Laboratoire lorrain de recherche
en informatique et ses applications

Year 1 Progress Report for PICS DIA-Tribe (2015)

Ref: PICS06758

Bart LAMIROY, Université de Lorraine, Loria (UMR 7503)
Daniel LOPRESTI, Lehigh University, Bethlehem, PA

Context

The project proposal, as submitted and accepted, aimed at leveraging a large collection of office documents at Lehigh University: business correspondence, technical white papers, notes, blueprints, measures and plots to bootstrap an international network initiative for Document Image Analysis (DIA) reproducible research benchmarking. Our goal is to provide an answer to the difficulty of correctly assessing progress in DIA as a scientific discipline and to reply to a broad demand of benchmarking data and tools. The data collection has the unique features of being diverse in the types of documents it encompasses; yet sufficiently coherent. It spans the whole spectrum of current DIA research.

The goals set in the proposal consist of :

1. Digitization and Preservation of Large Graphical Documents
2. Hosting and Distributing of the Data Collections
3. Enhancing Data Interaction and Providing Benchmarking Tools
4. Developing Performance Evaluation Metrics

Achievements

The work plan for the first funded year, defined in the intial proposal, was as follows :

Done 1. *Initial meeting at Lehigh University (February) to select small initial subset of documents for digitization, to determine first annotation and mark-up requirements and to set goals for first workshop.*

- B. LAMIROY and D. LOPRESTI met in February 2015 in San Francisco, while attending DRR XXII (Document Retrieval and Recognition)
- B. LAMIROY spent 1 week at Lehigh University with D. LOPRESTI in July 2015

In Progress 2. *Digitize, annotate and make documents available on existing DAE platform.*

During B. LAMIROY's stay at Lehigh University in July 2015, he and D. LOPRESTI selected and conditioned approximately **900 large engineering drawings** from the 4000 available. These documents have been conditioned and prepared for shipping and subsequent digitization.

Lack of **funding** has temporarily delayed the process. Funding for shipping and digitization has been secured in September 2015 through IEEE proceeds from the

Campus scientifique
BP 239 – 54506 Vandœuvre-lès-Nancy Cedex
Tél. : 03 83 59 20 00 – Fax : 03 83 27 83 19 www.loria.fr

GREC workshop (*cf*. item 4. below) and we are in the process of receiving the actual funds.

Partial Achievement

3. *Select one of more international contests announced for the IAPR International Conference on Document Analysis and Recognition (ICDAR 2015 – August, Tunisia) and adapt the dataset and the contests' performance evaluation metrics to the DAE platform.*

   Terrorist attacks (Bardo Museum and Sousse hotel resort) severely disrupted the organization of the ICDAR 2015 edition, to the point of the conference being relocated to Nancy, France on a very short notice. D. LOPRESTI was part of the conference steering committee, and B. LAMIROY was general chair of one of the satellite worshops (GREC – IAPR International Workshop on Graphics Recognition) as well as being the local arrangements chair for the entire set of workshops being relocated from Tunisia to France.

   Given these circumstances, it was decided to postpone this task to the next edition of ICDAR in 2017 (Kyoto, Japan).

   Part of the goals were addressed during the GREC workshop (*cf*. item 4. below) especially work by B. LAMIROY and P. PIERROT on developping **robust metrics** in absence of formal ground truth.

   Independently, B. LAMIROY has started a **M.Eng. thesis project** (student O. MEKKI) on designing an annotation process that would fit within the framework of crowd-sourcing as a GWAP (*Game With a Purpose*). First results expected in February 2016, proof of concept platform by Fall 2016.

Done

4. *Organize a first workshop at ICDAR 2015 and showcase the initial (partial) dataset and metrics and establish first network exchanges around the data collection.*

   In collaboration with the *Université de La Rochelle* and the Computer Vision Center at the Autonomous University of Barcelona, a call for contributions was launched and a special session was organized at the **GREC workshop**. Contributions and findings discussed during the workshop are currently under review for publication by Springer as an **LNCS volume** in 2016.

   Connected to the goal of establishing a network around the data collection, B. LAMIROY was the lead coordinator of an **ANR proposal** (consortium of Lehigh University, *Université de La Rochelle*, the Computer Vision Center at the Autonomous University of Barcelona and the *Université François Rabelais* in Tours) called DIA-Tribe. This proposal was selected for the second evaluation round, but was rejected at the final phase. Since the proposal received excellent grades and comments, it was resubmitted for the 2016 call.

References

Program for 2016

The initial proposal stated the following goals for year 2:

Maintained

1. Proceed with extensive selection and digitization of 200-300 documents and making them available.

We have already started and selected approximately 900 documents. All documents are conditioned and ready to be shipped for digitization. Funding is secured and we are awaiting transfer of funds.

We have access to appropriate infrastructure for making the data available once digitized.

Conditioned 2. Extending the DAE platform to a less centralized and more community-oriented, distributed solution.

We curently have a platform that is operational and has been on line for the last 5 years without noticeable incidents or flaws. The platform needs to be adapted for use in a more distributed environment.

Preliminary (upublished) work has been done related to the integration of NoSQL based storage support. It is estimated that 6 person months at an undergraduate level would be sufficient to develop a running proof of concept.

The achievement of this goal therefore is conditioned to the funding of 6 months of internship at the French legal rate (*indemnités de stage au taux légal*).

Maintained 3. Based on the experience of adapting experimental protocols of ICDAR contests, investigate new data interaction processes, and formalization of certified metrics.

This task has already started and reported before (M.Eng. O. MEKKI) with the work on GWAP based crowd-sourced annotation. This research will continue and extended.

Maintained 4. One or two extensive exchange stays between Loria and Lehigh University, most likely through graduate students working on the project topics.

No graduate students being currently enrolled on these topics, exchange stays will very likely concern the principal investigators B. LAMIROY and D. LOPRESTI or possibly the undergraduate students working on the program.

Open 5. Publication of results at the International Conference of Pattern Recognition (ICPR 2016, Cancun, Mexico)

This goal remains open, and will depend on progress reported during the project.

Funding Request

We maintain the funding request for year 2, especially given the fact that one of our initially stated goals is strongly conditioned by the funding of internship stipends, two major international events (Document Analysis Systems – DAS, in Greece and ICPR in Mexico) and the need for two extended (1 week) exchange stays.

	1st year (requested)	**1st year (funded)**	**2nd year (requested)**
Travel	5,000.00 €	5,000.00 €	5,000.00 €
Exchange stays			2,800.00 €
Seminars - Workshops	1,500.00 €		
Other	1,800.00 €		3,800.00 €
Total	8,300.00 €	5,000.00 €	11,600.00 €