# How Carefully Designed Open Resource Sharing Can Help and Expand Document Analysis Research

Bart Lamiroy[a], Daniel Lopresti[b], Hank Korth[b] and Jeff Heflin[b]

[a]Nancy Université – LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre Cedex, France
Bart.Lamiroy@loria.fr

[b]Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015
{lopresti,korth,heflin}@cse.lehigh.edu

## ABSTRACT

Making datasets available for peer reviewing of published document analysis methods or distributing large commonly used document corpora for benchmarking are extremely useful and sound practices and initiatives. This paper shows that they cover only a very tiny segment of the uses shared and commonly available research data may have. We develop a completely new paradigm for sharing and accessing common data sets, benchmarks and other tools that is based on a very open and free community based contribution model. The model is operational and has been implemented so that it can be tested on a broad scale. The new interactions that will arise from its use may spark innovative ways of conducting document analysis research on the one hand, but create very challenging interactions with other research domains as well.

## 1. INTRODUCTION

It is commonly accepted that the sharing of reference benchmark material is an essential practice in science domains where reproducible experiments play an important role in the global peer review process. Many areas and communities have structured themselves around such kind of resources and have greatly benefited from doing so.

In document analysis, there have been numerous attempts and initiatives[1–3] to produce common datasets for many of the problems that are addressed by the community. Some of them have had great impact, others less. What is certain is that all of them have had their hour of glory, and have then more or less quickly declined. Not that their intrinsic quality changed, but very often datasets have become progressively obsoleted by technology advances, new research focuses, or lack of support by their creators.

There is, indeed, a very high cost attached to the creation, maintenance and diffusion of useful research tools, datasets and benchmarks for a given community. However, they are snapshots of research topics, problems and state-of-the art at the time of their creation, unless an equivalent amount of effort goes into keeping them in line with the continuous evolution of knowledge.

Another aspect, one that is particularly related to document analysis, is that the research in this domain is very application and problem driven. Be it invoice routing, building the semantic desktop, digital libraries, global intelligence, or document authentication, to name a few, they all tend to generate very specific datasets, and produce very focused software solutions, often integrating a complete pipeline of cascading methods and algorithms. This most certainly does not affect the intrinsic quality of the underlying research, but it does tend to generate isolated clusters of extremely focused problem definitions and experimental requirements. This often even increases the cost of producing and maintaining shared available benchmarking resources, since the burden has to be supported by a very small community and cannot easily be distributed to other potential contributors. This also makes it difficult to cross borders and agree on what kinds of tools, formats *etc.* are actually the most useful ...

In this paper we address the above mentioned issues by proposing a model that tries to minimize the cost and burden to deliver and maintain datasets and benchmarking tools, keeping the model open to any kind of new contribution and even making it flexible and open to evolution with the needs of changing technology and knowledge. Our work is structured as follows: first we address the fundamental questions that arise and need

to be addressed when sharing common datasets, benchmarks and methods. We shall study them point by point thus creating the requirements a perfect platform should meet in order to appeal to and be used by the research community. We then present the our implementation choices, and give concrete and extensive examples of how it is and can be used for document analysis problems.

## 2. SCENARIO, SCRIPT AND SCREENPLAY

In order to develop our ideas, let's consider a scenario where we can have access of a well identified resource that can provide us with any kind of data related to a specific document analysis problem we would like to solve. For instance: *"Provide me with 1,000 random documents in 300dpi bitonal* `.tif` *format, predominately containing printed material and of which a reasonable number of pages contain at least one handwritten annotation."*.

### 2.1 Plot and Triggers

What we are going to describe, in essence, is the availability of a well identified, commonly available, resource (for convenience sake, let's assume this resource is centralized, we shall see further that this assumption can be greatly reduced) that offers the following services: storage and retrieval of document analysis data, complex querying of the document analysis data, collective, yet personalized markup, representation, evaluation, organization and projection of data, archival knowledge of uses and transformations of data, certified interaction with data.

The first item might, at a first glance, not look quite different from what already is available to some extent: document analysis datasets.[1–3] In the light of the second item, it proves to be completely different, however. Rather than to offer *monolithic* chunks of data and meta-data or *interpretations*, the envisioned resource treats data on a far finer grained level. This level of detail is illustrated in the scenario by the kinds of queries the system would be capable of answering: *"1,000 random documents in 300dpi bitonal* `.tif` *format predominately containing printed material and of which a reasonable number of pages contain at least one handwritten annotation."*

Furthermore, data representation (not only the document images, but extracted meta data, annotations and *interpretations* ...) needs not be conforming to a predefined format, but can be polymorphic, originating from both specific individual annotation initiatives as from collective agreed upon contributions or even algorithms. Not only is it stored, but it can also be retrieved and reprojected into any format. All these data are not necessarily human-contributed but can actually be the result of complete document analysis pipelines[4] and algorithms. As a result, the resource can hold apparently contradictory *interpretations* of identical documents, when these *interpretations* stem from different transformation and analysis processes. This means that there cannot be an absolute and unique notion of *ground-truth*, since the resource is hosting the interpretations for multiple contexts. This corresponds to recent debates on the existence (or rather lack thereof) of ground-truth or universal interpretations.[5–8]

The last item in our list describes a service where, certification set apart, not only data is provided as a resource, but also interactions with this data are available, as through documented benchmarks, state-of-the-art reference algorithms, *etc.*

The next sections will describe in detail what specific roles and actors can be defined to realize the scenario we have depicted above.

### 2.2 Featured Roles

Before considering the more scientific and technical underlying key points of our scenario, here are the essential roles it has to fulfill. Each of these points will then be further developed in section 2.3.

**Formats and Representation** are to be examined before addressing the issue of storage and access to the data. In order for our scenario to have a chance of realization, it seems to us that it cannot, ever, take any assumption on what data representation or formats should be used. Past experiments have shown too often that an approach consisting in inciting or "coercing" a community into using a single set of representation conventions does not work or contributes to locking it into a limited sub-part of uses of the generated data, hampering creative and extended uses beyond their initial purpose. This is clearly contradictory to our standpoint with respect to *ground-truth* and our preferring of the term *interpretations*.

This clearly advocates for an as open as possible way of representing any kind of data. On the other hand, *à trop étreindre, mal embrasse*, and abandoning any kind of imposed structure may prove rendering the querying part of our plot hard to realize. This will not be an issue if there is a widespread use and sharing of data, since it will eventually lead to the emergence of *de facto* formats, especially since we are targeting a rather well defined community. The less constraining representations formats are, the higher the chance there will indeed be a widespread use.

**Storage and Access** is one of the key elements to making our resource and associated services available and accessible. It seems, however, this is "merely" a question of material resources, rather a conceptual problem (this is an understatement, but availability, redundancy and consistency discussions are beyond the scope of this paper); especially when remaining under the assumption that the storage remains centralized (condition, as already mentioned, that can be alleviated). The quantity of data is potentially huge, and needs to be constantly available. The supporting infrastructure needs to be sufficiently dimensioned for handling multiple concurrent accesses and provide high enough bandwidth. The platform also needs to be able to host the versatile data-model implementing the previous point, allow for the rich semantic querying described in the next, and remain capable of hosting and executing the algorithms described after as well as providing the interfaces for commenting, evaluating, rating or re-projecting anything already stored.

**Querying and Retrieval** needs to allow an as broad as possible spectrum of interactions. Since we cannot really rely on a canonical representation of all stored data – mainly because there probably isn't any universal representation of all possible annotations and interpretations of document analysis results and document contents – querying and retrieval cannot just be keyword based. We need a much more semantic way of interacting with our data ... raising "data" to the level of "information", essentially. This task is challenging. The illustrative query mentioned before: *"1,000 random documents in 300dpi bitonal* `.tif` *format predominately containing printed material and of which a reasonable number of pages contain at least one handwritten annotation."* raises quite some exciting challenges. While "bitonal `.tif`" might be the less controversial part of the query, it still may be subject to discussion: the `.tif` format may support compression, while given algorithms require their `.tif` input to be uncompressed; what exactly does bitonal mean ? and if we allow on-the-fly conversions of image formats into `.tif`, would we allow also conversions of graylevel or color into bitonal ? (but then what binarization algorithm should we use ?) *etc.* The further downstream use of the retrieved data may require the handwritten annotations markup and localization data need to be *compatible* with the algorithm is is going to be tested on. This implies that this *compatibility* be somewhere formally expressed, and that, at the same time, the input type of the algorithm be equally formalized.

This means that querying and retrieval most definitely need to operate on a semantic level. The corollary of that is that these semantics need to be represented somewhere. This is especially the case here, where we want to have automated interaction between data and algorithms. In order to be applied algorithms to any available document in our repository, the platform needs to have the capability to relate these algorithms to data, and therefore requires input formats and output semantics to be formalized and stored in some way, thus joining the point made previously.

**Provenance and Certification** are two supplemental benefits we get from correctly realizing the previous roles, and especially from provenance mentioned in section 2.1. They strongly relate to (and depend on) recording of the origin of all information that is stored, and can offer an interesting level of interaction and service. One of the potential uses might be certification of experiments and algorithm results. These are, in fact, a specific instances of what global provenance tracking can offer. By assuring that all information is uniquely referenced, totally persistent (cannot be removed) and traceable, one can solve many data-algorithm-result interdependence problems. For instance, widely used benchmarks suddenly prove biased or flawed: provenance makes it possible to mark or invalidate depending data. Copyright claims may result in having datasets to be retracted or recomposed, but how about generated results ? Provenance information provides tools to keep the level of availability as close to optimal as possible, by analyzing which data is impacted by the claim.

### 2.3 The Cast

Describing what a perfect and hypothetical environment should provide in order to assure a new and stimulating way of doing document analysis research is rather easy. Actually starting to implement it and making it available to the community is more of a challenge. The DAE server implementing a big part of what is described in this paper exists, and is continuously expanding its features. This section describes the more technical and scientific challenges that have been met, or need to be met, in order to completely transform above descriptions into reality.

#### 2.3.1 The DAE Platform

Our platform is available at `http://dae.cse.lehigh.edu`. It runs on a 48TB storage, 12 core 32G RAM server,

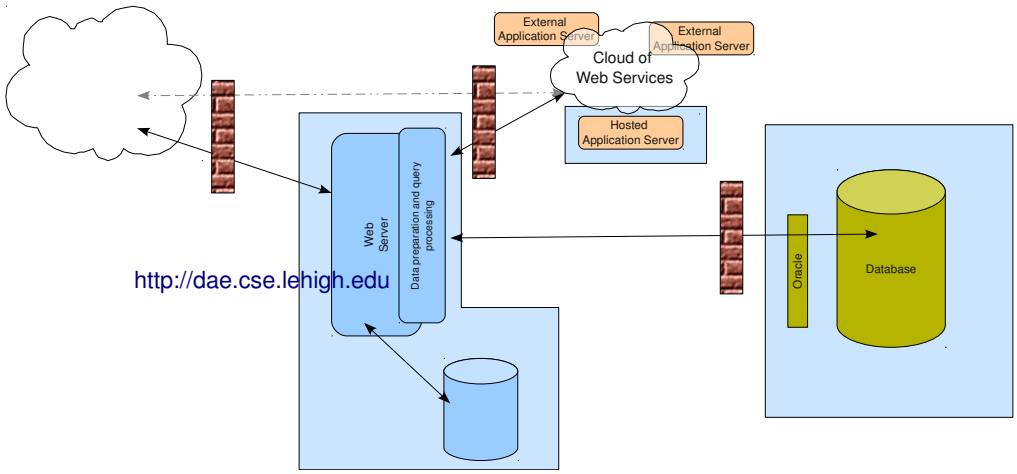


Figure 1. The DAE platform architecture

and as such, is to be more considered as a seriously endowed proof of concept*, rather than as the ultimate infrastructure to support all claims and goals expressed in this paper.

Besides the technical architecture of the server, as depicted in Fig. 1, the DAE Platform is first of all the implementation of a *data model*.[9] The data model is based on the following claims:

- all data is typed; users can define new types;
- data can be attached to specific parts of a document image (but does not need to),
- both data and algorithms are modeled; algorithms transform data from one type into data of another type;
- full provenance of data history is recorded;

It has been fully implemented on an Oracle 11.2 back-end database management system (right – in green in Fig. 1). It is accessed by a web front-end that provides a Web 2.0-like interface (left – in blue Fig. 1) and encapsulates SQL queries to the back-end. It also relies on independent "application" servers that are used for executing registered algorithms on the data (middle – in orange Fig. 1). The data model can be downloaded from `http://dae.cse.lehigh.edu/Design/ER.pdf`. All source code is GPL licensed and freely available from `http://sourceforge.net/projects/daeplatform/` and are open to contributions and extensions by the community.

This platform fulfills most of the roles described in section 2.2:

*Ideally, the platform should evolve into a distributed community-managed resource, rather than remaining a centralized platform.

**Formats and Representations** are transparently handled by the system, since the user can define any format, naming or association convention within our system. Data can be associated with image regions, image regions can be of any shape and format, there is no restriction on uniqueness or redundancy, so multiple interpretations are not an issue. Furthermore, in order to avoid generating a bulk mass of incomprehensible annotations and data-structures, data can be conveniently grouped together in sets. These sets can in their turn be named and annotated as well. Data items do not need to belong exclusively to a single set, so new sets can be created by recomposition or combination of existing data sets.

The core elements of our solution consist of considering most of the stored information as `data_items` which in their turns instantiate as more specific subtypes. The principal ones are depicted in the partial data model view of Fig 2.

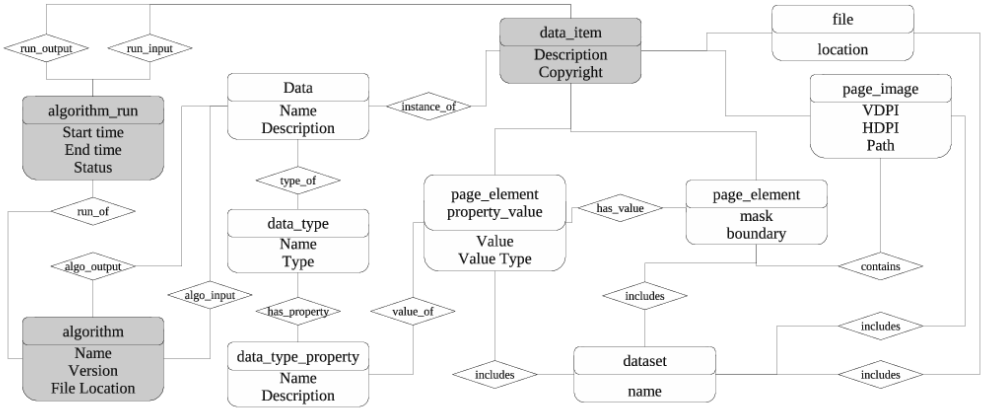


Figure 2. A partial view of the DAE platform data model

In our approach, `page_images` are considered as single, stand-alone, document analysis objects in a given context. They may be related to other images, known to represent the same physical object, but captured under other conditions (not represented in Fig. 2).

`page_images` can contain any number of `page_elements`. These are areas of pixels contained within the image. Our model is currently capable of representing these areas as bounding boxes or pixel masks, but can be extended in a very straightforward way to handle polygons or other geometric shapes. It is noteworthy to mention that there is no real need to extend beyond those types, however. One might be tempted (especially when considering graphical documents) to introduce complex vectorial representations, for instance, like ellipses, lines *etc.*. We argue that those belong to the domain of *interpretations* (and therefore `page_element_property_values`) rather than `page_elements`. `page_elements` are "physical" sub-parts of `page_images` and therefore just pixels.

`page_element_property_values`, on the other hand, can be any interpretation of a `page_element`: layout labels, OCR results, ...

Using this approach we have been able to incorporate a great amount of widespread and less widespread document analysis data formats into our platform. We are hosting and compatible with the native formats of the UNLV dataset,[1] Tobacco800[2] using the GEDI[10] format, and many others. We provide XSLT conversion tools to as well as a full XML Schema of our upload format.

The results are fully browsable and interactive datasets, as shown in Fig. 3.

**Storage and Access** are covered by the system by the fact that the storage (48TB raw storage, equivalent to an effective 28TB once formated and secured) and computing capacities are largely sufficient for making a actually operational proof of concept that scales to a few hundreds of concurrent connections, and capable of storing most currently available datasets. Furthermore its architecture makes it extremely

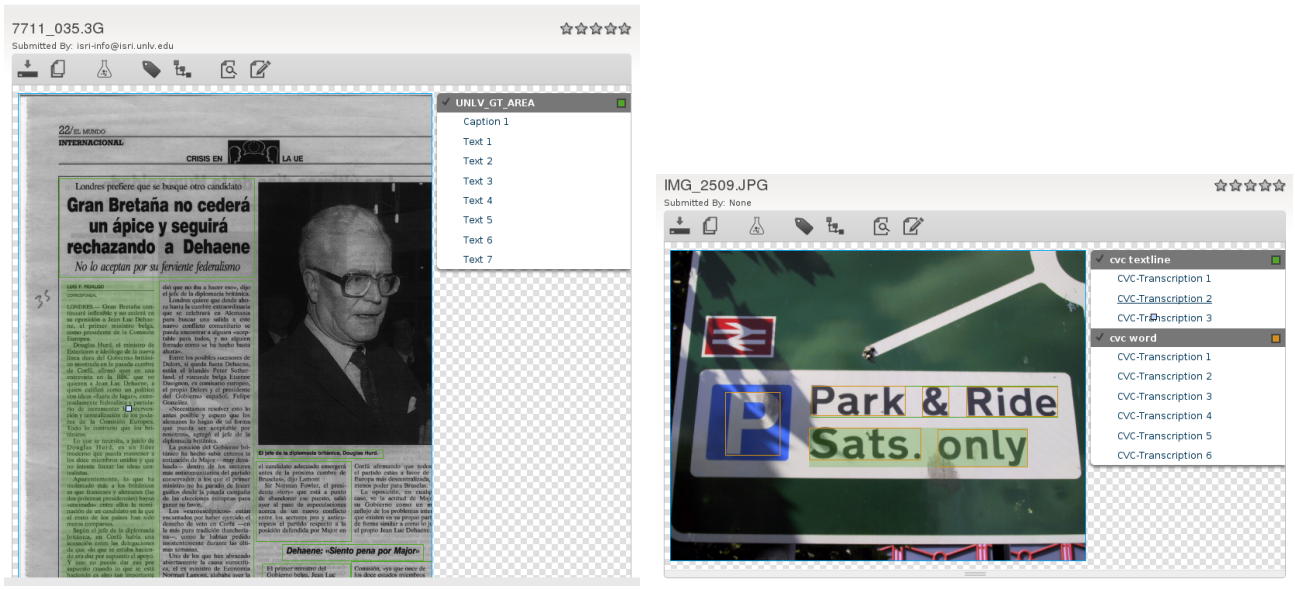


Figure 3. Example of UNLV (left) and custom (courtesy of CVC) imported `page_element_property_values` being consulted through our platform.

easy to scale to higher demands, as both the storage and the computing infrastructures are conceived as physically separate entities. The current version already spawns some of its computing onto a separate high performance computing cluster for specific algorithms and offers access to the others as Web-services. This guarantees that the part of the platform that manages the execution of algorithm is totally scalable.

Likewise, the storage of the data may also be distributed over multiple sites. Seamlessly integrating the meta data into a fully distributed environment is likely to require further investigation and research efforts, however.

**Querying and Retrieval** is the great gain that the DAE platform offers. Because of its underlying data model and architecture, everything is "SQL-queryable". The standard datasets that can be downloaded from the platform are no longer monolithic `.zip` files, but actually potentially complex queries that generate these datasets on the fly. Because of the extreme degree of flexibility in annotating and supplementing existing data with meta-data, the potential uses are endless.

The first effect is that every single data item stored on our platform has a unique identifier, be it a `page_image`, a `page_element` or `page_element_property_value` (*i.e.* an interpretation), they all are referenced through a unified URL like this:

`http://dae.cse.lehigh.edu/DAE/?q=browse/dataitem/download/99262`.

This means our platform can be considered as an on-line storage engine and that algorithms can directly access the stored data for their use, rather than requiring that it is downloaded and locally stored before use.

The second effect is that one can retrieve images on the one hand or other any information (like segmentation data, OCR results, recognition labels ...) on the other hand, or any combination of these. While this clearly allows for a classical approach to retrieving the data (*i.e.* and filter out parts of a previously pre-packaged dataset, it also allows for more creative uses and crossing information coming from various and different datasets, possibly even conceived for other application contexts than the one under consideration.

**Examples** of what our system offers are shown below.

- To find out what data sets that are declared on the platform (data sets may be recursively contain other datasets, so we restrict ourselves to top-level ones):

```
select distinct NAME,ID from DATASET where ID not in
   (select ASSOCIATING_DATASET_ID from ASSOCIATE_DATASET);
```

- To retrieve the filenames of all document images in a dataset (here the dataset defined by its id 633):

```
select PATH from INCLUDES_PAGE_IMAGE, PAGE_IMAGE where
   DATASET_ID = 633 and PAGE_IMAGE_ID = ID;
```

  It is noteworthy to mention that the `PATH` attribute retrieved is actually the URL where the image can be downloaded from. This, in its turn means that the actual image can be stored virtually anywhere on the Internet thus providing the opportunity to create *virtual* datasets, consisting of collections of remote and distributed storage elements[†].

- To retrieve all data produced by a particular algorithm (here algorithm 66):

```
select DATA_ITEM_ID from ALGORITHM, ALGORITHM_RUN_OUTPUT, ALGORITHM_RUN_OF where
   ALGORITHM_ID = 66 and ALGORITHM_ID = ID and
   ALGORITHM_RUN_OUTPUT.ALGORITHM_RUN_ID = ALGORITHM_RUN_OF.ALGORITHM_RUN_ID;
```

- To discover all user defined data types and their descriptions:

```
select * from DATATYPE_PROPERTY;
```

Although the previously mentioned query *"1,000 random documents in 300dpi bitonal `.tif` format predominately containing printed material and of which a reasonable number of pages contain at least one handwritten annotation."* requires some additional work in order to implement the randomization, and needs to handle some more subtle handling of the underlying semantics (*cf.* section 2.3.2) an SQL query implementing an approximation of it is perfectly within the scope of our platform. There is a catch, however, as will immediately become clear from the following query. Without loss of generality, let us consider only one part of the previous query: *pages containing handwriting*.

Since our data model is quite straightforward, this means that we are looking for `page_images` that have some `page_element` that has somehow a `page_element_property_value` that is *handwriting*. The model perfectly handles this, and the corresponding `page_element_property_value` can be human provided or come from specialized detection algorithms (*e.g.* page segmentation[?]). The aforementioned catch lies in the fact that there is no unique representation – and therefore no unique query – that yields the answer to *handwriting*.

**Note:** from this point on, we assume the reader is sufficiently convinced of the validity of the data model, and that there is no need to actually give all SQL queries *in extenso*. We shall therefore only provide shorter, and more readable pseudo-queries, that focus on the core of the argument.

Let's consider the two following pseudo-queries:

1. The following query returns pages having a property value labeled *handwriting*.

```
select PAGE_IMAGE such that PAGE_ELEMENT_PROPERTY_VALUE.VALUE = 'handwriting';
```

2. Similarly, the following query returns pages of which the property is of a type labeled *handwriting*.

```
select PAGE_IMAGE such that the PAGE_ELEMENT_PROPERTY_VALUE's
   DATA_PROPERTY_TYPE = 'handwriting';
```

[†]Although this opens a broad new scope of problems related to archiving and guaranteeing perennity of the provided link, we are conveniently considering, for argument sake, that the platform can guarantee that the provided `PATH` is always valid, for instance by only serving data stored it its local storage.

Indeed, as shown by queries like 1, some annotations or interpretations can be classification labels (*e.g. handwriting, table, logo ...*). Others, like 2 can actually be related to, for instance, handwriting recognition, and may yield the transcription of a `page_element` that is of type *handwriting*. As a matter of fact, queries like 1 may actually return unwanted results since they don't really take into account the type of the returned value. They might be OCR transcriptions of a text containing the printed word *handwriting*.

These differences may appear unnecessarily subtle for the average reader but they address a fundamental point related to the sharing of image interpretations. We shall discuss them in further detail in section 2.3.2.

**Interaction** with the data can be considered on multiple levels.

1. The first level is integrated in the data model, which represents algorithms, their inputs and outputs. Our platform goes further by actually providing access to binaries that can be executed on the stored data, thus producing new meta data and *interpretations* and expanding the range of available information. Queries like finding all OCR results produced by a specified algorithm, can either be used as an *interpretation* of a document, but can equally serve as a benchmarking element for comparison with competing OCR algorithms. Because of the fact that all are hosted in the same context, it becomes possible to "certify" (or *cite*, since they become unique references) evaluation or benchmarking runs.
2. The second level is more human-centered and concerns the interface that is provided to access the data. Besides the raw SQL access described previously, the platform offers a complete interactive environment that allows users to browse through stored document collections, upload data, run algorithms, tag, comment and rate data, *etc.* These functions are illustrated in Fig. 4.

### 2.3.2 Interpretations and Semantics

**Handling multiple interpretations** As mentioned before (*cf.* p. 2) the design of our platform makes it possible to have multiple interpretations of the same set of data. This is a necessity due to multiple contexts in which the same data can be used, and the intrinsic difficulties to define absolute interpretations anyway.[5–8] The major side-effect is that semantically equivalent interpretations may be stored in various ways in our platform.

Let's return to the example developed in 2.3.1, where we considered the following two following pseudo-queries:

1. `select PAGE_IMAGE` *such that* `PAGE_ELEMENT_PROPERTY_VALUE.VALUE = 'handwriting';`
2. `select PAGE_IMAGE` *such that the* `PAGE_ELEMENT_PROPERTY_VALUE's DATA_PROPERTY_TYPE = 'handwriting';`

Both are defining areas in `page_images` that contain the string *handwriting*. As already mentioned earlier, this does not really guarantee that the semantics of these areas actually cover handwritten visual information, since they might as well be the transcription of a printed text spelling the word *handwriting*.

Furthermore, the contributors of interpretations and labeling might have provided other labels, such as *manuscript* or *hand written* or *annotation*, *etc.* This is the cost we pay for a deliberate design choice. One might argue that using a thesaurus of authorized labels for annotation would solve the problem, but this is not necessarily true. This is related to the fundamental non-existence of absolute interpretations[11] and their relation with application contexts. One can easily conceive application contexts where *handwriting* is referring to totally handwritten documents, while other applications only relate to handwritten annotations on printed documents. The latter category can even be split into contexts where the annotations are guaranteed to be actual written text, opposing others where they can also be markup glyphs or other non-textual signs and marks.

The current state of development of our platform does only allow retrieval or interaction with the data on the grounds of syntax-oriented queries. However, its design already integrates ways of extending its possibilities to a more semantic-oriented interaction.

1. Since all data provenance is recorded and accessible, one can easily retrieve data produced by specific algorithms using specific runtime parameters. This guarantees that the obtained data share the same interpretation context, and thus has some significant overlapping semantic value.

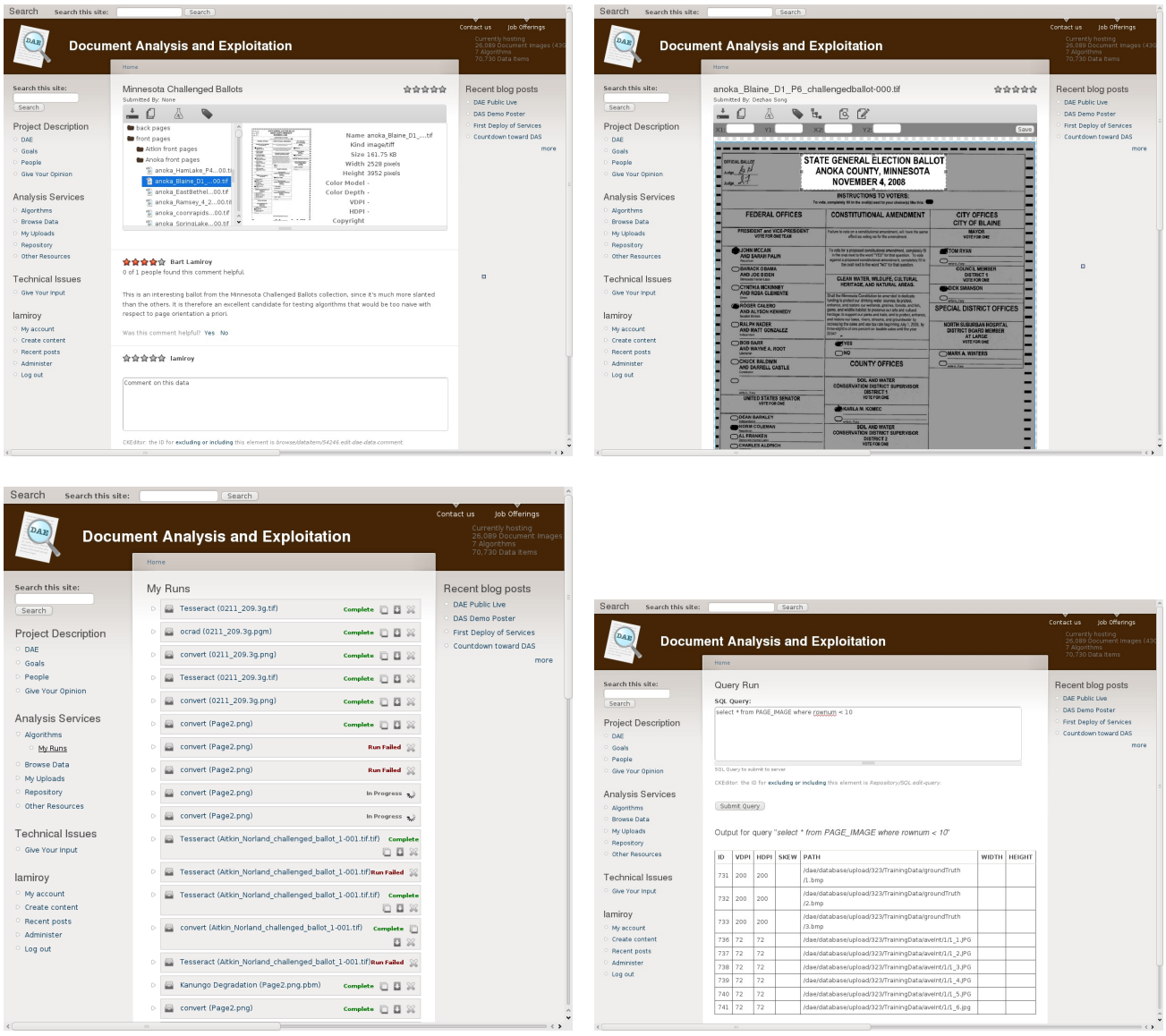


Figure 4. Data Interaction using the user interface: browsing data sets, commenting and rating (top left), image area markup creation (top right), algorithm execution (bottom left) and SQL querying (bottom right).

2. Since users are free to use and define their own, personal annotation tags, and since, again, provenance is recorded, one can assume that identical users, using identical tags, refer to identical semantic values.

3. Ongoing work currently consists in exporting our data model in OWL[12] and allowing contributors to explicitly define the semantics of their algorithms and data-relations.

**Semantic Web and Ontologies** Semantic Web[13] and ontologies are immediate future actors in our scenario. In order to make the querying really semantic in an as automated way as possible, and by correctly capturing the power of expressiveness as people contribute to the resource pool, the DAE platform will need to integrate adequate knowledge representations. This goes beyond the current storage of attributes and links between data. Since individual research contexts and problems usually require specific representations and concepts, contributions to the system will initially focus on their own formats. However, as the need for new *interpretations* arises, users will want to need to combine different representations of similar concepts to expand their experiment base. In order to allow them doing that, formal representations and semantic web tools are being developed in this framework.

When data needs to be shared, there are two common approaches: standardization the schema or design transformations. The first approach is very rigid, is slow to adapt to new needs, and may result in useful data being lost because there is no way to represent it. It has been tried in the past in the document analysis and its connected domains, without emergence of a global standard. The second approach can adapt to changes somewhat more easily, but because the transformations are often procedural in nature, it is difficult to reuse and/or compose them. Using semantic web technologies, ontologies can be used to provide not only schema information but additional semantic information that can aid in transformations. Also additional mapping ontologies can be used to provide declarative alignments between different domain ontologies (for instance different interpretations of the same data in other application contexts). Logical reasoners can then be used to check the correctness of the mappings and to compose them to produce indirect mappings between ontologies for which no direct mapping exists. A recent project has demonstrated how the technique works by integrating e-commerce taxonomies.[14]

### 2.3.3 SOA Architectures, Web-Services and the Cloud

The fact that the platform integrates the use of a fully SOA architectured set of algorithms extends the data model in a similar way as the previous points. By opening the supporting architecture using carefully designed tools of remote, distributed storage and execution (aka *the Cloud*) the DAE platform may greatly reduce all risks related to scalability, availability, centralization and cost, and eventually become a community governed and distributed resource, not only hosted at Lehigh University, but shared by all its users, both in its use as in its infrastructure support.

Furthermore, the use of semantically labeled services can greatly increase the level of interaction and querying of all stored data, as hinted by in sections 2.3.2 and 2.3.2.

## 3. RELATED PROJECTS

The document-analysis community is not unique in its need for a managed resource repository. Various scientific communities maintain shared datasets that face issues similar to ours regarding the distributed sources of data, the derivation of data from other data, and the identification of the specific version of a specific dataset used in a specific publication.

The three most recent ACM SIGMOD Conferences have sought to gather the data and implemented algorithms behind accepted papers. Authors of accepted papers were invited (not required) to provide code, data, and experiment setup information to permit testing for repeatability of the reported experiments and the ability to run additional experiments with different parameter settings.

The problem of integrating multiple interpretations of a specific document is related to the problem of integrating scientific observations whether done by multiple scientists or gathered from multiple remote sensors. The Sloan Digital Sky Survey[15] was pioneering work that integrated world-wide astronomical observations into

a single database. This work allows an astronomer to access data quickly without travel and to compare related observations made at a variety of times and from a variety of places. Microsoft's SciScope[16] is a tool that allows search of a wide variety of environmental databases and integrates them into a consistent format (`www.sciscope.org`).

Standing in contrast to the two previously mentioned projects, which are specific to particular domains, Google Fusion[17] provides a framework for sharing data and visualizations of data, and for discussions pertaining to data. While Google Fusion is not domain-specific, it lacks our dual focus on both data and algorithms. However, the cloud-hosted, open framework of which Google Fusion is an example, might be a desirable long-term evolution path for our repository.

The digital-library community has studied citations to evolving documents and subparts thereof (see, e.g.,[18]). However, there is normally an underlying assumption in such systems that the hierarchical structure of a document (chapter, section, etc,) is well known. In document-analysis, the document structure itself may be a matter of interpretation.

Provenance, though having been studied in a variety of contexts over several years, is now emerging as a research area in its own right. While some work in provenance focuses on data and their derivation, other work focuses on workflows, or the process of data manipulation. The Panda project at Stanford,[19] is attempting to unify these two approaches and comes closest among prior work to capturing our goal of managing not only data but also analysis algorithms.

## 4. DISCUSSION

Apart from opening a whole new range of research practices that can be imagined from the availability of this platform, and by offering a set of fundamental tools to expand document analysis to interact with new communities like the Semantic Web and knowledge representation or databases and provenance,[20] to name only those, this work also addresses a number of basic concerns that have been around for a long time related to sharing common datasets (G. Nagy[21] acknowledges datasets dating back to 1961 in the OCR domain), "ground-truths" or *interpretations* and conducting objective benchmarking.

Upto now, datasets have always been homogeneous, either because having been produced in a common context,[22–24] targeted to solving a specific, well identified problem (like OCR training,[1,25] layout analysis[26] *etc.*) or both. Because of this, some of them have grown obsolete as technology has changed and evolved, and the notion of *document* itself has morphed over the years. Others have not been able to be kept available, since resources to do so have been consumed, or either the personal commitment and involvement of individuals or supporting groups has faded. This has always hampered reuse and combination of existing data to have it evolve.

Furthermore the costs of generating "ground truth" (or rather *reference interpretations*) remain extremely high, since, by construction, they need extensive human intervention. This makes them generally extremely and specifically problem focused and rarely generic. This makes very good sense, since they often serve as a basis for well defined contexts. However, their very specialized nature makes it difficult to re-use them in other contexts, and, because of their cost and their very specialized nature, they can produce a *lock-in syndrome*, simply because there are no resources to re-engineer new, more appropriate ground truth data, as the domain and knowledge evolves.

These issues are largely solved by our approach in the sense that data sets are merely partial projections of the global data pool that can evolve naturally as new contributions, both document images as annotations and meta-data or *interpretations*, are added. Furthermore, since they can be used and reused in any context that is appropriate, they can more easily be used beyond their initial intended scope. Finally, since the platform is completely open, it can be reproduced, distributed and cloned so that global persistence and availability no longer are a problem.

Our current focus is also to provide interaction interfaces with a very smooth learning curve such that widespread use is not hindered by constrained and complex representation formats.

## 5. HOW THIS CAN IMPACT RESEARCH

The platform presented in this paper is a first step in a more global direction towards a more integrated and comprehensive vision of what research in Document Analysis, and more globally Machine Perception, benefit from. A more detailed description of it is open to contributions and is available at `http://dae.cse.lehigh.edu/WIKI`.

### 5.1 Improve Experimental Reproducibility and Evaluation

The main advantage of the DAE platform is that it can fulfill a role of independent certification authority, and thus contribute to enhancing reproducibility and evaluation of published results. For argument's sake, lets consider the case of a researcher developing a method operating on a specific class of documents (*e.g.* "300dpi bitonal `.tif` images predominately containing printed material and of which a reasonable number of pages contain at least one handwritten annotation."). Consider the following cases (in increasing order of both added value and complexity):

1. When publishing her first results the researcher can register her experimental dataset to the platform, and produce a unique and commonly accessible reference to the dataset she used to validate her work. This is not very different from current practices of publishing used datasets.
2. When her work becomes more mature and robust, she might be interested in not just using her own personal (and perhaps biased) datasets, but rely on the platform to provide her with a set of random documents corresponding to her context criteria. Hence the focus on "1000 random images ..." in the previously developed examples. Not only can the plaform recall exactly what random images it provided, but it can actually provide the same set of images, as well as the original query, to anyone who requests it. This has several advantages:
   - other researchers can have access to the query that was issued; the previously potential and implicit bias to the dataset now disappears, since the operating conditions are made explicit (*e.g.* the method expects bitonal images);
   - the query result can still be stored for future reference and re-use;
   - in order to test the robustness of the published method, other, similar datasets can be produced, sharing the same set of properties.
3. Since the platform can easily keep track of uses and downloads of data, one can imagine a feature consisting in providing certified evaluation data sets that are guaranteed having never been provided to the user who requested them, while maintaining reproducibility and open referencing. This would allow researchers to publish "certified" evaluation results on data they are very likely never to have used before.
4. Given that the platform also hosts, or provides access to state-of-the-art algorithms, as well as certified data, it also becomes easy to imagine that this framework may serve as an independent benchmarking and comparison platform, allowing any user to compete against best of breed algorithms, potentially even integrated in complex end-to-end pipelines. Of course, as for the datasets, its provenance features would allow it to issue to guarantee the objectiveness and to certify the comparison results and rankings.

Furthermore, its open sourced code base, as well as its scalable features, notably its capacity to seamlessly integrate in a distributed environment, prevent it from being trusted by a single player in the field, and allow it to evolve in a truly community managed resource.

### 5.2 Open Pathways to Cross-Domain Research

As a last, but important item it is to be pointed out that, although originally concerning document analysis resources and research, many of the issues addressed in this paper open interesting research opportunities beyond: correctly and efficiently storing all provenance data is not completely solved, making and guaranteeing time-invariance of the queries mentioned in the previous section remains a challenge, making the data repository truly distributed even more so. The issues of multiple interpretations, the underlying questions of aligning the

associated implicit or explicit ontologies are another major question; as is the one of automatically discovering or learning ontology-like structures from usage observation of data and algorithms. And what about the quality of added data and interpretations ? Since the system both allows for user-provided ranking and commenting on the one hand, and monitors usage on the other hand, to what degree would social media inspired algorithms for reputation and ranking apply to assess the quality of crowd-sourced contributions to the hosted data ...

This platform offers the opportunity to leverage a large number of collaborative research initiatives around the concepts of Document Analysis and Interpretation in a very broad, yet concrete way.

## Acknowledgements

The DAE project and resulting platform is a collaborative effort hosted by the Computer Science and Engineering Department at Lehigh University and is funded through a Congressional appropriation administered through DARPA IPTO via Raytheon BBN Technologies.

The project has involved, over the year 2010, the following members (in alphabetical order) Chang An, Sai Lu Mon Aung, Henry Baird, Austin Borden, Michael Caffrey, Siyuan Chen, Brian Davison, Jeff Heflin, Hank Korth, Michael Kot, Bart Lamiroy, Yingjie Li, Qihan Long, Daniel Lopresti, Dezhao Song, Pingping Xiu, Dawei Yin, Yang Yu and Xingjian Zhang.

The authors would like to acknowledge their diverse contributions through discussions, shared thoughts, code development, *etc.*

## REFERENCES

[1] "UNLV data set." http://www.isri.unlv.edu/ISRI/OCRtk.

[2] "Tobacco800 data set." http://www.umiacs.umd.edu/~ zhugy/Tobacco800.html.

[3] "UW english document image database I: A database of document images for OCR research." http://www.science.uva.nl/research/dlia/datasets/uwash1.html.

[4] Lamiroy, B. and Najman, L., "Scan-to-XML: Using Software Component Algebra for Intelligent Document Generation," in [*4th International Workshop on Graphics Recognition - Algorithms and Applications - GREC'2002 Lecture Notes in Computer Science*], Blostein, D. and Kwon, Y.-B., eds., *Lecture Notes in Computer Science* **2390**, 211–221, Springer-Verlag, Kinsgton, Ontario, Canada (10 2002).

[5] Hu, J., Kashi, R., Lopresti, D., Wilfong, G., and Nagy, G., "Why table ground-truthing is hard," *International Conference on Document Analysis and Recognition* , 0129, IEEE Computer Society, Los Alamitos, CA, USA (2001).

[6] Lopresti, D., Nagy, G., and Smith, E. B., "Document analysis issues in reading optical scan ballots," in [*DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*], 105–112, ACM, New York, NY, USA (2010).

[7] Smith, E. H. B., "An analysis of binarization ground truthing," in [*DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*], 27–34, ACM, New York, NY, USA (2010).

[8] Clavelli, A., Karatzas, D., and Lladós, J., "A framework for the assessment of text extraction algorithms on complex colour images," in [*DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*], 19–26, ACM, New York, NY, USA (2010).

[9] Korth, H. F., Song, D., and Heflin, J., "Metadata for structured document datasets," in [*DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*], 547–550, ACM, New York, NY, USA (2010).

[10] David Doermann, Elena Zotkina, and Huiping Li, "GEDI – A Groundtruthing Environment for Document Images," *Ninth IAPR International Workshop on Document Analysis Systems (DAS 2010)* (2010). Submitted.

[11] Eco, U., [*The limits of interpretation*], Indiana University Press, Bloomington : (1990).

[12] Bizer, C., "D2r map - a database to rdf mapping language," in [*WWW (Posters)*], (2003).

[13] Feigenbaum, L., Herman, I., Hongsermeier, T., Neumann, E., and Stephens, S., "The semantic web in action," *Scientific American* **December** (2007).

[14] Yu, Y., Hillman, D., Setio, B., and Heflin, J., "A case study in integrating multiple e-commerce standards via semantic web technology," in [*ISWC '09: Proceedings of the 8th International Semantic Web Conference*], 909–924, Springer-Verlag, Berlin, Heidelberg (2009).

[15] Szalay, A. S., "The sloan digital sky survey and beyond," *SIGMOD Record* **37**(2), 61–66 (2008).

[16] B. Beran, C. v. and Fatland, D., "Sciscope: a participatory geoscientific web application," *Concurrency and Computation: Practice and Experience* **22**, 2300–2312 (2010).

[17] Gonzalez, H., Halevy, A. Y., Jensen, C. S., Langen, A., Madhavan, J., Shapley, R., and Shen, W., "Google fusion tables: data management, integration and collaboration in the cloud," in [*ACM Symposium on Cloud Computing*], 175–180 (2010).

[18] Buneman, P. and Silvello, G., "A rule-based citation system for structured and evolving datasets," *Data Engineering Bulletin* (September 2010).

[19] Ikeda, R. and Widom, J., "Panda: A system for provenance and data," *Data Engineering Bulletin* (September 2010).

[20] Davidson, S. B. and Freire, J., "Provenance and scientific workflows: challenges and opportunities," in [*SIGMOD '08: Proceedings of the 2008 ACM SIGMOD international conference on Management of data*], 1345–1350, ACM, New York, NY, USA (2008).

[21] Nagy, G., "Document systems analysis: Testing, testing, testing," in [*DAS 2010, Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems*], Doerman, D., Govindaraju, V., Lopresti, D., and Natarajan, P., eds., 1 (2010).

[22] Agam, G., Argamon, S., Frieder, O., Grossman, D., and Lewis, D., *The Complex Document Image Processing (CDIP) test collection.* Illinois Institute of Technology (2006).

[23] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., and J.Heard, "Building a test collection for complex document information processing," in [*Proc. 29th Annual Int. ACM SIGIR Conference*], 665–666 (2006).

[24] Thoma, G. R., "Automating data entry for an on-line biomedical database: A document image analysis application," in [*International Conference on Document Analysis and Recognition*], 370–373 (1999).

[25] Rice, S. V., Nagy, G. L., and Nartker, T. A., [*Optical Character Recognition: An Illustrated Guide to the Frontier*], Kluwer Academic Publishers, Norwell, MA, USA (1999).

[26] Wang, Y., Phillips, I. T., and Haralick, R. M., "Table structure understanding and its performance evaluation," *Pattern Recognition* **37**(7), 1479 – 1497 (2004).