Dataset Submission Form

I. Overview – Message from TC-10 and TC-11

It is extremely important for the Document Image Analysis and Recognition community to be able to cross check and reproduce results described in published papers in the field. In order to achieve this, any datasets used as the basis for publications should be publicly available, as is the norm in many other disciplines.

Authors are actively encouraged to submit the datasets they used to train and/or evaluate their algorithms to their TC(s) in order for them to be published on the corresponding Web sites.

This initiative is not restricted to datasets. We are interested in archiving online any piece of data (ground-truth data, software, *etc.*) which would allow to easily reproduce results, set new targets, foster healthy competition, encourage collaboration and generally advance the DIAR field as a whole.

The Web site of TC-10 is http://www.iapr-tc10.org.
The Web site of TC-11 is http://www.iapr-tc11.org.

II. Submission Protocol

The process of submitting a dataset is the following:

1. Please fill in the form below, and send it by email to Dimosthenis Karatzas (dimos@cvc.uab.es), the TC-11 dataset curator, or to Bart Lamiroy (bart.lamiroy@loria.fr) the TC-10 dataset curator.
2. The dataset curators will review the submission request, and ensure that all information is clear and complete, and any copyright issues are properly addressed.
3. As soon as all information is in place, they will ask you to sign and return by fax the final submission form.
4. The dataset curators will work with you to upload the dataset to the TC Web site. Depending on the nature of the dataset this might be as easy as sending a CD or directly uploading the required files.

The TCs are actively working towards a more comprehensive way of dealing with datasets and associated information.

III. Copyright Note

Both TC-10 and TC-11 provide dataset hosting services as a benefit to the international research community. If it is determined that copyrighted material is improperly included in a dataset submitted to inclusion on their website, the offending material will be immediately removed upon notification by the copyright holder.

By submitting a dataset for inclusion to the TC-10 or TC-11 Web site, the author certifies that he/she has the right to publish the dataset and any associated data in the public domain and the act of doing so does not violate intellectual property rights or copyrights of some third party.

The TCs will provide a service through which the submitted dataset and any associated data will be made public to the Document Analysis community worldwide. In case any legal dispute arises in the future in relation to the publishing of this dataset and associated data in the public domain, the author will hold TC-10 and TC-11 free from any wrongdoing and accept responsibility for the publication of these data.

By submitting a dataset and associated data, you explicitly accept that any third party can independently submit additional information that relates to the original dataset (*e.g.* additional ground-truth data, software, *etc*).

We strongly encourage the authors, where they own the copyrights of the submitted information, to consider offering it to the community under a creative commons license. See the links below for further information:

http://wiki.creativecommons.org/Before_Licensing
http://creativecommons.org/choose/

IV. Useful Definitions

Dataset: A collection of data along with *metadata* information, as required to use these data.

Metadata: *Metadata* is information specific to a particular *dataset*. *Metadata* are usually tightly structured within the *dataset* itself (e.g. information encoded within the filenames of submitted images). *Metadata* can only be submitted at the time of submission of the *dataset*.

Interpretation Specification (Ground Truth Specification): The definition of the required information that accurately describes a particular aspect of the data at a high level where reasonable agreement between different human observers can be established, as well as the definition of an appropriate structure (format) for storing this information.

Interpretation (Ground Truth) Data: A set of data conforming to a particular *interpretation specification* and relating to a specific *dataset*.

Task: A well defined process to evaluate algorithms in the context of a specific scientific problem. A *task* would typically provide a specific evaluation protocol, and link to specific resources as required (a *dataset*, and usually related *Interpretation data*).

Resources: Any other type of related *resources* that are not specifically covered by the above definitions. Examples would include software to browse and visualise a dataset, software to create Interpretation data, algorithms to do performance evaluation, codecs, reports, publications, etc.

V. Submission Form

The submission form has multiple sections. Please fill in the sections applicable to your situation. Not all sections need to be filled in. It would be helpful to have a look at already published datasets to get an idea of the information that is needed (http://www.iapr-tc11.org/mediawiki/index.php/Datasets). Feel free to use as much space as you need, but generally a couple of paragraphs are more than enough to describe every aspect of the submitted material.

1 Submission Information

The following choices are offered to quickly target the general domain of the submitted data, as to have it correctly referenced and reviewed. As such, they are mainly convenience choices. They have no immediate impact on accessibility or availability of the data, once hosted.

1.1. This submission is of interest to: multiple choices may apply.

- ☐ TC-10 Graphics Recognition
- ☐ TC-11 Reading Systems

1.2. I would like this submission to be reviewed by:

- ☐ TC-10 Graphics Recognition
- ☐ TC-11 Reading Systems
- ☐ Don't care

2 Dataset Information

2.1. Are you submitting a new dataset? It is possible to only submit Interpretation data for an existing public dataset. If this is the case then just fill in the title and Web address of the existing public dataset in this section.

◉ Yes ○ No

2.2. Title and Acronym. Please provide a title and acronym for the dataset. This is how your dataset will be identified on the TC-10 or TC-11 Web site. Avoid generic titles and prefer titles that would identify the dataset uniquely like "The University of Duckburg Dataset of Scanned Tables".

2.3. Keywords. Please provide a list of keywords that will be used to categorise and search for your dataset.

2.4. Description. Write a short description (a couple of paragraphs max) of what is included in the dataset. Importantly, include information about how the dataset was constructed and the data collected. For example if 100 scanned pages are submitted, it is important to know if they are sourced from a single book, or they are 100 pages from 100 distinct books. If data were collected from human subjects describe the task they were given, and any statistics that you find relevant (*e.g.* nationality, mother tongue, age groups, gender etc).

2.5. Metadata and Technical Details. Please describe here any metadata that will be submitted along with the dataset and how they are encoded. Any naming conventions or internal structure used should be described here. Also, include information (as applicable) regarding the number of samples in the dataset, the file formats used etc. Also, please include

information about the total size of the dataset in question, this will help us identify the best way to upload the data.

3 Interpretation Information

3.1. **Do you intend to submit Interpretation (Ground Truth) Data along with the dataset**? If you do, then please fill in this section otherwise proceed directly to the next one.

○ Yes ◉ No

3.2. **Title.** Please provide a short title for your Interpretation data (*e.g*. "Skew Angles for the ACRONYM dataset" or "Layout and Transcription for the ACRONYM dataset").

3.3. **Keywords.** Please provide a list of application domains / research tasks where these Interpretation data could be useful for (e.g. layout-analysis, character recognition, text detection, word spotting).

3.4. **Description.** Please provide a short description about the Interpretation data submitted. When Interpretation data were created manually, the ground-truthing protocol/guidelines should be included here. In the case of synthetic data, any parameters used should be listed here. Also describe the Interpretation format used to store the data and if applicable include a sample file. Please include information about the total size of the Interpretation data.

4 Research Tasks Definition

4.1. **Do you want to define any research Tasks based on the dataset and/or Interpretation data you have submitted?** We strongly encourage you to define as a research task the research question you set out to answer in your related publications. If you do not want to define any tasks, then please proceed directly to the next section.

○ Yes ◉ No

4.2. **Task Title.** Please provide a title for the research task you propose to the community (e.g. "Character Recognition in Typewritten Historical Documents").

4.3. **Task Description.** A short description of the task you propose: what is the purpose of this, its importance to the future of the community, the state of the art at the time of defining this, citations to any existing publications treating this problem with this or other datasets, etc.

4.4. **Evaluation Protocol.** When defining a research task, you should define explicitly the evaluation protocol to be used to assess the results of algorithms implemented for this task, so that comparisons of different contributions can be made. This could be as simple as defining training and test sets and suggesting the use of well known metrics, or you could instead propose new metrics and providing supporting software to facilitate evaluation. If any type of evaluation software is provided for this task, please provide detailed technical information on the expected output formats needed for the evaluation.

5 Software

5.1. Do you intend to submit any accompanying software?

○ Yes ◉ No

5.2. Description. Describe briefly the functionality of the submitted software. Consider sending a thumbnail of the software along with the form.

5.3. Technical Details. Please include information regarding minimum requirements, operating system needed, any special installation instructions etc.

6 Other Material

6.1. **Related Publications.** Please provide a list of published contributions that relate in some way to this dataset (they introduce/describe the dataset; they make use of it for training or evaluation etc).

6.2. **Submitted Files.** Please provide a list of files that you submit along with this form. These could be samples from the dataset, sample Interpretation files, screenshots of the submitted software, etc. We strongly advice you to submit sample files for every aspect of the submission you describe in this form. Also try to send screenshots or other representative images that we can use in the Web site to illustrate the material submitted.

7 Copyright Ownership

7.1. **Contact Author.** Give the name and contact details of the person responsible for this submission. This information will be published online along with the material submitted.

7.2. **Do you own the copyright to all the material you have submitted?**

○ Yes ◉ No

7.3. **If you answered No to the question above, please let us know who owns the copyright and why you think you are able to place this data in the public domain (e.g. if the material is already publicly available under a creative-commons license).**

Signature _____

Name _____

Address _____

Date _____