Ground Truth Data for Document Image Analysis

Glenn Ford **George R. Thoma**
Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland

Abstract

The ground truth data described here is collected from the production operation of MARS (Medical Article Records System), a system combining scanning, OCR, document image analysis and lexical analysis techniques. Developed by an R&D division of the National Library of Medicine (NLM), MARS automatically extracts bibliographic data from paper-based biomedical journals to populate the Library's flagship database, MEDLINE®, used worldwide by biomedical researchers and clinicians. The bibliographic data extracted include the article title, author names, institutional affiliations and abstracts.

This ground truth data includes document images, OCR output and operator-verified data at the page, zone, line, word, and character levels. It is accessible online via a public website to enable researchers to develop innovative and efficient algorithms for automatic zoning (page segmentation), labeling (field identification), lexical analysis techniques to correct OCR errors, and techniques for reformatting syntax to adhere to established conventions. In addition, we offer a tool (Rover) to visually compare the results of such programs to the ground truth data. The ground truth and results data are in XML, and Rover is written in Java. The overall website development uses MacroMedia Dreamweaver UltradDev 4 to provide a rich interface and extensive database connectivity.

1 Introduction and objective

Research in document image analysis is greatly dependent on ground truth data for the design, training and testing of algorithms for data identification and extraction. However, ground truth datasets and their associated analysis and visualization tools are usually created to analyze problems in specific applications and datatypes: skewed document images (Okun et al.)[1], handwritten documents (Cha and Srihari)[2], video sequences (Doermann and Mihalcik)[3], statistical data (Swayne et al.)[4], and speech signals (Barras et al.)[5]. Moreover, apart from the domain-specific nature of these datasets and tools, they are usually limited as to operating platforms and data formats, as described in an excellent taxonomy on this subject by Kanungo et al.[6]

To our knowledge no ground truth dataset exists that represents the corpus of biomedical journals, and none with the goal of extracting the text representing the bibliographic fields descriptive of the articles within these journals. Such bibliographic data is extracted automatically from scanned biomedical journals by the Medical Article Records System (MARS) built and operated at the National Library of Medicine to populate its flagship database, MEDLINE®. MARS relies on rule-based algorithms for automatic zoning, labeling and reformatting.[7-9]

In the course of normal production operations, MARS generates vast amounts of document images and OCR-converted and operator-verified data. This data, as ground truth, would be invaluable in the design of innovative and efficient algorithms for automatic zoning, labeling and reformatting by the computer science and medical informatics communities. To aid in this effort, we are developing MARG (*Medical Article Records Groundtruth*), a ground truth database accessible via the Web.

This ground truth data includes page, zone, line, word, and character level information. In addition to providing a public site for researchers worldwide to develop and test their algorithms, we propose a tool to enable them to graphically visualize the ground truth data and employ an automated analysis assistant. Code-named Rover (*g**RO**undtruth **V**s. **E**ngineered **R**esults*), this automated analysis assistant will compare the results of a researcher's program to the ground truth data. The ground truth and results data are in XML, and Rover will be written in Java. The overall website development uses MacroMedia Dreamweaver UltradDev 4 to provide a rich interface and extensive database connectivity.

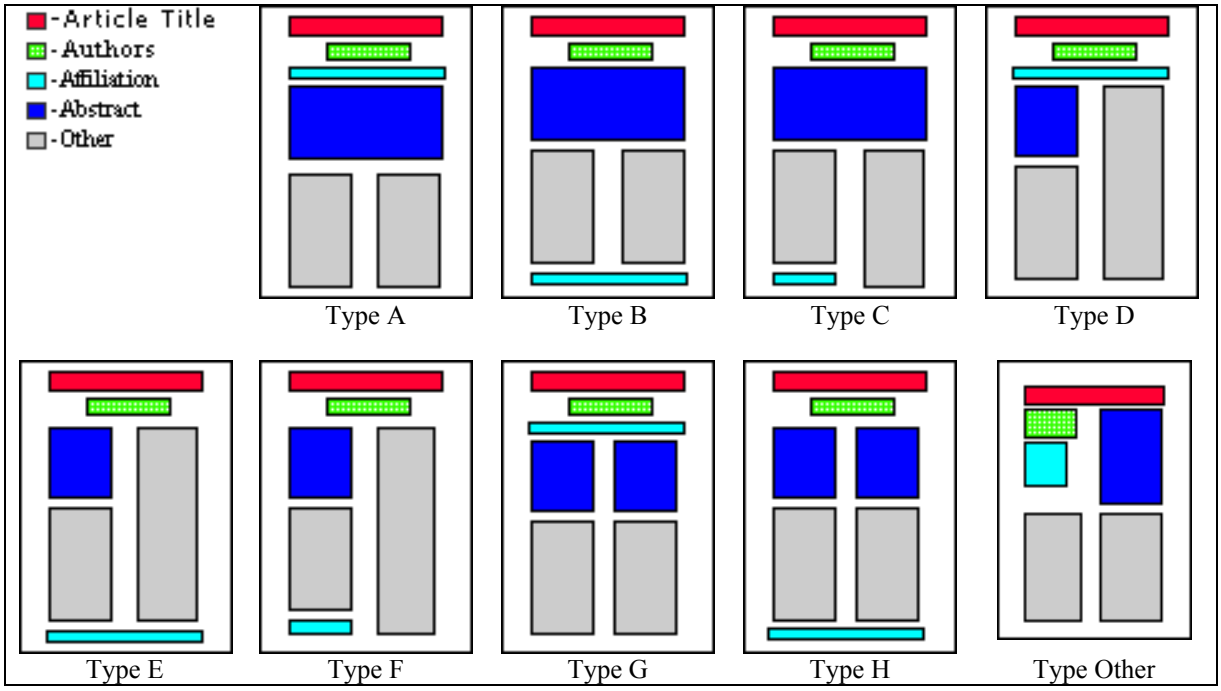

Figure 1 Page Layout Types

2 Design considerations

2.1 Page layout

Identifying geometric features to design algorithms for automated data extraction is a non-trivial task since there is a variety of layout geometries in the 4,300 journal titles indexed in MEDLINE, though most follow the reading order paradigm of article title-author names-author affiliation-abstract. Ground truth data representative of the corpus of biomedical journals must therefore include samples of all significant layout types.

The MARG site contains examples of 9 layout types (Figure 1) grouped according to the placement of the Article Title, Author(s), Affiliation, and Abstract. The general description each type is as follows:

- Type A - The Title, Author, Affiliation and Abstract appear in the defined order and are located in the upper half of the page.
- Type B - The Title, Author and Abstract are in the upper portion of the page. The Affiliation is located at the bottom.
- Type C - The Title, Author and Abstract are in the upper half of the page. The Affiliation is single columned and located in the left column of double column text. Variations include: body of text below the double column affiliation area that is material not relevant to MEDLINE citations.
- Type D - The Title, Author, and Affiliation are in the upper half of the page. The Abstract usually is in the first column. Variations include the abstract continuing into a portion of the second column.
- Type E - The Title, Author, and Affiliation are in the upper half of the page. The Abstract is single-columned, but above the body text of the article in most cases.
- Type F- The Title and Author are in the upper half of the page. The Affiliation is along the bottom-left. The Abstract usually is in all or some of the first column. Variations include the abstract continuing into a portion of column 2.
- Type G - The Title, Author and Affiliation are in the upper half of the page. The Abstract is in two adjacent columns.
- Type H - The Title and Author are in the upper half of the page. The Affiliation is along the

bottom. The Abstract is double columned, but above the body text of the article in most cases.

- Type Other - This category, holding all unusual layouts encountered, account for approximately 23% of the journal collection. Layouts in this category have not been categorized further as yet.

2.2 Distribution of page layouts

Once the layouts were classified in the types shown in Section 2.1, it was of interest to determine their frequency of occurrence. Figure 2 shows a distribution of these defined types.

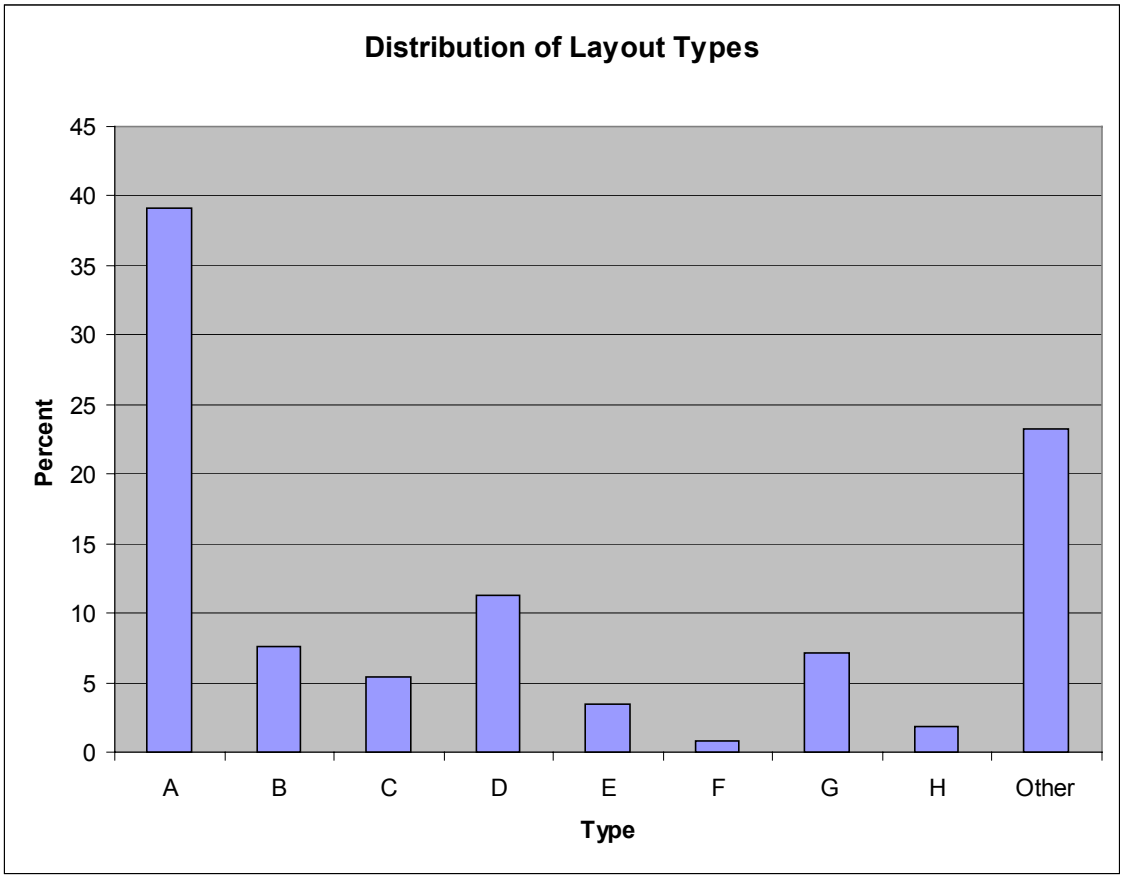

Figure 2 Distribution of layout types

2.3 Data format

The format chosen for all data is XML: for images, OCR-converted data and operator-verified data since XML excels in adaptation (to accommodate changes in data), maintenance, linking (from one piece of data to another), simplicity, and portability over networks, operating systems and development environments[10-12].

2.4 Modifying existing data

Despite its undeniable value, the existing data has certain deficiencies. For example, although OCR output characters in error are corrected as part of the text verification stage in production, their attributes (e.g., *italics* or **bold**) are not. But these attributes can serve as features to create algorithmic rules to correctly identify zones or labels. Attribute information will be included in subsequent releases of the ground truth dataset.

3 Rover: a visualization and analysis tool

Once the data is available in XML format along with the original TIFF image, researchers need the functionality to visualize and analyze the data. Specifically, the following functions are needed:

- Load a TIFF image and the corresponding XML file into an application, and display the XML data, where appropriate, overlaid on the image.
- Modify and add new ground truth data.
- Compare the results of new algorithms against the ground truth data.

The first two functions are provided by TrueViz, described in a survey of visualization tools[6]. TrueViz is a public domain tool developed at the University of Maryland for visualizing, creating and editing ground truth and metadata. It is implemented in Java and works on Windows and Unix platforms (specifically, it has been successfully tested in the Windows 2000 and Sun Solaris 2.6 environments). It reads and stores ground truth and metadata in XML format, and reads the corresponding TIFF images. It allows the user to inspect ground truth data at many levels, viz., at the page, zone, line, word, and character levels, and provides pertinent information at each level. For example, at the character level, such information includes the character code, font type and style, and bounding box (x1,y1,x2,y2) coordinates.

TrueViz is a suitable platform to initiate the design of Rover, an analysis assistant that will offer all three functions mentioned above. To serve as an effective analysis assistant, Rover will extend TrueViz to provide researchers the capability to *compare* their XML data to ground truth XML data graphically (Figure 3), and thereby help them iteratively refine their algorithms. In addition, Rover will provide statistics and visual presentations on specified areas. For example, the user would be able to use Rover to compare all characters in the dataset that are **bold,** and to enumerate errors. Rover would visually locate the mistakes and report statistics on the query. In this example it would report the number of bold characters, the number detected correctly, and the number detected incorrectly, both as absolute numbers and as percentages. Rover would also export this information to a database or spreadsheet for further analysis.

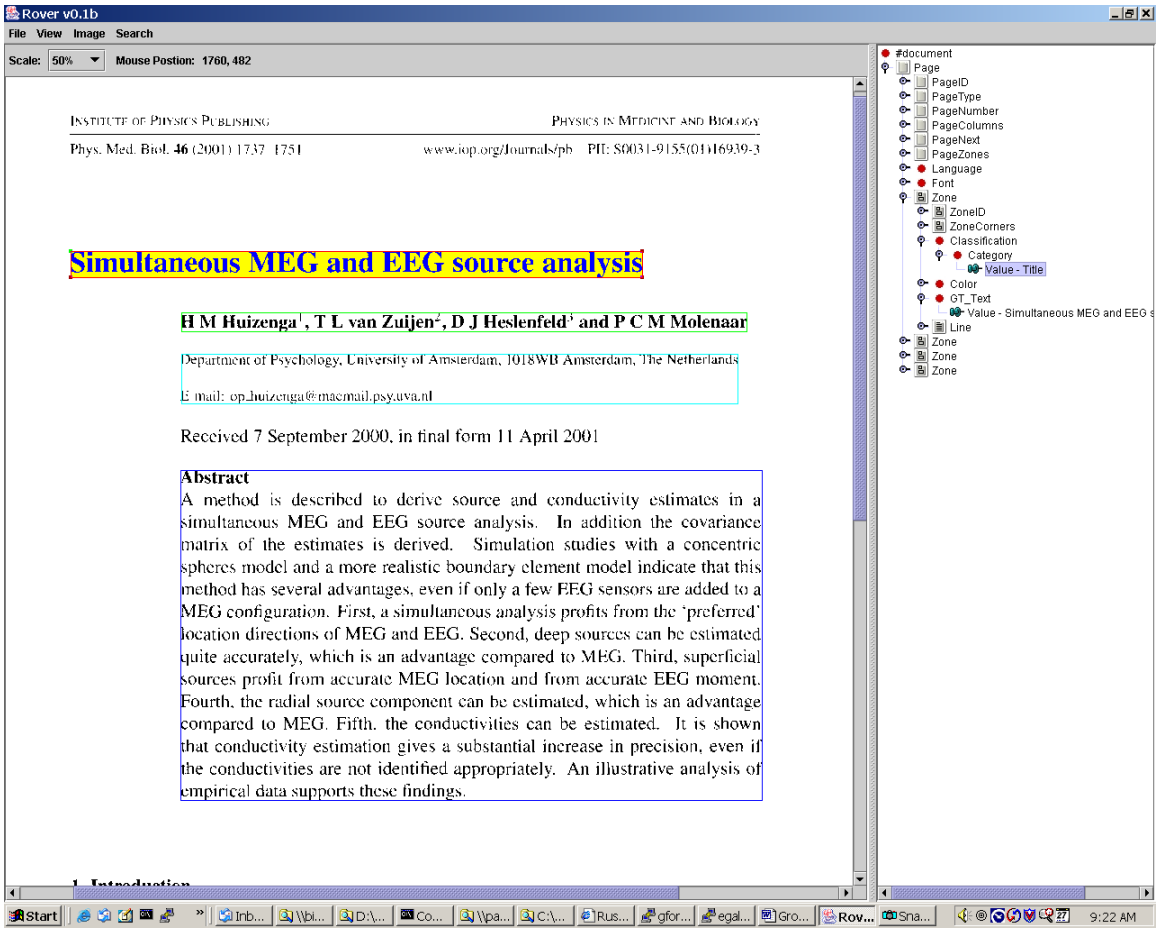

Figure 3 Rover screen snapshot of TIFF page and zone segments.

4 Ground truth website

A website provides access to the ground truth data, though it will go beyond serving as a simple data repository. Our objective is to encourage researchers to share and exchange ideas, as well as provide feedback to our development team for new desirable features. Figure 4 shows the organization of the website and how the elements interconnect. This section presents an overview of the website layout and functionality.

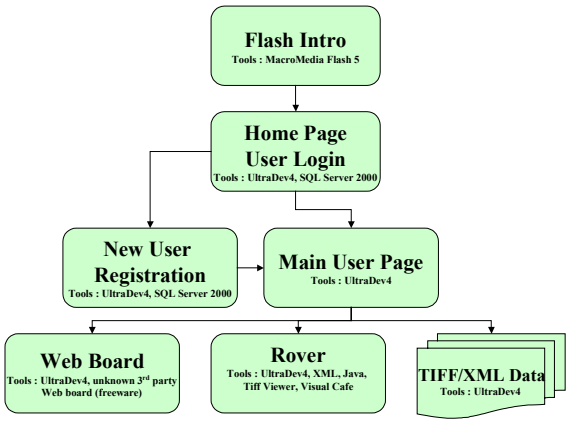

Figure 4 Main sections of website and tools

4.1 Product introduction

A Macromedia Quick Flash "movie" is displayed to the users, though this may be bypassed. The metaphor used is a Sherlock Holmes style magnifying glass moving over a rare biomedical document.

4.2 User login

On the main page of the website, users are asked for a name and password. This data is stored in a SQL Server 2000 database along with the website data. We use Macromedia Ultradev as the design tool, because it readily allows for password protection and database connectivity. While we do not anticipate restricting anyone from accessing this site, we intend to require all users to register. When registered users log in, they will be taken to the main user page. Unregistered users will be presented with a registration page before being allowed access.

4.3 New user registration

The system will require first time users to enter a few important items of information that will enable us to provide security to the system, access to the Bulletin Board, and provide the ability to contact users with important new data or features added to the website in the future. The registration page was developed using Ultradev, which allows the developer a graphical design environment and provides rapid page development.

4.4 Data access

Once logged in or registered, the user would have access to all the ground truth data and tools. Since the data collection will be quite large, the system will allow users to download the entire data set or any subset they choose. For example, some users may only be interested in certain types of journal layouts, such as double column abstracts.

4.5 Data analysis

The users will be provided a link to launch Rover. While the initial version of this tool will possess the current functionality of TrueViz, the complete analysis support discussed earlier will be designed in a later phase.

4.6 Bulletin board

This is for the user community to report bugs, and use as a forum for discussions related to algorithmic development. Here the users may also upload and download files, such as technical papers written, algorithm source code, and new ground truth data.

5 Summary

We describe here a system for the distribution of ground truth data collected from the production operation of MARS, a system for the automated extraction of bibliographic data from scanned biomedical journals. This ground truth data includes document images, OCR output and operator-verified data at the page, zone, line, word, and character levels. It is accessible online via a public website to enable researchers to develop innovative and efficient algorithms for automatic zoning (page segmentation), labeling (field identification), lexical analysis techniques to correct OCR errors, and techniques for reformatting syntax to adhere to established conventions.

References

1. Okun O, et al. An experimental tool for generating ground truths for skewed page images. Proc. SPIE Document Recognition and Retrieval VIII. Vol. 4307, San Jose CA, January 2001, 22-33.

2. Cha S-H, Srihari SN. Handwritten document image database construction and retrieval system. Proc. SPIE Document Recognition and Retrieval VIII. Vol. 4307, San Jose CA, January 2001, 13-21.

3. Doermann D, Mihalcik D. Tools and techniques for video performance evaluation. Proc. 15th International Conference on Pattern Recognition. Barcelona, Spain, September 2000, 167-70.

4. Swayne DF et al. Xgobi: interactive dynamic data visualization in the X window system. Journal of Computational and Graphical Statistics 7, 1998.

5. Barras C, et al. Transcriber: a free tool for segmenting, labeling and transcribing speech. Proc. 1st Int. Conf. Language Resources and Evaluation. Granada, Spain, May 1998; 1373-76.

6. Kanungo T, et al. TRUEVIZ: A groundtruth/metadata editing and visualizing toolkit for OCR. Proc. SPIE Document Recognition and Retrieval VIII. Vol. 4307, San Jose CA, January 2001, 1-12.

7. Thoma GR. Automating the production of bibliographic records for MEDLINE. Internal R&D report, CEB, LHNCBC, NLM; September 2001; 92pp. Available: archive.nlm.nih.gov/pubs/biblio/biblio.php

8. Thoma GR. Document image analysis and understanding R&D. Internal R&D report to the Board of Scientific Counselors, CEB, LHNCBC, NLM;

October 2001; 55pp. Available: archive.nlm.nih.gov/pubs/biblio/biblio.php

9. Kim J, Le DX, Thoma GR. Automated Labeling in Document Images. Proc. SPIE, Vol. 4307, Document Recognition and Retrieval VIII, San Jose CA, January 2001, 111-22.

10. Desai G and Fenner J. Unleash the power of XML. Imaging and Document Solutions, Vol. 9, No. 12, Dec 2000, 29-32.

11. Pearson G, Moon CW. Bridging two biomedical journal databases with XML: A case study. Proc. 14th IEEE Symposium on Computer-Based Medical Systems. Los Alamitos, CA: IEEE Computer Society. July 2001, 309-14.

12. World Wide Web Consortium, XML Working Group, "XML 1.0 Specification 10-February-1998", ed. T. Bray, J. Paoli, C.M. Sperberg-McQueen. Available in various forms: www.w3.org/TR/1998/REC-xml-19980210.