Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments*

C. Clausner, S. Pletschacher and A. Antonacopoulos

PRImA Lab, School of Computing, Science and Engineering, University of Salford,
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract - **Large-scale digitisation has led to a number of new possibilities with regard to adaptive and learning based methods in the field of Document Image Analysis and OCR. For ground truth production of large corpora, however, there is still a gap in terms of productivity. Ground truth is not only crucial for training and evaluation at the development stage of tools but also for quality assurance in the scope of production workflows for digital libraries.**

This paper describes Aletheia, an advanced system for accurate and yet cost-effective ground truthing of large amounts of documents. It aids the user with a number of automated and semi-automated tools which were partly developed and improved based on feedback from major libraries across Europe and from their digitisation service providers which are using the tool in a production environment.

Novel features are, among others, the support of top-down ground truthing with sophisticated split and shrink tools as well as bottom-up ground truthing supporting the aggregation of lower-level elements to more complex structures. Special features have been developed to support working with the complexities of historical documents. The integrated rules and guidelines validator, in combination with powerful correction tools, enable efficient production of highly accurate ground truth.

I. INTRODUCTION

The digitisation of printed documents has been subject for many years now and is yet far away from being at an end. Driven by the ever increasing need for the accessibility of information, huge efforts are being made to transform the heritage of books, newspapers and other documents into modern digital media. The sheer amount of the targeted material has led to new challenges in the area of Document Image Analysis. Methods have to be highly efficient and reliable to be competitive in the context of mass digitisation.

The process of document analysis mainly comprises Image Enhancement, Layout Analysis (region segmentation and classification), Optical Character Recognition (OCR) and various post processing steps (e.g. lexicon based correction). Numerous systems have been developed over previous decades, all claiming to be the best for specific tasks. This stresses the demand for means for objective comparison of different approaches and ways to uncover individual strengths and weaknesses. Evaluation systems are the answer; by comparing automatically generated data of the method in question against the ideal results (Ground Truth), they produce meaningful performance indicators and quality measures.

The production of ground truth is a crucial part of any evaluation system. Only if the ground truth data is accurate and complete, are the evaluation results reliable. Document datasets selected for ground truthing have to be representative, realistic and as large as possible, especially in the context of mass digitisation. One of the main goals of libraries, archives and other institutions involved in large-scale digitisation is cost effectiveness. Ground-truthing tools therefore have to be highly efficient, easy to use, reliable, flexible and accurate at the same time.

Ground truth for layout analysis evaluation encompasses the definition and labelling of regions (or zones) within a document page, logical relations between these regions and possible further information such as metadata and text content (for OCR text evaluation). Several systems for ground truthing have been reported in the literature, mostly seemingly intended for research and small datasets. The most prominent of the past approaches is Pink Panther [1], a combined system for ground truthing and document layout evaluation. Although it already included many good concepts (polygonal regions, flexible metadata, partial reading order), it has never been developed further. Another well-known system is TRUEVIZ [2]. Apart from regions it also supports text lines, words and glyphs together with their textual content. Nevertheless, it only supports rectangular region outlines which are insufficient for more complex document layouts. The system PerfectDoc [3] incorporates advanced features such as multi page reading order and semantic annotations but regions can only be represented by bounding boxes and document images have to be in JPEG format (often regarded as inadequate due to lossy compression). A rather recently reported system is the GEDI Ground Truth Editor [4]. Although it comprises an up-to-date XML format and many sophisticated features, its Java based user interface leaves room for improvements regarding usability and efficiency. Especially for larger document images (> 30 mega pixels) the tool becomes quite unresponsive (refresh rate < 5 frames per second on a system with a quad core CPU at 2.4 GHz, 3 GB RAM and a screen resolution of 1920*1200 pixels). The also Java based PixLabeler [5] allows labelling on pixel level. The provided drawing tools seem suitable for marking regions efficiently, but apart from different labels, no further data (attributes or text) can be entered for elements. The authors plan to support layers in future versions.

* This work has been supported in part through the EU 7th Framework Programme grant IMPACT (Ref: 215064).

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.19

II. THE SYSTEM

In this paper we present the ground-truthing system *Aletheia* (from the Greek word for 'truth'). Is has to be noted that it is not just an incremental update of the tool reported by the authors in [6]; it is rather a complete redesign.

Aletheia has been partly developed in context of IMPACT (IMProving ACcess to Text) [7], a major EU-funded project aiming at improving technologies for mass digitisation of historical documents, comprising libraries and universities across Europe. This can be regarded a significant advantage, as feedback from professional ground-truthing service providers had considerable influence on the development. The main target of the system is efficiency, accuracy, flexibility and usability, all with regard to large-scale evaluation.

Aletheia is part of a complete performance analysis infrastructure, encompassing XML formats, ground-truthing and evaluation tools, validators, converters, viewers and more. Ground truth and segmentation results are represented according to a sophisticated XML schema which is component of the PAGE (Page Analysis and Ground truth Elements) Format Framework [8]. It enables an accurate and detailed definition of layout regions by supporting polygonal outlines and a rich set of attributes and metadata. Furthermore, high-level relations between regions can be modelled through features as reading order, layers and hierarchies. Text regions can be further structured into text lines, words and glyphs and have fields for text ground truth or OCR results. The XML schema is flexible, extensible and can be regarded as mature. It is used for all IMPACT related datasets and a large collection of contemporary documents, presented in [9].

III. GROUND-TRUTHING STRATEGIES AND TOOLS

Ground truthing page layout always starts with a document image. Aletheia supports the usage of colour/grey-scale and black-and-white images at the same time (the user can switch between the two views). If a black-and-white image is provided, a connected component analysis is run internally. The components are stored in an efficient run-length encoding and are fundamental for a number of features included in Aletheia. Integrated binarization algorithms also allow the generation of black-and-white images on the fly.

The tool uses a Multi Document Interface (MDI) aiming to meet user expectations concerning usability and look-and-feel. It is highly customisable through a range of user-defined settings and it fits in seamlessly into the multiple user model of the operating system. The user interface is designed to be compliant with the Microsoft User Experience Guidelines [10].

The main viewing area of Aletheia (see fig. 1) shows the document image, overlaid with regions or other page elements, depending on the selected viewing mode (border, print space, region, text line, word or glyph). A range of zooming tools and viewing options facilitate working with documents in high detail. Floating dialogues and toolbars that can be undocked to increase available space for the document image and are beneficial for multi screen configurations.

The next sections present Aletheia's features and tools in the context of different ground-truthing strategies and applications.

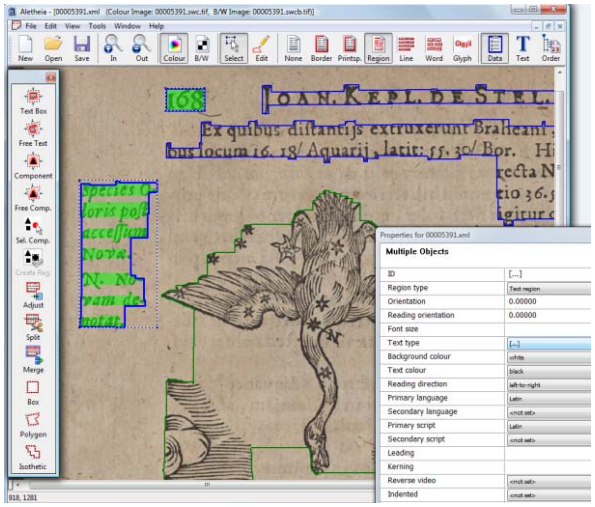

Figure 1. Work area of Aletheia and properties dialogue.

A. Top-Down Strategy

Top-down ground truthing is the process of defining page element hierarchies by starting with the highest level – regions – and obtaining lower levels (text lines, words and glyphs) by splitting parent elements. It is possible to use a pre-produced layout segmentation (by any layout segmentation method directly supporting PAGE XML or through conversion) as a starting point for ground truthing. However, it is the experience of the authors that correcting pre-produced data can be more time-consuming and error-prone than starting a completely new ground truth file. An experiment with 10 randomly selected documents from the IMPACT dataset confirms this view. Layout ground truth was produced using both approaches. Carried out by an experienced ground truther, it took 3 minutes and 10 seconds per document to finish when starting from an empty document. Correcting the results produced by ABBYY FineReader Engine 9.0 (together with an exporter to PAGE format) took 4 minutes and 32 seconds per document on average. Nevertheless, the decision which strategy to use has to be based on the chosen segmentation method and its strengths and weaknesses in context of the intended application and the quality of the material.

Layout regions are represented by unconstrained polygons. Aletheia is equipped with a range of manual and semi-automated tools to define region outlines. The latter make use of the connected component data extracted from the black-and-white image and can be divided into three categories:

1. Bounding box based outline shrinking: Usually used for text regions. The user roughly marks the region of interest (using the mouse) and an algorithm shrinks the defined outline step-by-step until it hits an obstacle (bounding box of connected component). The algorithms decision when to enter a gap between two components can be influenced by several parameters. Finally, the outline is refined to fit more closely using the run-length data of the components. This tool generates isothetic outlines with few corner points, benefitting efficient further processing (as illustrated in fig. 2).

Figure 2. Semi-automated outline generation for text regions: top: Marking a region; bottom: Resulting outline.

2. Smearing based outline generator: Usually used for images, graphics and other objects with a less regular appearance than text. Equally to the previously mentioned tool, the user firstly marks the region using the mouse. Then all connected components are determined that lie completely within the selected area. An algorithm using iterative horizontal, vertical and diagonal smearing then internally connects all components to one compact shape. Then, a tracing method is used to follow the outline of the shape and create a polygon. Fig. 3 illustrates the process.

Figure 3. Semi-automated outline generation for graphics: left: Marking a region; right: Resulting outline.

3. Component selection: This tool is closely related to the previously mentioned outline generator. The connected components that should be part of the region are directly selected by the user. The region polygon is calculated by the aforementioned algorithm based on smearing and outline tracing.

Once a region is created, it can be modified easily by moving, adding and deleting polygon points. Furthermore, mismarked regions can be corrected using wizard based merge and split tools.

Following the top-down strategy, text regions need then to be split into text lines. Aletheia provides manual and semi-automated tools for this task as well. A split can be accomplished by simply drawing a split line between two text lines in the document image. If the spacing is sufficient, this can even be achieved by only one click. A shrinking algorithm then calculates text lines with tightly fitting polygons. This process can be repeated in a similar way down to the level of glyphs.

B. Bottom-Up Strategy

The bottom-up strategy follows the concept of aggregating elements of lower levels to larger elements in higher levels. As a first step the glyphs of a document need to be marked manually or by selecting corresponding connected components and generating the outline automatically. Words can then be created by joining the corresponding glyphs. Text lines and regions can be produced accordingly. Fig. 4 gives an overview for this process.

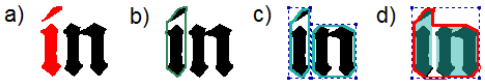

Figure 4. Creation of glyphs and words: a) Selection of components; b) Created glyph; c) Selection of glyphs; d) Created word.

C. Region Types, Attributes and Metadata

Regardless of the chosen strategy, regions need to be classified and possibly enriched with additional information. Aletheia supports eleven region types: text, image, graphic, line drawing, chart, separator, table, maths, noise, frame and unknown. For some region types there exists also a sub-type denoting the logical function of a region (e.g. for text regions: paragraph, heading, page number, drop-capital and more). Furthermore, a number of additional attributes allow the specification of very detailed information on each region. For text regions for instance, there is (among others): language, font, reading direction, background colour and text colour. All types and attributes can be changed in a properties dialogue with checkboxes, drop-down boxes and input fields. Moreover, the dialogue also allows the modification of multiple regions at the same time (for attributes the selected regions have in common), increasing efficiency drastically for documents with many similar regions. For changing the type of 10 regions to 'Heading' for instance, the mentioned feature reduces the number of required clicks from 40 to 14.

D. Text Ground Truth

Aletheia supports the input of text ground truth on each structure level of page text elements (text region, text line, word and glyph). Unicode is fully supported and encouraged, however, it is also possible to input plain ASCII encoded text in parallel. The dialogue to enter text is

equipped with a customizable virtual keyboard for special characters (see fig. 5). Aletheia includes its own TrueType font, enabling even the use of characters that are not part of the Unicode standard yet. Special characters are of great importance in the context of historical documents which often contain specific symbols, ligatures or abbreviations that are not part of modern alphabets. Initiatives such as MUFI (Medieval Unicode Font Initiative) [11] help to standardise the usage of the private use areas in Unicode.

Since text recognition is one of the major concerns in document analysis, text content can exist on all structure levels (region, line, word and glyph). Although this carries the risk of redundant data, it is necessary to achieve maximum flexibility (text ground truth might be available on glyph level whereas a specific OCR system outputs text only for whole regions). Text propagation tools have been implemented to allow moving text between different levels. If text is already available on region level, it can be propagated down to the glyphs of a page (assuming the elements of all levels exist and match the text). In addition, text of low-level regions can be used to compose the text of their parent regions.

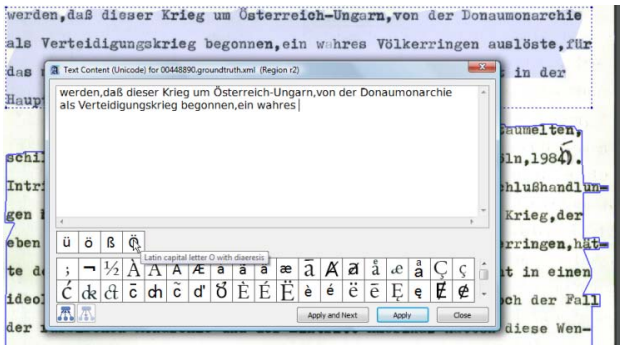

Figure 5. Text input dialogue with virtual keyboard.

E. *Logical Relations Between Regions*

Advanced evaluation systems take into account logical relations between layout regions to refine the performance analysis. Aletheia supports the definition of reading order and layers. The reading order in general describes the logical order in which the text regions of a page are supposed to be read. Most systems reported in the literature merely allow a rather limited strictly sequential order. This is insufficient for a wide range of document layouts. The reading order model used in Aletheia is based on groups of ordered and unordered elements. When placed in an ordered group, regions have sequential relation (e.g. consecutive paragraphs) whereas regions placed in an unordered group are rather loosely related (e.g. advertisements in a newspaper). Fig. 6 shows the dialogue for creating and modifying the reading order and the visualisation in the main document view.

Layers are required for document layouts with overlapping regions (e.g. stamp on text or watermark in background). No other ground-truthing system (the authors are aware of) supports layers at the point this work has been published. Fig. 7 shows the definition of layers in Aletheia for an example where layers are required to accurately describe the document layout.

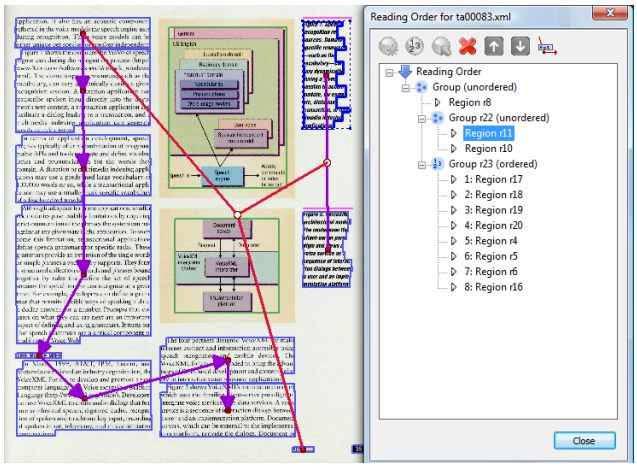

Figure 6. Reading order definition and visualisation.

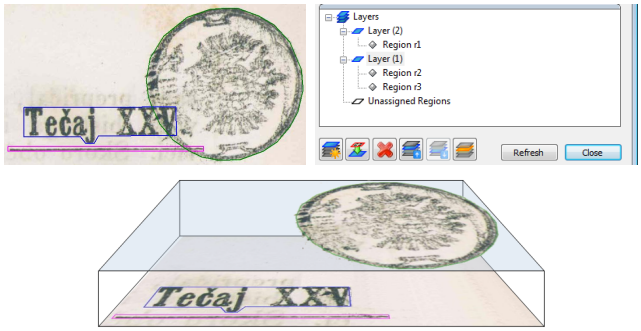

Figure 7. Layers: top-left: Stamp over text and separator; top right: Layer dialogue; bottom: Illustration of layers.

IV. Rules and Guidelines Validator

In the scope of large-scale ground truth production the need for guidelines and rules become apparent. Such rules are crucial for consistent results with a high and stable quality. A well written set of guidelines is the first important step and has been accomplished for Aletheia in the context of IMPACT. However, since ground truthing is often subcontracted to service providers, means for automated quality assurance are of great importance. This has been achieved by integrating a Rules and Guidelines Validator in Aletheia. The validator includes a variety of rules such as: Check for missing data, check for overlapping regions, reading order related checks, etc. Validation results are presented in an easy-to-navigate tree structure, allowing tracing errors to their source (i.e. the corresponding regions). Fig. 8 demonstrates the validator.

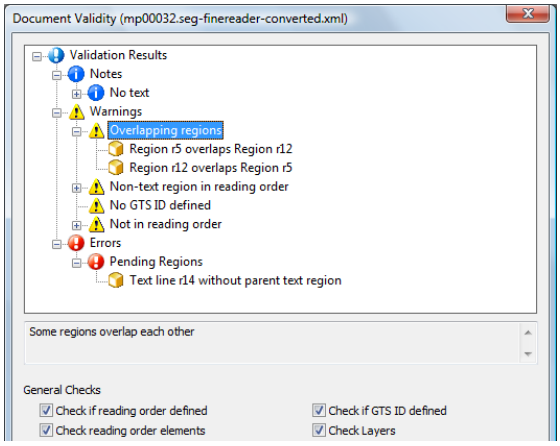

Figure 8. Rules and Guidelines Validator (results).

V. Application in Production Environments

Aletheia has been designed with regard to large-scale ground truth production. In the context of IMPACT it has been employed successfully for ground truthing over ten thousand historical documents. Major libraries all across Europe contribute to vast datasets of documents. The actual task of ground truthing is usually carried out by service provider companies, which are supplied with the document images and Aletheia. After automated and manual validation against rules and guidelines, the results (in form of PAGE XML files) are ingested into the document repository.

Valuable feedback for Aletheia has been provided by the service providers and libraries which led to further improvements of the system. A feature realised by this means is the 'Quick Open' dialogue which improves the workflow for multiple documents.

It has been experienced that most ground-truthing provider companies prefer the strategy of pre-producing segmentation by using professional layout analysis systems. Since the results of any automated approach can never meet the expectations of manually created ground truth, Aletheia has been equipped with correction tools to deal with the most frequent issues. Under-segmented regions can be corrected by a flexible split tool. A merge tool including region ordering and text preview enables straightforward correction of over-segmentation.

VI. Conclusion and Future Work

Aletheia targets to meets all expectations for a ground-truthing system designed for large-scale application. It makes use of a flexible and mature XML schema, which is used for the dataset comprising several thousands of documents in context of IMPACT. Region outlines are represented by flexible polygons, allowing accuracy and efficient further processing at the same time. The state-of-the-art multi document user interface is efficient and reliable and encompasses a variety of highly productive features (e.g. advanced zooming, undo engine), increasing the usability. This is further promoted through a large set of settings and options, enabling individual customization. Aletheia avoids constraints in ground truthing by allowing top-down as well as bottom-up strategies. An integrated validator ensures the conformity of ground truth with rules and guidelines. Moreover, the system is well documented (detailed user guide for all features and tools, as well as rulebooks and examples for ground truthing). Due to the rich set of functions, Aletheia can also be used as viewer for QA and even as a semi-automated segmentation tool for smaller projects, pilot trials, case studies and show cases.

A full repetitive development cycle is used for the system, encompassing planning, modelling, implementation and several testing stages. The usage of a version control system and bug tracking tools further underline the professional orientation. Feedback from libraries and digitisation service providers is constantly considered when planning incremental updates for the tool. Aletheia has proven its usefulness in real world applications, especially in the context of IMPACT. Furthermore, it will was made publicly available in scope of the ICDAR 2011 Historical Document Layout Analysis Competition.

Planned for future development are (among others) the support of nested frame regions and the integration of existing state-of-the-art segmentation methods.

References

[1] B. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation" Pattern Recognition, Vol. 31 (1998), pp. 1191–1204.

[2] C. H. Lee and T. Kanungo, "The Architecture of TRUEVIZ: A GroundTRUth/Metadata Editing and VISualiZing Toolkit," Pattern Recognition, vol. 36, no. 3, pp. 811-825, 2003.

[3] S. Yacoub, V. Saxena, S. N. Sami, "PerfectDoc: A Ground Truthing Environment for Complex Documents", Proc. of the 8th International Conference on Document Analysis and Recognition (ICDAR2005), Seoul, South Korea, Aug. 29–Sep. 1, 2005, pp.452-456.

[4] W. Seo, M. Agrawal and D. Doermann, "Performance Evaluation Tools for Zone Segmentation and Classification (PETS)," 20th International Conference on Pattern Recognition (ICPR), 2010, pp.503-506.

[5] E. Saund, Jing Lin, P. Sarkar, "PixLabeler: User Interface for Pixel-Level Labeling of Elements in Document Images", Proc. of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), Barcelona, Spain, July 26-29, 2009 pp.446-450

[6] A. Antonacopoulos and H. Meng, "A Ground-Truthing Tool for Layout Analysis Performance Evaluation", Proceedings of the International Association for Pattern Recognition (IAPR) Workshop on Document Analysis Systems (DAS2002), D. Lopresti, J. Hu and R. Kashi (Eds.), Springer Lecture Notes in Computer Science, LNCS 2423, pp. 236-244.

[7] IMPACT (IMProving ACcess to Text), http://www.impact-project.eu/

[8] S. Pletschacher, A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260.

[9] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009), Barcelona, Spain, July 2009, pp.296-300.

[10] Microsoft User Interaction Guidelines, http://msdn.microsoft.com/en-us/library/aa511258.aspx

[11] MUFI (Medieval Unicode Font Initiative), http://www.mufi.info/