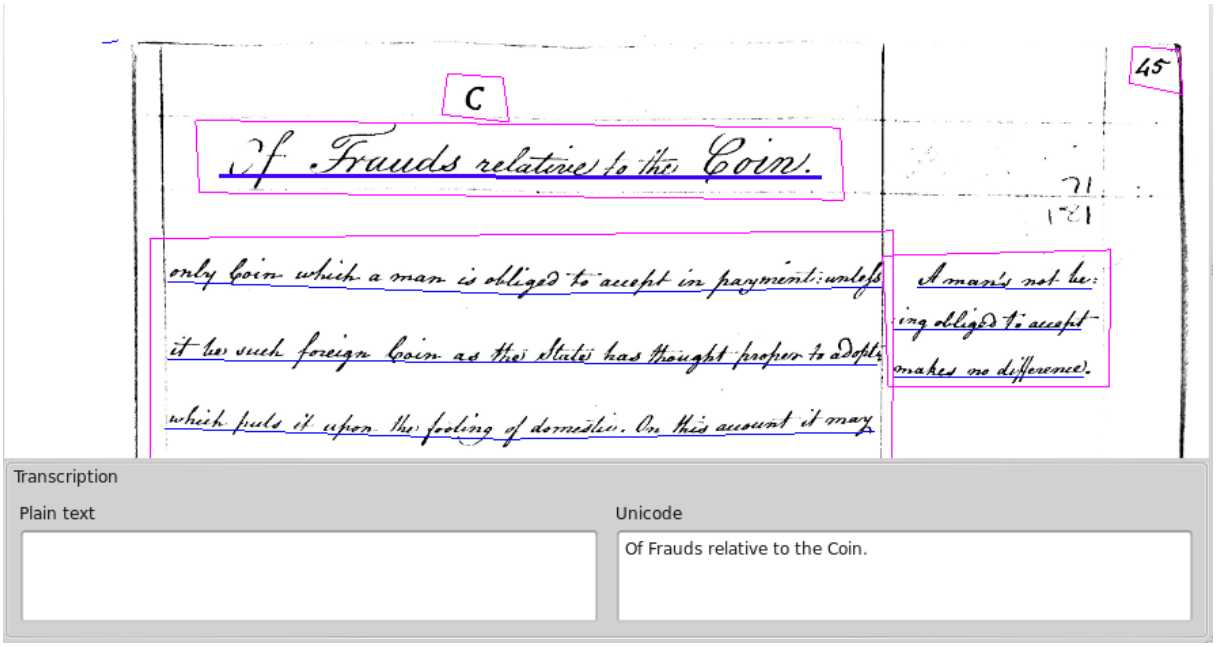

Fig. 15 GT Tool PAGE used for generating ground truth

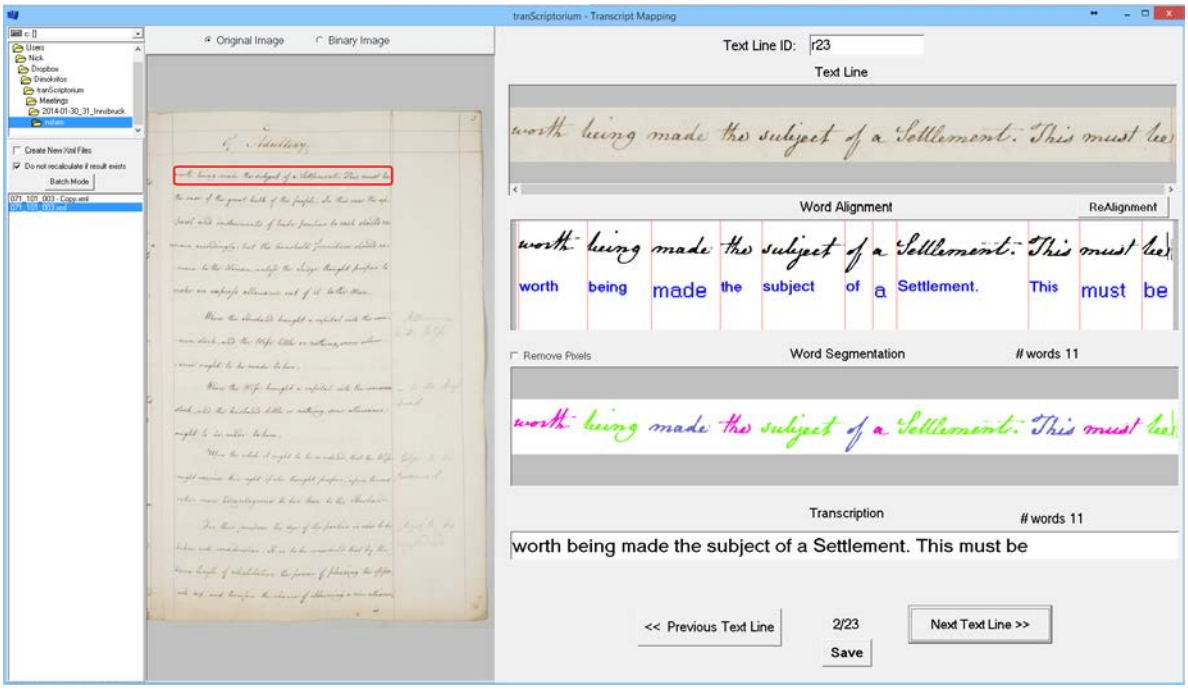

Fig. 16 A screenshot of the Word Transcript Mapping tool

4. Further aspects

2.5. Accessibility of GT

The ground truth generated in tranScriptorium will be made available to researchers on the website of the project. Dependent on the legal restrictions posed by the owners or copyright holders of the page images and the transcriptions a direct download of the ground truth packages will be possible.

A model contract was already drafted by UPVLC and will be provided together with the packages. A first version is provided here as an annex to this document.

2.6. Competitions

In order to attract more researchers to use the ground truth data sets of tranScriptorium the UPVLC team applied for a competition at the 14th International Conference on Frontiers in Handwriting Recognition (ICFHR-2014). A decision is still pending.

2.7. More ground truth

As already indicated in our introduction ground truth production will not stop after the first phases of the project, but is understood as an ongoing process.

In order to enlarge our data sets two main target groups need to be tackled:

a) Archives and libraries

Though archives often lack progress in the digitisation of their holdings there are currently several initiatives for large scale digitisation in archives. For instance the German Science Funds (DFG) has launched a large pilot project with six well-known state and federal archives from Germany to investigate all aspects connected with digitisation issues in archives. So it is expected that in a few years larger amounts of images will become available. This is true even more since archives ran microfilm campaigns for several decades, and e.g. the German State Archive has microfilmed several running kilometres of their holdings. Digitisation from microfilm is extremely cheap and therefore potentially millions of page images can be produced with a reasonable amount of money. The documents which will become available by these digitisation projects will very much be similar to the images from Zwettl, Bozen and the Reichsgericht: administrative documents stemming from all kinds of official units such as communalities, municipalities, regional and national governmental bodies. Since archives (and libraries) are in most cases the owners of these documents, they are usually also the owners of the related digital copies. Consequently archives need to be tackled for providing large data sets of digitised archival material. It has to be emphasized that the pure digitisation of source material does not create any copyright claims on the digital images, but certainly the ownership stays with the archives.

a) Researchers/historians

The situation is completely different if we look at the availability of transcriptions for the above mentioned sources. Only in rare cases employees of an archive are actually transcribing documents, most often the driving force comes from outside, mainly historians from various disciplines, such as literary, juridical, theological, technical, or

medical history. Therefore the availability of the transcribed text depends on the agreement of the person who carried out the transcription, or – as it was usually the case until very recently – with the publishing house responsible for the edition of the transcription as a book. Though this practise is now changing and more and more editorial projects are carried out online on the basis of an Open Access policy, still a large number of interesting transcriptions will be locked due to copyright restrictions.

We are confident that within the course of the tranScriptorium project we will be able to attract the interest of many archives and historians and to use this interest for convincing them to provide more data for technical research.

3. References

Antonacopoulos, A., Clausner, C., Papadopoulos, C., & Pletschacher, S. (2013). ICDAR2013 Competition on Historical Book Recognition – HBR2013 †. In *12th International Conference on Document Analysis and Recognition ICDAR2013* (pp. 1491–1495). doi:10.1109/ICDAR.2013.294

Balk, H., & Conteh, A. (2011). IMPACT : Centre of Competence in Text Digitisation. In *HIP'11, September 16 - September 17 2011, Beijing, China* (pp. 155–160).

Clausner, C., Pletschacher, S., & Antonacopoulos, a. (2011). Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments. *2011 International Conference on Document Analysis and Recognition*, 48–52. doi:10.1109/ICDAR.2011.19

Fischer, A., Frinken, V., Fornés, A., & Bunke, H. (2011). Transcription alignment of Latin manuscripts using hidden Markov models. In *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing - HIP '11* (pp. 29–36). New York, New York, USA: ACM Press. doi:10.1145/2037342.2037348

Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., Stolz, M., & Bern, C.-. (2010). Ground Truth Creation for Handwriting Recognition in Historical Documents. In *DAS '10, June 9-11, 2010, Boston,MA, USA* (pp. 3–10).

Kondermann, D. (2013). Ground Truth Design Principles An Overview. In *VIGTA '13 July 15 2013, St. Petersburg, Russia* (pp. 1–4).

Pletschacher, S., & Antonacopoulos, A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. *2010 20th International Conference on Pattern Recognition*, 257–260. doi:10.1109/ICPR.2010.72

Plötz, T., & Fink, G. a. (2009). Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, *12*(4), 269–298. doi:10.1007/s10032-009-0098-4

Rahnemoonfar, M., & Antonacopoulos, A. (2011). Restoration of Arbitrarily Warped Historical Document Images Using Flow Lines. In *2011 International Conference on Document Analysis and Recognition* (pp. 905–909). Ieee. doi:10.1109/ICDAR.2011.184

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., … Lladós, J. (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, *46*(6), 1658–1669. doi:10.1016/j.patcog.2012.11.024

Scheutz, M., & Weigl, H. (2004). Ratsprotokolle. In J. Pauser, M. Scheutz, & T. W. Winkelbauer (Eds.), *Quellenkunde der Habsburgermonarchie. Ein exemplarisches Handbuch (16.–19. Jahrhundert)* (pp. 590–610). Wien.

4. Appendix

4.1. Model agreement for access to GT

Bentham collection DATABASE SUBSCRIBER AGREEMENT

The *Bentham collection* database consists of a set of images of a collection of works on law and moral philosophy written by yhe philosopher Jeremy Bentham.

The SUBSCRIBER:

Name:

Affiliation:

Address:

e-mail:

The DATABASE is property of the University College London (UCL) "Owner" from hereon in.

The use of the information contained in the DATABASE is under the terms of this agreement.

- The DATABASE must be used exclusively for non-commercial research purposes in the field of computer science. For research purposes in other fields (e.g. philoshopy, laws, etc.), the SUBSCRIBER must contact UCL first.
- The Owner grant a license to use the DATABASE to the SUBSCRIBER and his/her research group/lab only. It is non-transferable, therefore the SUBSCRIBER agrees not to disclose nor sell any part of the DATABASE to a third party.
- The SUBSCRIBER takes no ownership rights in the DATABASE. The SUBSCRIBER agrees not to duplicate the DATABASE, nor to create a derivative work of the same.
- The SUBSCRIBER is allowed to include the information of the DATABASE in scientific publications, including the following statement:

Any use of the DATABASE not expressly authorized in this Agreement is strictly prohibited.

By signing this agreement, the SUBSCRIBER accepts the aforementioned terms and conditions,

Appendix: Basic concepts in tS

This table provides an overview of the basic concepts for GT production in tS. The table is an updated version of a working paper that was provided to the partners during spring 2013 and discussed among the partners. The basic idea of this table was to define basic concepts within the project in order to enable a common understanding of the GT task.

NAME		
COLLECTION		
	Definition in tS	A number of documents as they are usually put together in an archive. E.g. the Bentham collection consists of hundreds of single documents.
	Definition in PAGE	n.a.
	Definition in TEI	<teiCorpus> contains the whole of a TEI encoded corpus, comprising a single corpus header and one or more TEI elements, each containing a single text header and a text.
	Relevance	GT documents will be available on collection basis if applicable.
DOCUMENT		
	Definition in tS	In the case of the tS project documents are mainly understood as “manuscripts” which form a discrete unit and can be distinguished from similar documents by their content, appearance or authorship.
	Definition in PAGE	n.a.
	Definition in TEI	Every piece of text may be a TEI document.
	Relevance	For GT purposes it is important to have a large number of metadata available on document level for identification as well as processing purposes.
FOLIO or LEAF		
	Definition in tS (from Wikipedia)	"folio" is used in terms of page numbering for most manuscripts that are bound but without page numbers as an equivalent of "page" (both sides), "sheet" or "leaf", using recto and verso to designate

		the first and second sides, and (unlike the usage in printing) disregarding whether the leaf concerned is actually physically still joined with another leaf.
	Definition in PAGE	n.a.
	Definition in TEI	n.a.
	Relevance	Not relevant for HTR and GT but it plays a role for the design of the Transcription Platform since the way to count folios (recto-verso) instead of pages is common to many archives and collections.
PAGE		
	Definition in tS (from Wikipedia)	A page is one side of a leaf of paper. It can be used as a measurement of documenting or recording quantity
	Definition in PAGE	The main unit.
	Definition in TEI	A page is defined as the text between two page breaks.
	Relevance	A page is the most important unit for aligning text with page images.
BORDER		
	Definition in tS	Follows PAGE
	Definition in PAGE	Border of the actual page (if the scanned image contains parts not belonging to the page).
	Definition in TEI	n.a.
	Relevance	Important for document understanding.
	Comment	In manuscript digitisation it is common sense not to crop a page, but to leave (black) borders for indicating the actual end of the page.
PRINTSPACE		
	Definition in tS	The effective area on a page where the main part of the text can be found.
	Definition in PAGE	Determines the effective area on the paper of a printed page. Its size is equal for all pages of a book (exceptions: titlepage, multipage pictures). It contains all living elements (except marginals) like body type, footnotes, headings, running titles. It does not contain pagenumber (if not part of running title), marginals, signature mark, preview words.
	Definition in TEI	n.a.

	Relevance	Might be of some relevance for document understanding, but currently not exploited in the tS project.
MARGIN		
	Definition in tS (from Wikipedia)	A margin is the area between the main content of a page and the page edges. The margin helps to define where a line of text begins and ends. The top and bottom margins of a page are also called "head" and "foot".
	Definition in PAGE	n.a.
	Definition in TEI	n.a.
	Relevance	Might be of some relevance for document understanding, but currently not exploited in the tS project.
REGION		
	Definition in tS	Follows PAGE
	Definition in PAGE	All elements of interest are considered to be represented in a region. he most important region types are text, image, line drawing, graphic, table, chart, separator, maths, noise and frame. Regions may be covered as a rectangle, isothetic polygon or as a polygon.
	Definition in TEI	Regions are similar to the <zone> element in TEI. <zone>defines any two-dimensional area within a surface element.
	Relevance	Regions play an important role for the document understanding and segmentation task in tS.
	Comment	Unfortunately PAGE does not make a difference between “region” as a graphical unit and “paragraph” as a logical unit. This may cause some problems since TEI understands paragraphs as logical units.
TEXTREGION		
	Definition in tS	Follows PAGE
	Definition in PAGE	Pure text is represented as a text region. This includes drop capitals, but practically ornate text may be considered as a graphic. Text regions may contain line regions, words and glyphs.
	Definition in TEI	See above.
	Relevance	The relevance is high since it should not happen, that any content gets lost from a page.

OTHERREGIONS		
	Definition in tS	Follows PAGE.
	Definition in PAGE	There are several other region types in PAGE which are mainly important for printed documents.
	Definition in TEI	Other elements are encoded e.g. as “figure”, “formula” or similar.
	Relevance	Other regions do not play a role in tS since the focus is clearly on text recognition.
READING ORDER		
	Definition in tS	Follows PAGE
	Definition in PAGE	The reading order determines the logical relationship between the elements of a page or document. To express a reading order between elements they have to be included in an OrderedGroup. Elements which do not fit into a reading order (e.g. running titles) can be included in an UnOrderedGroup.
	Definition in TEI	Reading order is only implicitly defined in TEI by the sequence of the TEI document itself.
	Relevance	Reading order is highly relevant for document understanding but may not play a big role in tS.
PARAGRAPH		
	Definition in tS (from Wikipedia)	A paragraph is a self-contained unit of a discourse in writing dealing with a particular point or idea.
	Definition in PAGE	A paragraph is the default type of a text region.
	Definition in TEI	The paragraph is the fundamental organizational unit for all prose texts, being the smallest regular unit into which prose can be divided.
	Comment	See above “region”.
LINE		
	Definition in tS	Defined as in PAGE. In more detail: In tS we allow the overlapping of LineDrawingRegions, due to the nature of handwritten texts and their overlapping characters.
	Definition in PAGE	A line is not explicitly defined in PAGE but understood as a LineDrawingRegion which can be seen as the area that contains all glyphs of a given line. The LineDrawingRegion can be defined as rectangle, isometric polygon or polygon.

	Definition in TEI	n.a.
	Relevance	Together with the baseline this is the most important feature for GT in HTR. Currently in HTR processing LineDrawingRegions are the common input to the HTR engines. The correct distinction of these regions is therefore of high importance.
BASELINE		
	Definition in tS (from Wikipedia)	The baseline is the line upon which most letters "sit" and below which descenders extend.
	Definition in PAGE	Multiple connected points that mark the baseline of the glyphs. The baseline can therefore take the form of a poly-line.
	Definition in TEI	n.a.
	Relevance	The baseline is another important feature for GT in HTR.
	Comment	The baseline was introduced to the PAGE format on request of the tS project.
WORD		
	Definition in tS (from Wikipedia)	A word is the smallest element that may be uttered in isolation with semantic or pragmatic content (with literal or practical meaning).
	Definition in PAGE	not defined
	Definition in TEI	<w>(word) represents a grammatical (not necessarily orthographic) word.
	Relevance	For HTR processing word segmentation is not necessary.
	Comment	Since within tS some attempts are also made to word spotting on the basis of pre-segmented words some GT is also produced on word level.
LANGUAGE		
	Definition in tS	Follows TEI
	Definition in PAGE	Important concept that is used within PAGE but not explicitly defined.
	Definition in TEI	<language>characterizes a single language or sublanguage used within a text.
	Relevance	Since in tS a lexicon based approach for HTR is used, languages play an important role.
SCRIPT		

	Definition in tS (from Wikipedia)	A script is a distinctive writing system, based on defined elements or symbols.
	Definition in PAGE	"Script" is used but not explicitly defined. A list of basic scripts is provide.
	Definition in TEI	In TEI script is not an element but only an attribute for the "handFeatures".
	Relevance	Only of general relevance for HTR. Currently not used.
IMAGE (DIGITAL FACSIMILE)		
	Definition in tS	An image represents usually a single page within tS.
	Definition in PAGE	Important concept but not explicitly defined.
	Definition in TEI	<facsimile>contains a representation of some written source in the form of a set of images rather than as transcribed or encoded text. There are several attributes to describe images in more detail.
	Relevance	Obviously GT in tS must contain (correct) text and related images.
DOCUMENT FEATURES		
	Definition in tS	n.a.
	Definition in PAGE	PAGE describes several document features, such as headlines, marginalia, page numbers, footnotes, etc.
	Definition in TEI	TEI has a large feature set, this is one of its core competences.
	Relevance	In tS only basic document understanding is carried out, therefore advanced features do not play a role.
	Comment	In order to cope with TEI elements that are not relevant for HTR processing, it is expected to include TEI elements in the "custom" element of PAGE.