The DAE Platform: A Framework for Reproducible Research in Document Image Analysis

Bart Lamiroy[1(✉)] and Daniel P. Lopresti[2(✉)]

[1] Université de Lorraine – Loria (UMR 7503), Campus Scientifique – BP 239, 54506 Vandœuvre-lès-Nancy Cedex, France
bart.lamiroy@loria.fr

[2] Packard Laboratory, P.C. Rossin College of Engineering and Applied Science Computer Science and Engineering, Lehigh University, 19 Memorial Drive West, Bethlehem, PA 18015, USA
lopresti@cse.lehigh.edu

Abstract. We present the DAE Platform in the specific context of reproducible research. DAE was developed at Lehigh University targeted at the Document Image Analysis research community for distributing document images and associated document analysis algorithms, as well as an unlimited range of annotations and "ground truth" for benchmarking and evaluation of new contributions to the state-of-the-art.

DAE was conceived from the beginning with the idea of reproducibility and data provenance in mind. In this paper we more specifically analyze how this approach answers a number of challenges raised by the need of providing fully reproducible experimental research. Furthermore, since DAE has been up and running without interruption since 2010, we are in a position of providing a qualitative analysis of the technological choices made at the time, and suggest some new perspectives in light of more recent technologies and practices.

1 Introduction

The issue of reproducibility is a fundamental tenant of scientific research. It forms the basis by which a field advances, and fosters a research community that is both competitive and yet collaborative. Advances are made only when it is possible to build on trustworthy work that has come before. Despite the overall high quality of the research being conducted within the international pattern recognition community, and a general awareness of the importance of good scientific practice, our field has recently come to realize there is room for improvement. This observation was one of the motivating factors behind the First Workshop on Reproducible Research in Pattern Recognition, held in conjunction with ICPR 2016 in Cancun, Mexico [1].

Concerns about reproducibility are not limited to our research community; they have also arisen in other fields over the past several years, most famously in US biomedical research [4,6]. In their influential policy statement as leaders of the National Institutes of Health, Collins and Tabak write that a "*growing chorus*

B. Kerautret et al. (Eds.): RRPR 2016, LNCS 10214, pp. 17–29, 2017.
DOI: 10.1007/978-3-319-56414-2_2

of concern, from scientists and laypeople, contends that the complex system for ensuring the reproducibility of biomedical research is failing and is in need of restructuring" [6]. While they note that lack of reproducibility is rarely due to scientific misconduct, the system has evolved to push (or entice) researchers away from good practices. The list of failings they quote, while drawn from a completely different domain, could easily be adapted to apply to the field of pattern recognition. In particular, they call out [6]:

1. The need to make strong (perhaps unjustifiable) claims to get published in the top venues.
2. Missing technical details when papers are published.
3. Bad practices in experimental design, including improper "blinding, randomization, replication, sample-size calculation ..."
4. The use by some scientists of a "secret sauce" (their words) in getting their experiments to work which they fail to reveal in publications to preserve a competitive advantage.
5. Inaccessible and/or proprietary data used in published works.

Collins and Tabak also point out the difficulty in publishing negative results that identify flaws in previously accepted theories. Without the benefit of a publication at the end of the tunnel, few scientists will engage in the hard work of trying to confirm or disprove the outcomes of others. They also assert that datasets are a valuable intellectual contribution in their own right that should be citable.

Their list of simple reasons that work may fail to be reproducible, while drawn from another domain, should also sound familiar to those working in pattern recognition: "*different animal strains, different lab environments or subtle changes in protocol*" [6]. All of these forms of bad behavior have analogs in our field as well. We may think of using different test collections, different implementations of a standard machine learning technique, or different approaches to computing and reporting performance measures.

The DAE platform, to be discussed in this paper, is our attempt to address some of these issues [10,13]. In particular, DAE provides an open environment for researchers to "publish" their algorithms and their data, and to document and to serve the data used in past experiments so that new techniques can be compared relative to old ones.

Collins and Tabak propose a number of clear steps for addressing the shortcomings they see in biomedical research, including the adoption and enforcement of better experimental practices. We can take such steps in the pattern recognition community, too, many of which involve changes in the social processes we use (*e.g.* the standards by which papers are reviewed and accepted for publication). The DAE server provides some of the functionality they envision for a Data Discovery Index (DDI) [3].

The remainder of this paper is organized as follows: first we will redefine the notion of *Reproducible Research* in the context of Document Image Analysis and the relations it establishes with a broader concept of Open Research (Sect. 2). Section 3 develops the functional architecture of the DAE platform, and explains

how it addresses reproducible research, and offers solutions to several of the points raised in the previous sections. Finally, Sect. 4 considers the handling of more complex notions like "Truth" and reference interpretations.

2 Our Definition of Reproducible Research

Reproducible research is not only about documenting processes, availability of experimental data and software, benchmarking and performance evaluation. Is also (and essentially) a comprehensive process of interaction with information that is certified to be reliable, of traceability and provenance, accountable re-use, recycling and re-sampling of pre-existing sources, leading to better practices overall. We already developed these issues in [16].

Research goals for work in pattern recognition and machine perception generally consist of:

- developing algorithms that are robust and approach human levels of performance for specific tasks of interest;
- inventing new methods that are better than known techniques;
- generating experimental results that are well-documented, understandable in context, and reproducible by others;
- building on past knowledge to yield new insights moving us toward solutions for problems of vital importance.

For each if these goals the sections below contain a number of observations regarding general practices and their impact of the global quality of research outcomes.

2.1 Robustness and Human Levels of Performance

In *developing algorithms that are robust and approach human levels of performance*, we want algorithms to be general. This goal expresses the need for perception algorithms to perform well on tasks containing operational conditions or expected results that are difficult to formally define [9] and for which one expects the resulting algorithm to perform equally well on new, previously unseen data.

Measuring this robustness, or reproduce published results in controlled conditions is a challenging task, and often relies on benchmarking using reference data sets. One can observe, however, that too often methods are either tested on small, overused datasets, or – especially in the context of recent deep learning developments – in extremely large datasets that make it difficult to assess the breadth of scope they capture. As a result, many experimental results reported in the literature suffer form the intimate knowledge of the data the algorithm developers have acquired over time and therefore introduce a quite strong bias towards the specifics of the benchmarks. The result is that large segments of current practices lack convincing evidence of generality.

Furthermore, the notion of "human levels of performance" is not quite well defined. In many a situation human experts can disagree on all but the most

trivial of cases [9]. This, combined with the natural bias towards the benchmark data described above, means that performance metrics and conclusions from reference data should systematically undergo scrutiny and analysis with respect to the limitations of generality they induce.

While in itself this is not a restriction to reproducibility, it does raise the question of how well it influences (positively or negatively) the emergence of new, robust and assessable approaches and ideas, or on the other hand, may tend to push research towards niche problems.

2.2 Improve upon Known Techniques

The previous section relates to the aim of *inventing new methods that are better than known techniques*, and especially to the question of knowing whether we have succeeded [17]. We already mentioned the fact that the need to compare against previously published results creates over-reliance on standard datasets (which is counter-productive). Notwithstanding, comparing new approaches with previously published ones, under the same conditions and with the same data, is still important for the assessment of progress and ongoing improvement. However, attempts to re-implement a published algorithm are often problematic due to incomplete descriptions or even the inherent conflict of interest that arises when attempting to show that one's own methods improve upon existing techniques [19]. As we already pointed out in [13,16], there are many opportunities for improvement:

1. Access to source code and data used in reported results are part of the basic requirements, but do not solve everything. Code tends to be dependent on technological environments and context, and frequently becomes obsolete (due to API changes of dependency libraries; evolution of the standard version of compilers, interpreters, and frameworks; *etc.*). Notwithstanding, code repositories and initiatives like IPOL [2] are important contributions to reproducibility. There are however situations when source code cannot be made available, or where the execution environment and resources are as important as the code itself. In that case, having access to executable binaries or complete packaged virtualized environments [12,21] can supply supplementary or complementary tools for reproducibility.
2. Even though access to source code or executables is helpful in many cases, a complete description of experimental protocols is essential to guarantee that one measures comparable results when evaluating whether new approaches improve upon the state-of-the-art. This includes, besides the data, the selection criteria, pre-filtering *etc.*
3. Another way of measuring improvement over known techniques is the use of recurring, open competitions. In order for these competitions to fully accomplish their goals, they should be frequent, have consistent and well documented evaluations protocols and metrics, and maintain records over time and subsequent editions. This implies a significant investment of resources by community in question.

2.3 Well-Documented Reproducible Results and Knowledge to Build New Advances

Generating experimental results that are well-documented, understandable in context, and reproducible by others is a real challenge and requires carefully thought out and sufficiently explicit protocols. Some of the difficulties raised here have already be mentioned in the previous sections. It is important that published results clearly establish, describe and provide all relevant data (parameters, description of data selection and filtering process, post-processing, *etc.*) making it possible for others to reproduce experiments under the same conditions.

In many cases, the explicit and/or implicit bias in selecting and using data (*e.g.*, discarding hard cases) makes the full experimental context difficult to recover. Furthermore, "Publish or Perish" mindsets lead to overstated claims and a poor understanding of the generalizability of the published results.

One of the essential by-products of reproducible open research is that it simplifies *building on past knowledge to yield new advances* toward solving problems of importance. This is generally what drives experimental research. However, if all previously enumerated conditions for open reproducible research are not met, it often becomes quite difficult to know if the efforts dedicated to developing methods actually improve upon existing techniques, or if they are well suited for the task at hand. In other cases, much time is risked to be spent "reinventing the wheel." Again, "Publish or Perish" pressure often leads to precipitation that leaves insufficient time to think and construct upon previously existing achievements, very much like trying to build a pyramid out of shifting sand without first forming it into blocks. Fixing this will require a radical paradigm shift within the community, and like the NIH position paper by Collins and Tabak discussed earlier [6], here we propose some steps in that same direction.

2.4 Reference Data and Truth

One of the corollaries of aiming for human performance levels and clearly describing experimental conditions is that there can be no such thing as "ground truth" [9]. Rather, there is the intent of the author (which is hard to determine, although sometimes we have it) or the interpretation arrived at by a reader of the document [8] (which could be a human or an algorithm) within her personal reference frame or context. Subsequently, there may be no single right answer – interpretations may naturally differ – although for some applications, we expect that users who are fluent will agree nearly all of the time.

While this may seem to contradict commonly accepted approaches to the annotation of data and the verification of algorithms, our conviction is that maintaining a *status quo* on unique reference annotations is hindering broad and open extensible or reusable research.

On the other hand, it raises a number of practical questions. With multiple interpretations, how should we proceed in developing new methods that mimic a fluent human expert? Some may be more careful, fluent or expert than others; worse, this can depend entirely on context! How can this be handled in the

context of attempting to describe a reproducible experiment? We have suggested elsewhere that on-line reputation, as originally derived in social networking, can determine whose interpretations to trust and in what context [16]. Use of reputation is one key feature of the new paradigm we would like to promote. This is beyond the scope of this paper, however. Section 4 does address some of the more structural issues of handling multiple interpretations and lack of absolute ground truth.

3 The DAE Platform, a Technical Overview

The DAE platform was the outcome of a 2009–2011 DARPA funded project. The acronym DAE refers to Document Analysis and Exploitation: see http://dae.cse.lehigh.edu. We first reported it in [10]. Complete operational details can be found in [12].

The platform has been running without major interruption since 2010, and hosts a variety of Document Image Analysis data sets, as well as document analysis reference algorithm implementations. Its general architecture is represented in Fig. 1.

DAE was conceived from the beginning with the idea of reproducibility and data provenance in mind. In the following sections, we more specifically analyze

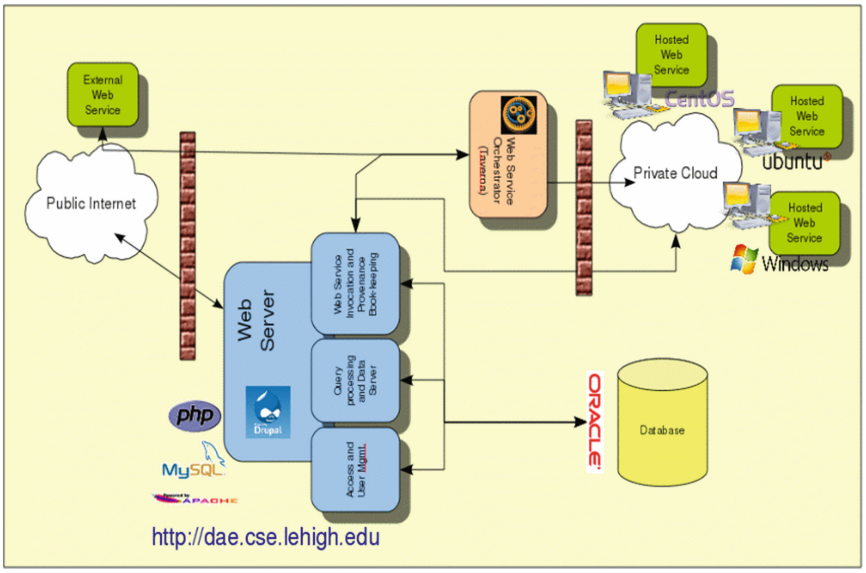

Fig. 1. General Architecture of the DAE Platform: browsing and authentication through a traditional web interface (blue – lower middle), WSDL interaction for querying and executing hosted applications (orange – upper middle); virtualized applications (green – upper right) and back-office database (yellow – lower right) (Color figure online)

how this approach answers a number of challenges raised by the needs of reproducible experimental research. Furthermore, because of the experience gathered by running the platform, we can provide a qualitative analysis of its technological impact and offer new perspectives in light of more recent technologies and practices.

3.1 General Features

From a general point of view, the DAE platform hosts a variety of data sets and algorithm implementations for document image analysis. However, the core feature of the platform is that all data is referenced in a central database on a fine-grained level. The full data model is described here [13]. We are not going to further detail the data structure, here.

This allows the platform to offer the following services:

1. It stores large data collections containing both "raw" document images, as well as an unlimited range of annotations. These annotations vary from high level content interpretations (like, for instance, text transcriptions, author identification or document structure) or pixel level segmentation information (binarization, shape outlines, ...).
2. All data can be hierarchically structured, and allows for convenient browsing through the collections.
3. It can host a wide variety of programs that interact with the data. They are intended to be reference snapshots of the state-of-the-art at some point in time. Thanks to virtualization they run in an isolated and well controlled environment. These programs are published as web services using a standard WSDL[1] interface. The WSDL API also offers SQL querying and interaction with the stored data.

These features contribute in a significant way to handling some of the more fundamental requirements for reproducible research, as shall be made clear in the next section.

3.2 Contributions to Reproducible Research

Data Is a Query: while the DAE data model is fully compatible with the more traditional approach to fixed datasets and reference annotations (or *ground truth*, for that matter) it actually has a much versatile approach to data.

As outlined in Sects. 2.3 and 2.4 data needs to conform to the following properties for efficient reproducible and open research:

1. For the sake of reproducibility, data used for published experiments should be freely and fully accessible in their exact same state as described in the referenced work; this will allow to replay and compare the results by third parties or with other approaches;

[1] https://en.wikipedia.org/wiki/Web_Services_Description_Language consulted January, 2017.

2. Since data (and more specifically their associated annotations) are open to various interpretations, contexts and possible disagreements, their annotations should be able to capture this multiplicity, while guaranteeing that every interpretation context can be accessed in a non-ambiguous and repeatable manner;
3. Data collections should be open to recomposition, extension, combination and selection in order to create new collections for other contexts and experimental setups while maintaining legacy and references to previous versions; this will enhance reuse and improvement in new contexts as the state-of-the-art evolves.

This is handled by the DAE platform by having a fine-grained data model [13]. It essentially consists of using a flat data model of *stuff*, and let users organize it in various ways. Most data stored in the platform (raw experimental data, annotations, *etc.*) is a `data_item`, and data items can be freely associated to one another or grouped (in a non-exclusive way) in collections.

As a consequence, all data, annotations, interpretations and collections are stored in such a way that they can be retrieved through well defined queries (SQL, essentially) and that, through the use of appropriate labeling, reference configurations can be frozen and their corresponding query made available as an archival reference. The platform provides a transparent mapping between URLs and queries. More detailed information is available in [13].

Software as a Service in a Controlled Environment: Since availability of experimental data is only one of the requirements for open reproducible research, algorithms and software need to be made available as well. We have made the choice that, where software is concerned, reproducibility through availability of source code is not necessarily the best guarantee for replication of results. Initiatives like IPOL [2], github and others make it possible to thoroughly describe source code such that previously published methods can be reused and reimplemented. DAE pursues another goal: benchmarking and comparison. It therefore offers the possibility to run published reference applications in a controlled environment, as close as possible to the one used at the time of publication.

These implementations are made available in a *Software as a Service* mode through WSDL interfaces (*cf.* [12] for further details). Interested parties can therefore freely launch remote executions of the software, either with their own data or with data provided by the platform itself. This approach has a number of advantages as well as some drawbacks.

- The main advantage is to offer the possibility for any type of software (regardless of complexity, programming language, or execution environment) to be run by the platform. It uses completely isolated virtualized environments, and is therefore immune to dependencies or technological obsolescence, as long as the virtualization remains available.
- It significantly lowers the barrier for contributors to provide their applications: no need for releasing source code, or to conform to specific restrictions related to supporting programming languages *etc.*

– The SAAS/WSDL approach offers the possibility to linearly scale the platform and to distribute it over multiple locations. Individual contributors may choose to host their own software, rather than upload it to the centralized DAE platform.
– The downside, on the other hand, is that the hosting facility supports the costs for all executions, rather than having them supported by the experimenter; popular applications may become a burden for the hosting facility, since execution resources are expended on the server side.

Experimental Protocols Made Explicit: As a by-product of the two previous features (data as a query/url and software as a service) it becomes easy to make experimental protocols explicit, sharable and reusable. We have explored [11] the use of web service orchestrating scripts like Taverna [20] and myexperiment.org.

3.3 Lessons Learned and Possible Upgrades

The DAE server has been up and running without interruption for more than five years. It is currently hosting 113,605 document images (totaling 287 gigabytes of data), 9 algorithms and a total of 357,925 `data_items`. We are currently in the process of uploading approximately 800 scanned technical drawings from the Lehigh Engineering Collection [5] (representing anther 400 gigabytes of raw image data). Our experience running the environment has suggested a set of possible improvements and extensions, as outlined below.

The general motivations behind these extensions consist of facilitating interactions with other, comparable or complementary initiatives, making access easier and making it possible to distribute the platform over multiple cooperating sites.

Evolution from a WSDL Interface to a REST Interface: The current architecture is based on web services with a Web Service Description Language (WSDL) interface. WSDL has the advantage of having well typed and formalized interfaces, but has the disadvantage of being synchronous. Representational State Transfer (REST) interfaces are much more flexible and have the advantage of allowing asynchronous interactions.

Services in the REST model are better-suited to the different uses of the platform than WSDL services, and will improve interactions with other services and other modern applications. This is what has been implemented by the DIVAServices platform [21], for instance.

Support for Standard Formats: There are a large number of initiatives for the annotation and structuring of data extracted from digital documents, each adapted (or adopted) according to the field of application. TEI[2], for example, is

[2] http://www.tei-c.org.

a standard used in various digital humanities projects; PAGE is a format advocated by the European project Impact and the associated European Competence Center [18]; GEDI is a format developed by the University of Maryland [7] and widely used in research circles in document analysis.

It is important for reproducibility and open research that users are not restricted to closed, proprietary formats; open, well documented exchange formats should be used.

Transition to a Distributed Data Model and Infrastructure: The platform currently relies on a back office consisting of an Oracle database and a centralized ZFS file system. In order to allow more flexible interactions with instances hosted at different sites, and thus allow easy extension and sharing of resources it would be appropriate to move to a more scalable structure of the NoSQL type.

Virtualization: One of the obstacles to the adoption of the platform is the need to use web services and encapsulate them in virtualized and safe environments. Virtualization, as deployed on the current version, is not optimal in terms of resource allocation, and will not scale well. It is therefore necessary to switch to more flexible and modernized virtualization technologies. In collaboration with the initiative at the University of Fribourg [21], we plan to employ their solution using Docker.

4 Handling Multiple Interpretations

As already mentioned before, data annotation for the verification and benchmarking of algorithms cannot be assumed to be unique [9]. The traditional approach to using *Ground Truth* for assessing the validity and performance of research generally consists of 3 phases:

- assemble a *representative* collection of reference documents;
- use human annotators to identify, select and provide the expected interpretations (we do not consider cases where test data can be synthetically generated);
- create the set of reference interpretations for the document collection (*Ground Truth*).

Performance evaluation then consists of

- providing part of the annotated data as indication and benchmark for the expected outcome (this allows researchers to define the scope of their algorithms, possibly train them, or otherwise configure them);
- running the resulting algorithms on the remaining part of reference documents (without providing associated *Ground Truth*);
- measuring discrepancy between algorithm outputs and expected *Ground Truth*;

– rank algorithms according to their measured performance.

This general paradigm is well understood, and largely adopted by the community to assess and measure the quality of the state-of-the-art. Notwithstanding, it has a number of limitations and drawbacks, some of which have already been studied [14].

1. Getting the annotations for constructing the ground truth is costly. It requires human intervention, takes time, is subject to human error. Furthermore, recent trends and techniques (essentially those based on supervised Machine Learning) tend to rely on (and require) larger and larger amounts of data. This creates a bottleneck situation.
   Crowd-sourcing has been advanced as a potential solution to this problem, but introduces issues itself. It requires a large commitment and involvement of a community, incentives and motivation are an issue and may largely affect annotation quality and reliability, extra quality control and/or processes for handling ambiguity or disagreement are needed (and may become as costly as the annotation itself) and sometimes ethical issues may arise.
2. Constituted reference annotations progressively get tainted over time: partially because they represent a snapshot of the data that was relevant at a given point in time; partially because once more and more people start to use them, it becomes more and more difficult to maintain the separation between known training data and unknown testing data. The latter eventually loses its neutral status, since it pervasively becomes known and may be used for training.
3. The way the traditional evaluation paradigm is used (very often through the organization of recurring annual or bi-annual contests) is sub-optimal in some situations, in a sense that it is unusual to see explicit loop-back mechanisms that help improving algorithms; that it is difficult to get a detailed account on how the performance of competing algorithms increases over time and sometimes *regression testing* from one contest edition to another would be useful.
4. *Ground Truth* is excessively context bound and it has been formally established [9] that it necessarily contains data that can either be considered as being mislabeled, or as being open to multiple legitimate (yet incompatible) interpretations. This induces the fact that performance evaluation and subsequent ranking may be statistically insignificant if the level of disagreement on the reference annotations is too high.

Our proposed solution consists in directly incorporating, measuring and thus leveraging the level of disagreement/uncertainty of the *Ground Truth* and actually stop calling it *Ground Truth* altogether – call it CRI: *Consensus Reference Interpretation*, for instance.

The DAE platform handles multiple concurrent annotations on the same data, and provides means to filtering, selecting and organizing these annotations. What we currently lack are the appropriate metrics and methods to efficiently handle the associated notions of fuzzy or context related "truth" and the reputation or confidence that can be associated with them [14,15].

5 Conclusion

As we pointed out in our introduction, the stakes and need for awareness regarding reproducible and open research are pervasive to all sciences. Large influential communities like the Health Sciences and Particle Physics are raising concerns and offering standards, in attempts to improve overall practices.

In this paper, we have represented the DAE Platform and highlighted its features in the context of reproducible research in a smaller and specifically targeted community. It was developed with Document Image Analysis research in mind and allows for distributing document images, associated document analysis algorithms as well as an unlimited range of annotations for benchmarking and evaluation of new contributions to the state-of-the-art.

Although DAE was conceived from the beginning with the idea of reproducibility and data provenance in mind, there are still quite a number of challenging technical developments that need to be incorporated on the one hand. On the other hand, the principal and most important challenge is to persuade large research communities that reproducible research is above all an attitude and a collection of practices that do not necessarily depend on technology, but more on collective adoption and enforcement on good practices.

Acknowledgment. The authors acknowledge support from the CNRS PICS-06758 Dia-Tribe. Early stages of this work were supported by a DARPA IPTO grant administered by Raytheon BBN Technologies.

References

1. Workshop on reproducible research in pattern recognition, Cancun, Mexico, December 2016. https://wrrpr2016.sciencesconf.org/
2. IPOL Journal - Image Processing On Line (2009)
3. Resource indexing, data science at NIH, January 2017. https://datascience.nih.gov/bd2k/funded-programs/resource-indexing
4. Alberts, B., Kirschner, M.W., Tilghman, S., Varmus, H.: Rescuing US biomedical research from its systemic flaws. Proc. Natl. Acad. Sci. **111**(16), 5773–5777 (2014)
5. Bruno, B., Lopresti, D.P.: The lehigh steel collection: a new open dataset for document recognition research. In: Document Recognition and Retrieval XXI, San Francisco, California, USA, 5–6 February 2014
6. Collins, F.S., Tabak, L.A.: Policy: NIH plans to enhance reproducibility. Nature **505**(7485), 612–613 (2014)
7. Doermann, D., Zotkina, E., Li, H.: GEDI - AGroundtruthing environment for document images. In: Ninth IAPR International Workshop on Document Analysis Systems (DAS 2010) (2010)
8. Eco, U.: The Limits of Interpretation. Indiana University Press, Bloomington (1990)
9. Lamiroy, B.: Interpretation, evaluation and the semantic gap... What if we were on a side-track? In: Lamiroy, B., Ogier, J.-M. (eds.) GREC 2013. LNCS, vol. 8746, pp. 221–233. Springer, Heidelberg (2014). doi:10.1007/978-3-662-44854-0_17

10. Lamiroy, B., Lopresti, D.: A platform for storing, visualizing, and interpreting collections of noisy documents. In: Fourth Workshop on Analytics for Noisy Unstructured Text Data - AND 2010. ACM International Conference Proceeding Series, Toronto, Canada. IAPR, ACM, October 2010
11. Lamiroy, B., Lopresti, D.: An open architecture for end-to-end document analysis benchmarking. In: 11th International Conference on Document Analysis and Recognition - ICDAR 2011, Beijing, China, pp. 42–47. International Association for Pattern Recognition, IEEE Computer Society, September 2011. ISBN: 978-1-4577-1350-7
12. Lamiroy, B., Lopresti, D.: The Non-Geek's guide to the DAE platform. In: DAS - 10th IAPR International Workshop on Document Analysis Systems, Gold Coast, Queensland, Australia, pp. 27–32. International Association for Pattern Recognition, IEEE, March 2012
13. Lamiroy, B., Lopresti, D., Korth, H., Heflin, J.: How carefully designed open resource sharing can help and expand document analysis research. In: Gady Agam, C.V.-G. (ed.) Document Recognition and Retrieval XVIII - DRR 2011, Part of the IST/SPIE 23rd Annual Symposium on Electronic Imaging. Document Recognition and Retrieval XVIII, vol. 7874. SPIE, San Francisco, January 2011. ISBN: 9780819484116
14. Lamiroy, B., Pierrot, P.: Statistical performance metrics for use with imprecise ground-truth. In: Lamiroy, B., Dueire Lins, R. (eds.) GREC 2015. LNCS, vol. 9657, pp. 31–44. Springer, Cham (2017). doi:10.1007/978-3-319-52159-6_3
15. Lamiroy, B., Sun, T.: Computing precision and recall with missing or uncertain ground truth. In: Kwon, Y.-B., Ogier, J.-M. (eds.) GREC 2011. LNCS, vol. 7423, pp. 149–162. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36824-0_15
16. Lopresti, D., Lamiroy, B.: Document analysis research in the year 2021. In: Mehrotra, K.G., Mohan, C.K., Oh, J.C., Varshney, P.K., Ali, M. (eds.) IEA/AIE 2011. LNCS (LNAI), vol. 6703, pp. 264–274. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21822-4_27
17. Lopresti, D., Nagy, G.: When is a problem solved? In: Proceedings of the 2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Washington, DC, USA, pp. 32–36. IEEE Computer Society (2011)
18. Pletschacher, S., Antonacopoulos, A.: The page (page analysis and ground-truth elements) format framework. In: Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR 2010, Washington, DC, USA, pp. 257–260. IEEE Computer Society (2010)
19. Salman, I.: Cognitive biases in software quality and testing. In: Proceedings of the 38th International Conference on Software Engineering Companion, ICSE 2016, New York, NY, USA, pp. 823–826. ACM (2016)
20. Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A.N., Balcazar Vargas, M.P., Sufi, S., Goble, C.: The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Res. **41**(W1), W557 (2013)
21. Würsch, M., Ingold, R., Liwicki, M.: SDK reinvented: document image analysis methods as RESTful web services. In: Document Analysis Systems (DAS), April 2016