Laboratoire lorrain de recherche
en informatique et ses applications

Year 2 Progress Report for PICS DIA-Tribe (2016)

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

Previous Achievements

During the first funded year (2015), we made progress on, or fully achieved the following goals, defined in the initial proposal (*cf.* 2015 progress report):

1. *Digitize, annotate and make documents available on existing DAE platform.*
2. *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.*
3. *Organize a first workshop at ICDAR 2015 and showcase the initial (partial) dataset and metrics and establish first network exchanges around the data collection.*

Announced Program and Results for 2016

At the end of year one we amended the initial program, based on the progress made in 2015, and the following goals were set year 2:

1. P*roceed with extensive selection and digitization of 200-300 documents and making them available.*

Partial Success

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

800 documents have been retrieved, transferred from Lehigh University to Loria and have been digitized. From this viewpoint the goals have been exceeded.

During D. Lopresti's stay in Nancy, 4-6 July, we conducted some quality tests with the digitization equipment of Stocomest, a local company. These tests were conclusive, but administrative regulations required us to have at least three competing offers. After some further testing, we contracted with the lowest bidder (Ingecap in Metz).

The quality of the 800 digitized documents was very disappointing and currently insufficient for publication and/or further global use. We have decided to amend our work program for 2017 (*cf*. next section)

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

Delayed to 2017

We currently 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.

As mentioned in the 2015 report. This task requires considerable engineering efforts, and was conditioned by the funding of at least 6 months of internship at the French legal rate (*indemnités de stage au taux légal*). We failed to get this funding.

However, following the DAS conference in Santorini, Greece[1], we have engaged in exchanges and discussions with the Université de Fribourg in Switzerland (M. WÜRSCH and M. EICHENBERGER-LIWICKI working on the DIVAServices platform). During D. LOPRESTI's stay in Nancy, we invited M. WÜRSCH and worked out plans for collaborating on merging the DAE and DIVAServices approaches. This has led to a joint application for funding on a specific call issued by the SATT Grand-Est. Results are to be announced first quarter of 2017 (*cf*. next section).

3. *Investigate new data interaction processes, and formalization of certified metrics*.

Fail + New Ideas

The core part of this was entrusted to a M.Eng. student (O. MEKKI) with the objective to develop a GWAP based crowd-sourced annotation framework. This task failed due to lack of involvement of the student. This research will be started over in 2017 (*cf*. next section).

However, work on the formalization of metrics has created spin-off research. In collaboration with the IECL Laboratory (UMR 7502) in mathematics and Dr. E. KOUDOU, we have started working on the specifics of quality metrics on imprecise data. This work has obtained specific funding through a CNRS PEPS for the second semester of 2016, has obtained funding for a Ph.D. thesis starting in 2017 through the Charles Hermite Federation (FR 3198), and has given rise to a proposal to the 2017 CNRS Mastodons call.

As a consequence, the focus and work on metrics has been transferred from the current PICS and the above IECL – Loria collaboration.

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

Done

[1] Marcel Wursch, Rolf Ingold, Marcus Liwicki, "SDK Reinvented: Document Image Analysis Methods as RESTful Web Services", 2016 12th IAPR Workshop on Document Analysis Systems (DAS, p. 90-95, , 2016

No graduate students being currently enrolled on these topics, D. LOPRESTI spent one week at Loria for work on the program.

Done

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

We presented a keynote at *RRPR 2016: 1st Workshop on Reproducible Research in Pattern Recognition*, 4 Dec 2016 Cancún (Mexico). D. LOPRESTI et B. LAMIROY, "*The DAE Platform: a Framework for Reproducible Research in Document Image Analysis*"

This work is currently being reviewed for an LNCS publication in 2017.

Work Program for 2017

In the initial submitted work program, we had listed for 2017 is listed below in italics. Given the evolution and achievements listed in the previous section we propose this program evolves as described below:

Refocus

1. *Pursuit of development of metrics, integration of state-of-the-art interaction tools for selecting, querying and automating annotation of the experimental data (both locally in Loria and Lehigh, and through exchange stays of graduate students)*

Given the lessons learned from the failure of implementing a GWAP, we intend to investigate this topic anew with new student teams, both at Lehigh University through their local Montaintop Experiences program, and in Loria with a team of Télécom Nancy PIDR students (already selected).

Maintained

2. *Publicizing a fully distributed and open version of the DAE platform, including the fully digitized set of selected documents and their annotations.*

As mentioned before, this task requires significant software engineering resources, and a significant part of its success will depend on the associated funding we will be able to leverage, particularly through the SATT Grand-Est grant mentioned before. Furthermore, part of it will be conducted in collaboration with the *Université de Fribourg.*

Another issue that has arisen concerns the quality of the digitization. We will be investigating two solutions. The first consists in leveraging the large level of redundancy and the relative consistency in the data to develop automated rectification and filtering algorithms. The second will consist in digitizing the dataset again with higher quality equipment. This will actually benefit the overall evaluation method and opens new perspectives for expressing evaluation metrics, since this will provide us with complementary experimental evaluation data.

Shifted

3. *Formal validation of developed metrics and experimental protocols and their implementation for the ICDAR 2017 Conference contests.*

Given the spin-off research collaboration with the IECL Laboratory and E. KOUDOU and the subsequent resources provided through the PEPS, possible Mastodons funding (acceptance results due Jan. 15) and the starting PhD. Grant, this topic will be essentially

handled in this context.

4. *Organization of a final workshop at ICDAR 2017 (Kyoto, Japan)*.

Maintained

D. LOPRESTI et B. LAMIROY are part of the organizing committee of the proposal "Workshop on Open Services and Tools", together with M. WÜRSCH and M. EICHENBERGER-LIWICKI from the *Université de Fribourg.* Acceptance results are due by Jan. 15 2017.

Funding Request

	1st year (requested)	**1st year (funded)**	**2nd year (requested)**	**2nd year (funded)**	**3rd year (requested)**
Travel	5,000.00 €	5,000.00 €	5,000.00 €	5,000.00 €	6,000.00 €
Exchange stays			2,800.00 €		
Seminars - Workshops	1,500.00 €				
Other	1,800.00 €		3,800.00 €		4,000.00 €
Total	8,300.00 €	5,000.00 €	11,600.00 €	5,000.00 €	9,000.00 €

We slightly reorient the initially presented funding request for year 3 we would like a full funding of 9,000 € covering the following costs:

- travel expenses for a stay at Lehigh University in July (allowing for a transition of the 1st semester work done in Loria on tasks 1. and 2. with the expected summer program at Lehigh University);
- travel expenses for meetings and coordination with the *Université de Fribourg*;
- participation at ICDAR 2017 in Kyoto, Japan, in November for the intended presentation and contribution to the Workshop on Open Services and Tools;
- an exceptional 4,000€ funding (> 20%) for the second digitization of the dataset (*cf.* mentioned difficulties).