A Real-World Noisy Unstructured Handwritten Notebook Corpus for Document Image Analysis Research

Jin Chen, Daniel Lopresti, Bart Lamiroy
CSE Department, Lehigh University
{jic207, lopresti}@cse.lehigh.edu, Bart.Lamiroy@loria.fr

Introduction

Traditionally, document image analysis (DIA) is conducted on datasets that are *prepared* for research purposes.

Handwriting Recognition:

CEDAR, CENPARMI, ...

Authorship Analysis:

IAM, Firemaker, ...

Prepared Datasets

spontaneous or *elicited*: whether handwritten samples are affected by data collectors.

- +: Elicitation simplifies the data collection.
- - : Differs from real-world scenarios.

raw or *curated*: whether the post-processing of datasets excludes any type of samples, e.g., hard cases.

- +: Curation simplifies solutions to the problem.
- - : Might overestimate system performance in real life.

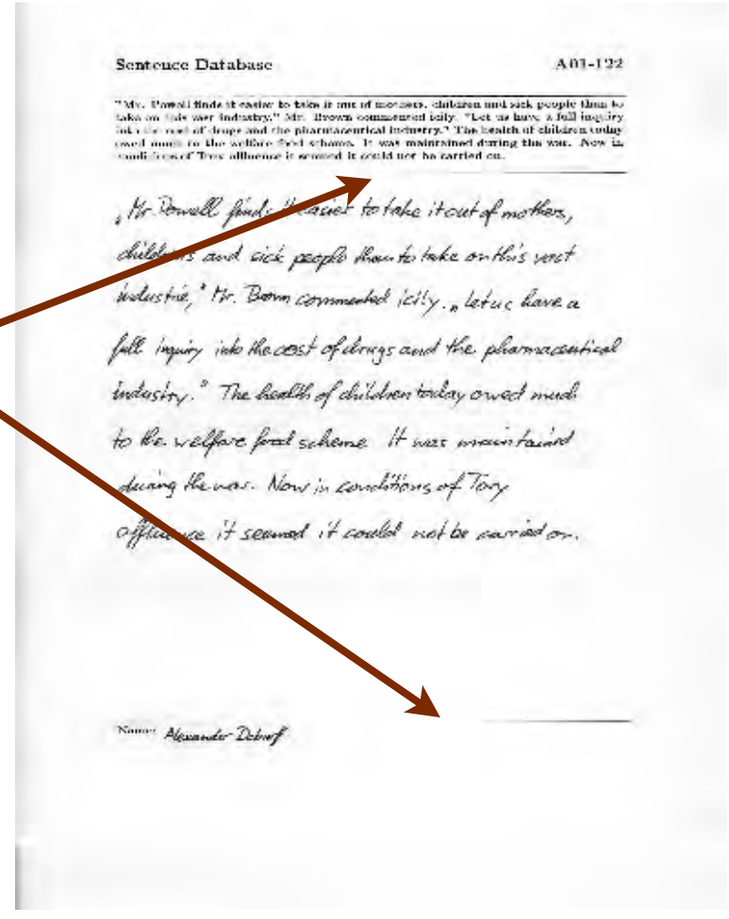
However, there are no absolute spontaneous and curation free datasets.

An Example

The IAM dataset is a large scale handwritten English dataset for handwriting recognition, writer ID, etc.

However, restrictions are applied in data collection:

- Employ pre-printed separating lines.
- Require the use of rulers and an 1.5cm spacing between lines.
- Subjects intervened if the supervisor observes limited space on page.

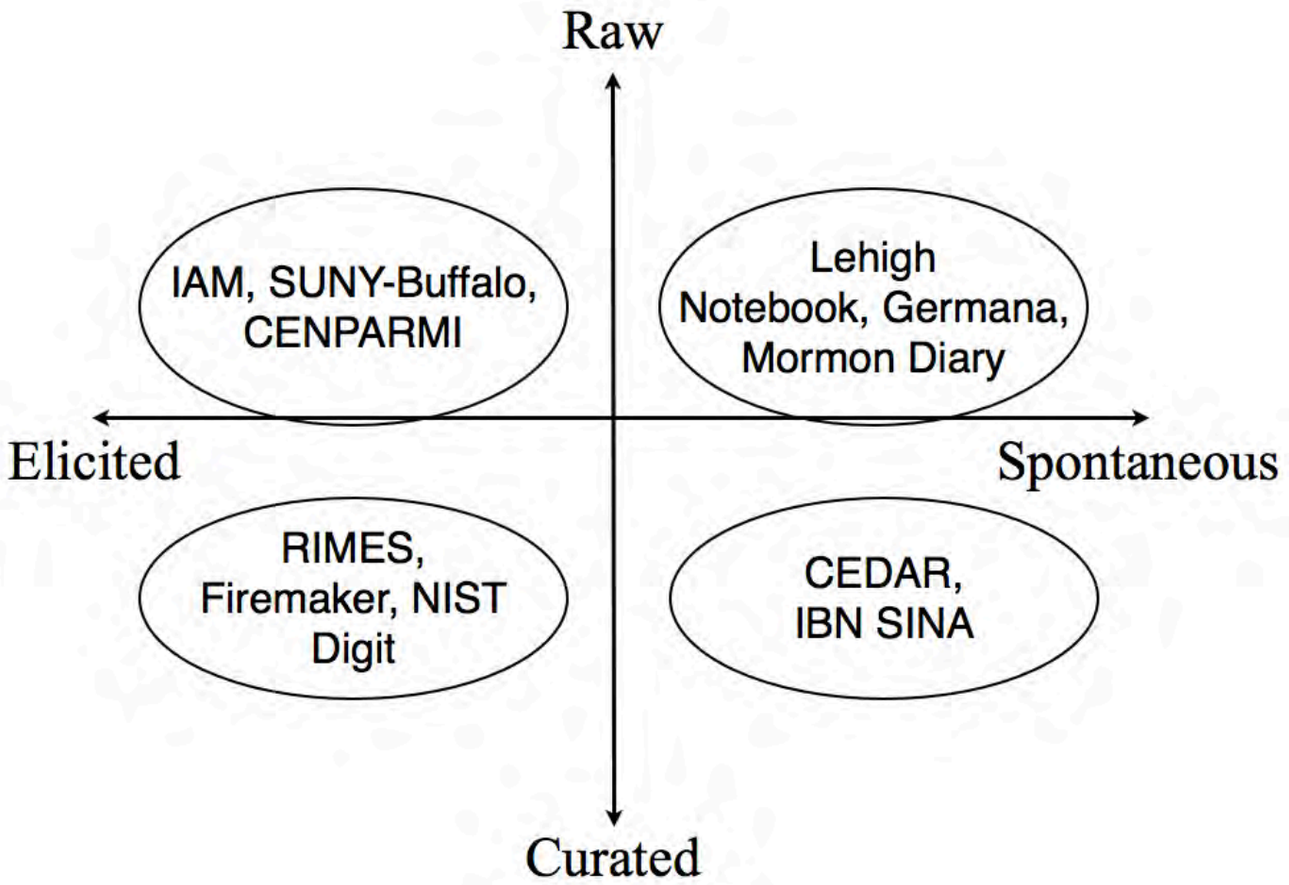


Existing Datasets

Datasets	Source	Process	Purpose
IAM	Elicited	Raw	HW Recognition, Writer ID
SUNY	Elicited	Raw	Writer ID
Firemaker	Elicited	Curated	Writer ID/Verification
NIST(SD3)	Elicited	Curated	Character Recognition
RIMES	Elicited	Curated	HW Recognition
IBN SINA	Spontaneous	Curated	Historical HW Recognition
CENPARMI	Elicited	Raw	U.S. Zip Code Recognition
CEDAR	Spontaneous	Curated	U.S. Zip Code Recognition
Mormon Diary	Spontaneous	Raw	Historical Document Analysis
Germana	Spontaneous	Raw	Historical Document Analysis
LU Notebook	Spontaneous	Raw	Various Document Analysis

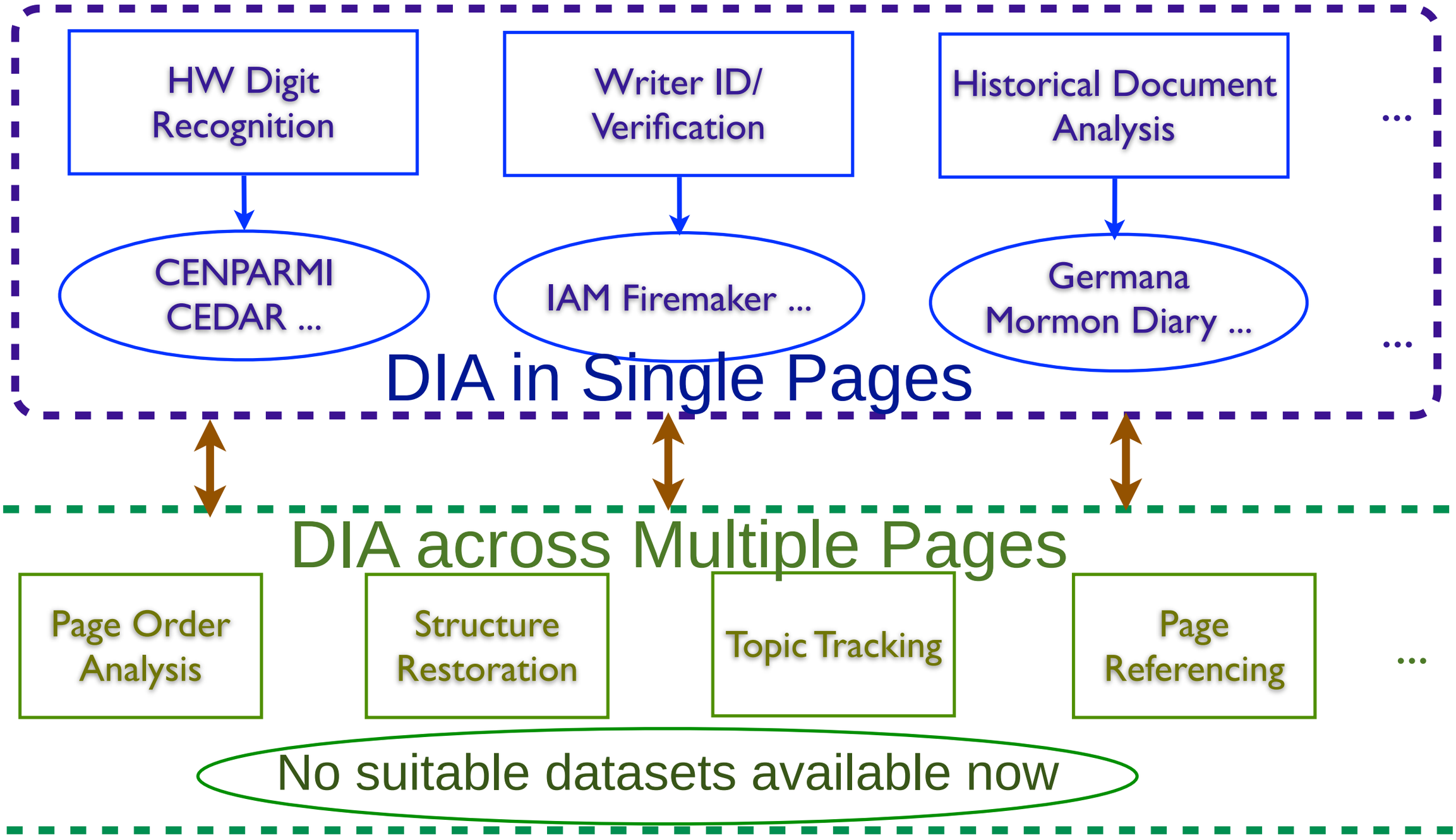
Motivation

- Most datasets are either elicited or curated.
- Germana and Mormon Diary datasets are historical handwriting datasets that are divergent from modern handwritten datasets.



We want to reduce as much as possible the elicitation and the curation during the process of building datasets.

Problem Space vs. Dataset Space

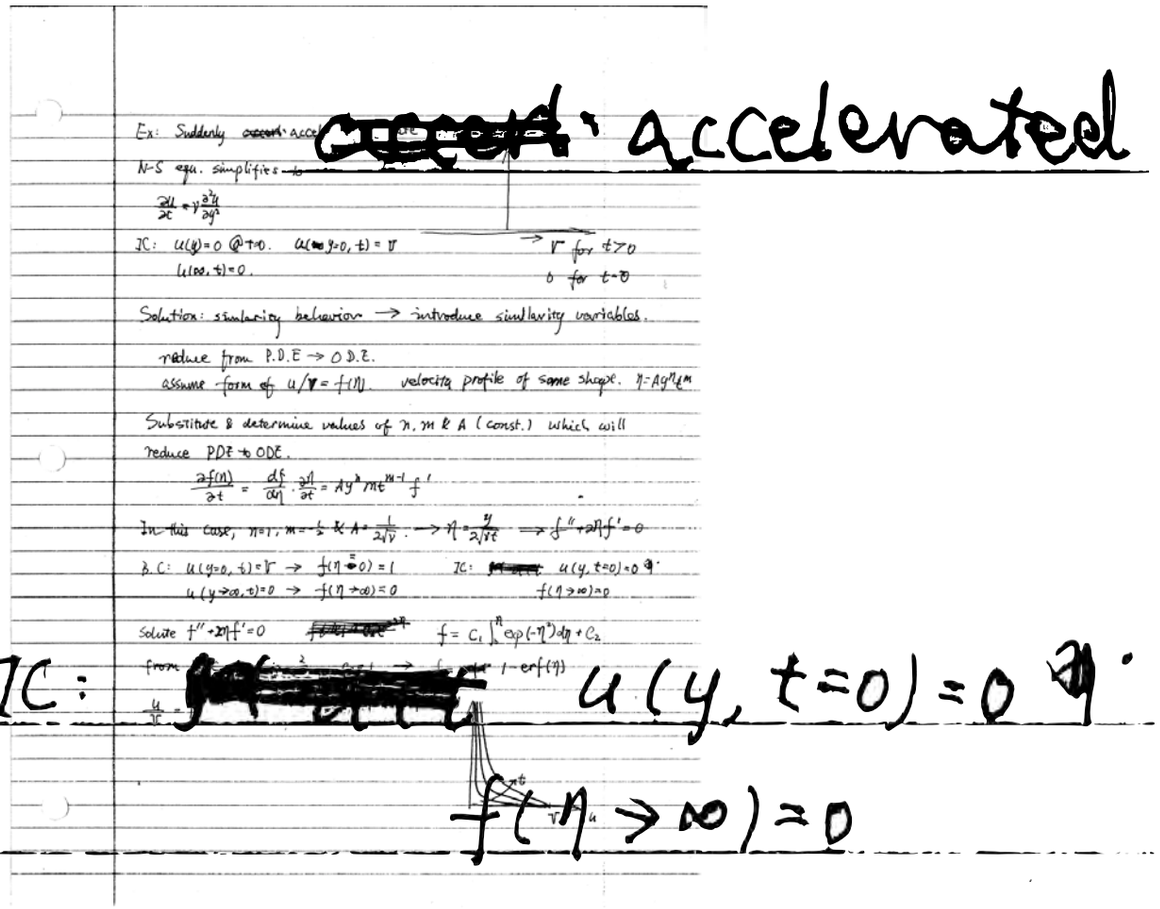


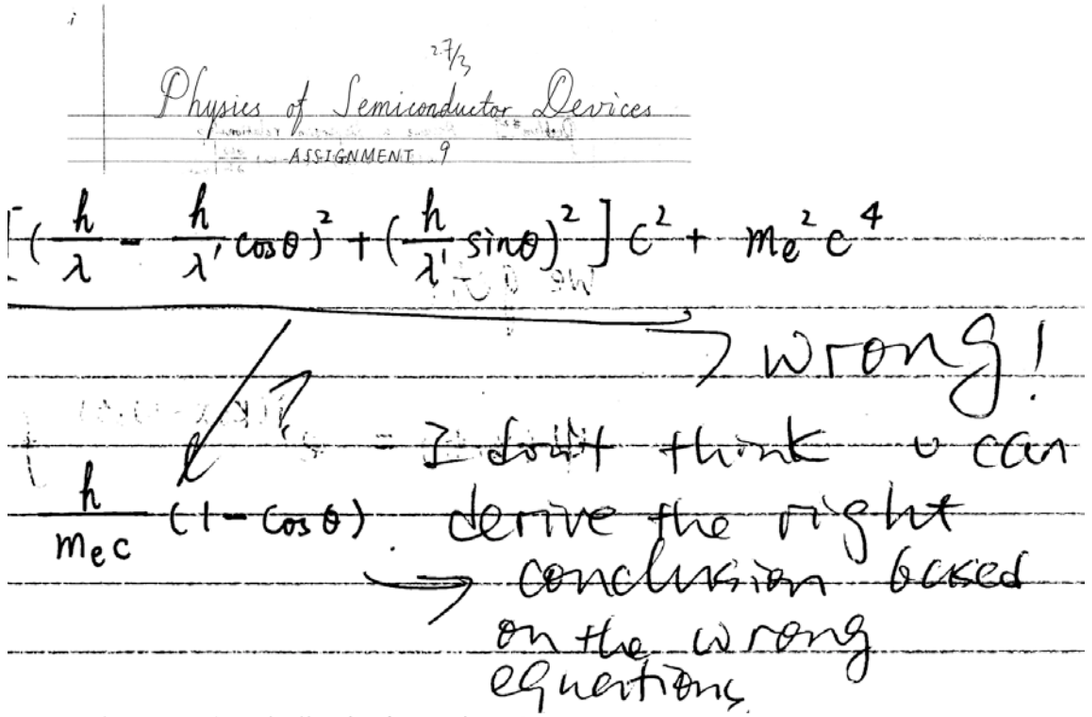
Lehigh Notebook Dataset

- All the notebooks were used by Lehigh students, thus ensuring minimum elicited handwriting.
- To scan notebooks, we separated pages while ensuring the page order.
- Each notebook page was scanned at 600dpi into PDF files, using a bitonal setting under plain text mode.
- All pages were converted into TIFF images with no compression, resulting in *5104w x 6600h*.
- So far, we have collected 499 pages from nine students and aim for 100 notebooks, 3k pages from >50 students.

Lehigh Notebook Dataset

Differences from existing handwriting datasets: corrections, annotations, arrows, doodles, etc.

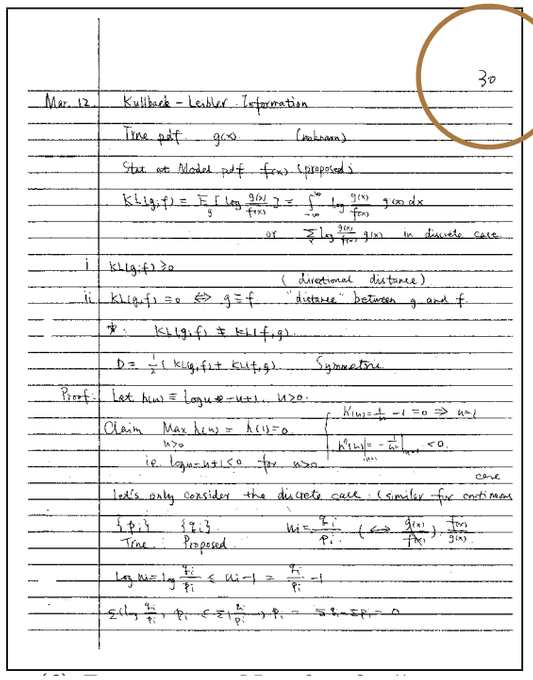


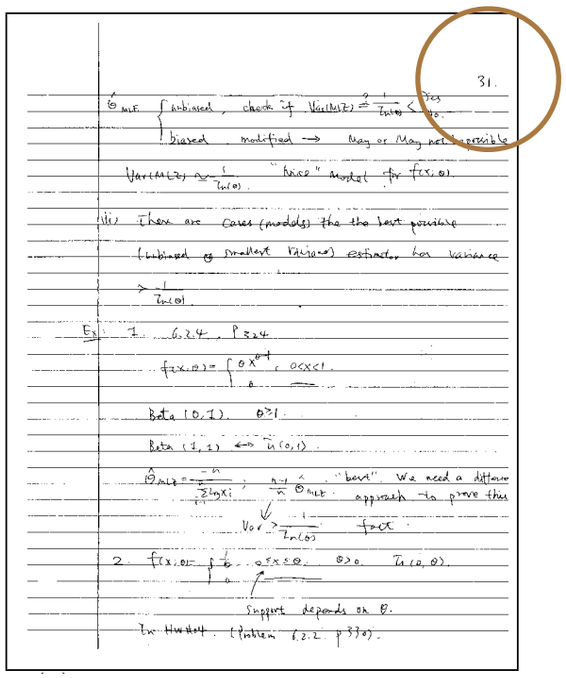


Minimum elicited and curated handwriting!

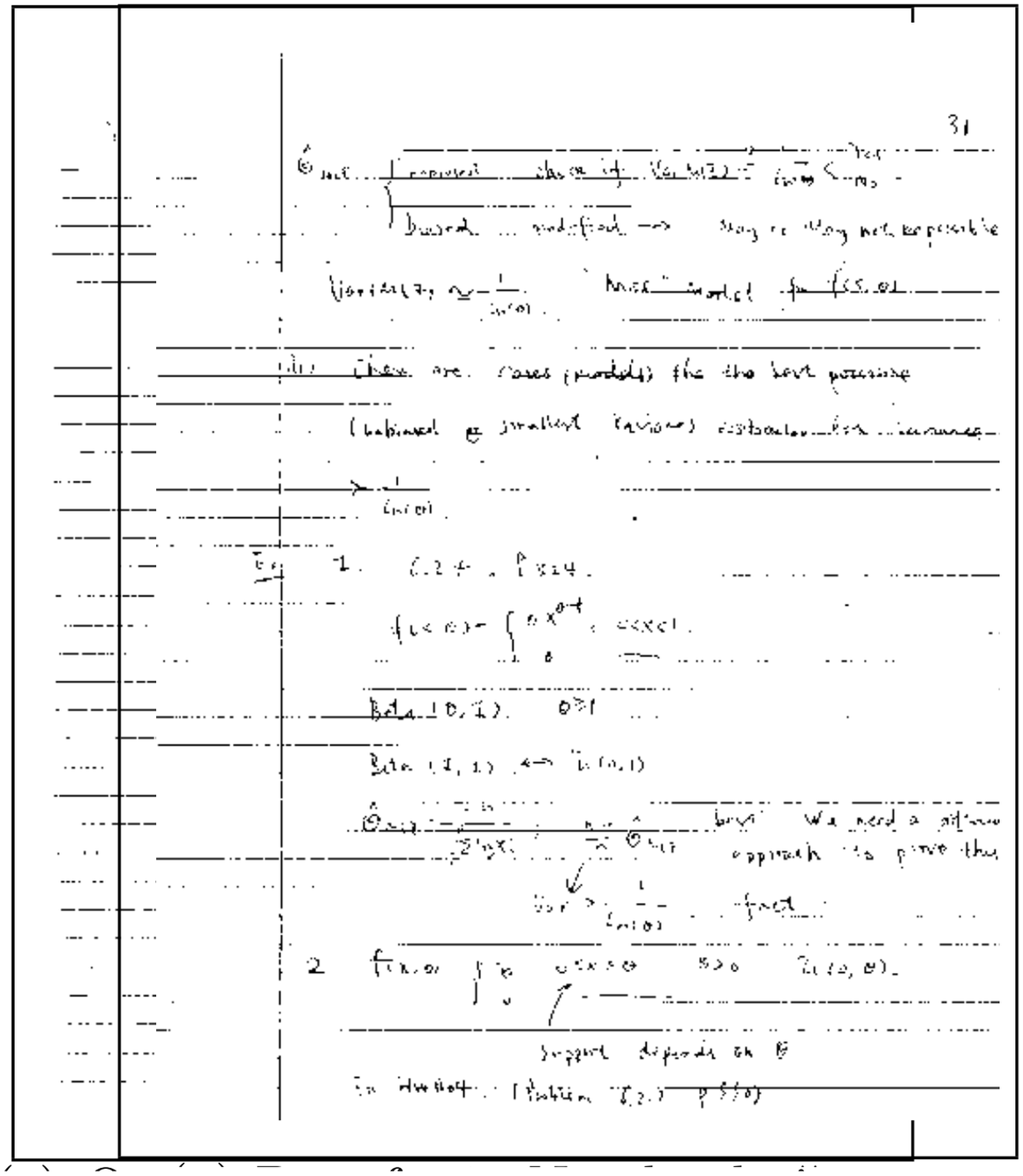
Page Order Analysis

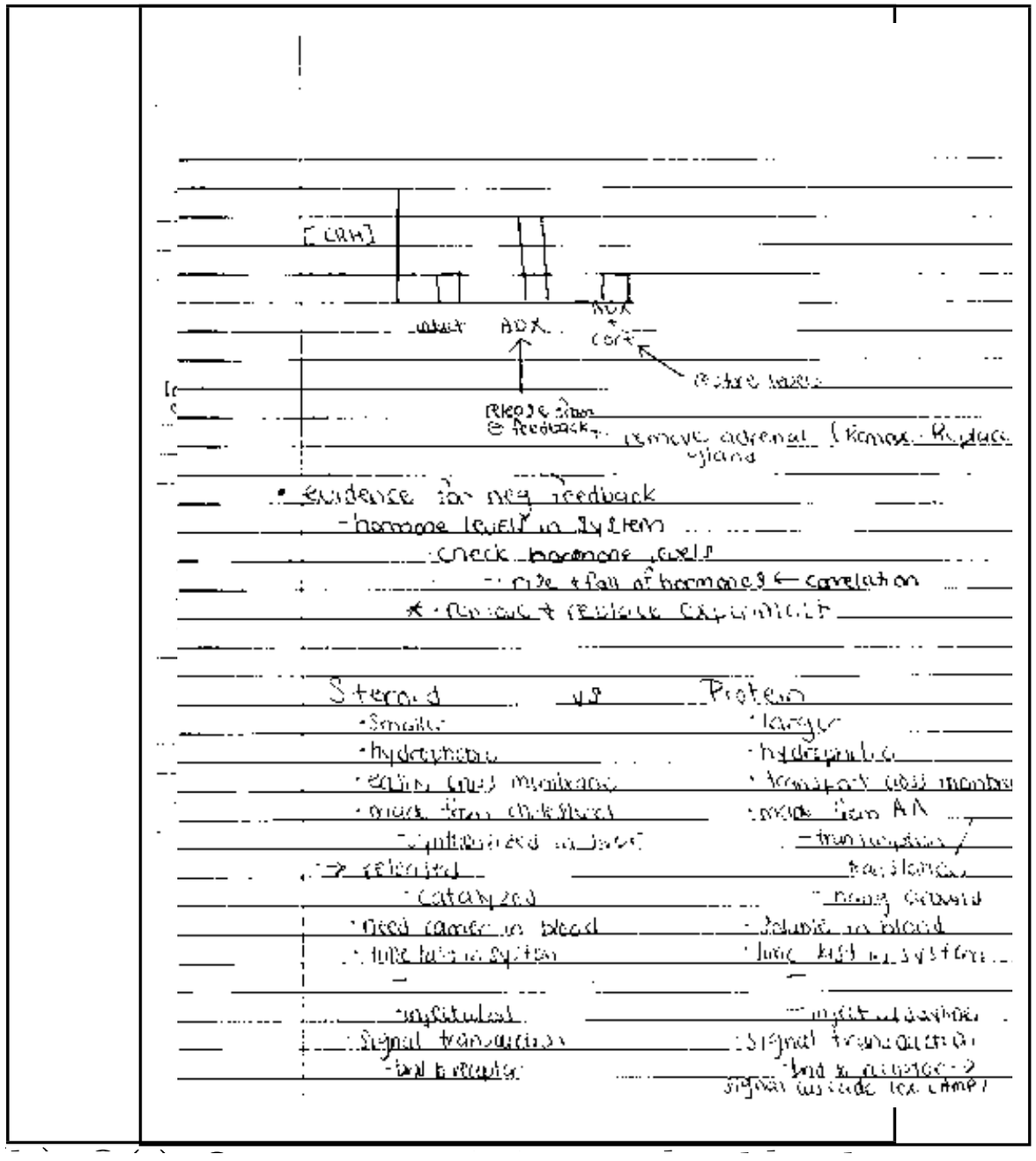
- *Page order*: logical sequence of pages that ought to be interpreted sequentially.
- In real life, page order is important for understanding an unstructured document collection, e.g., a set of loose pages.





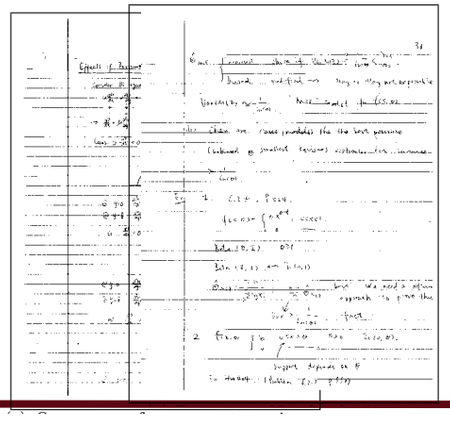
Structure Restoration

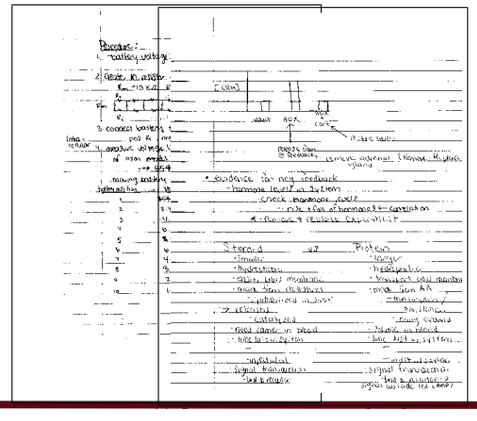




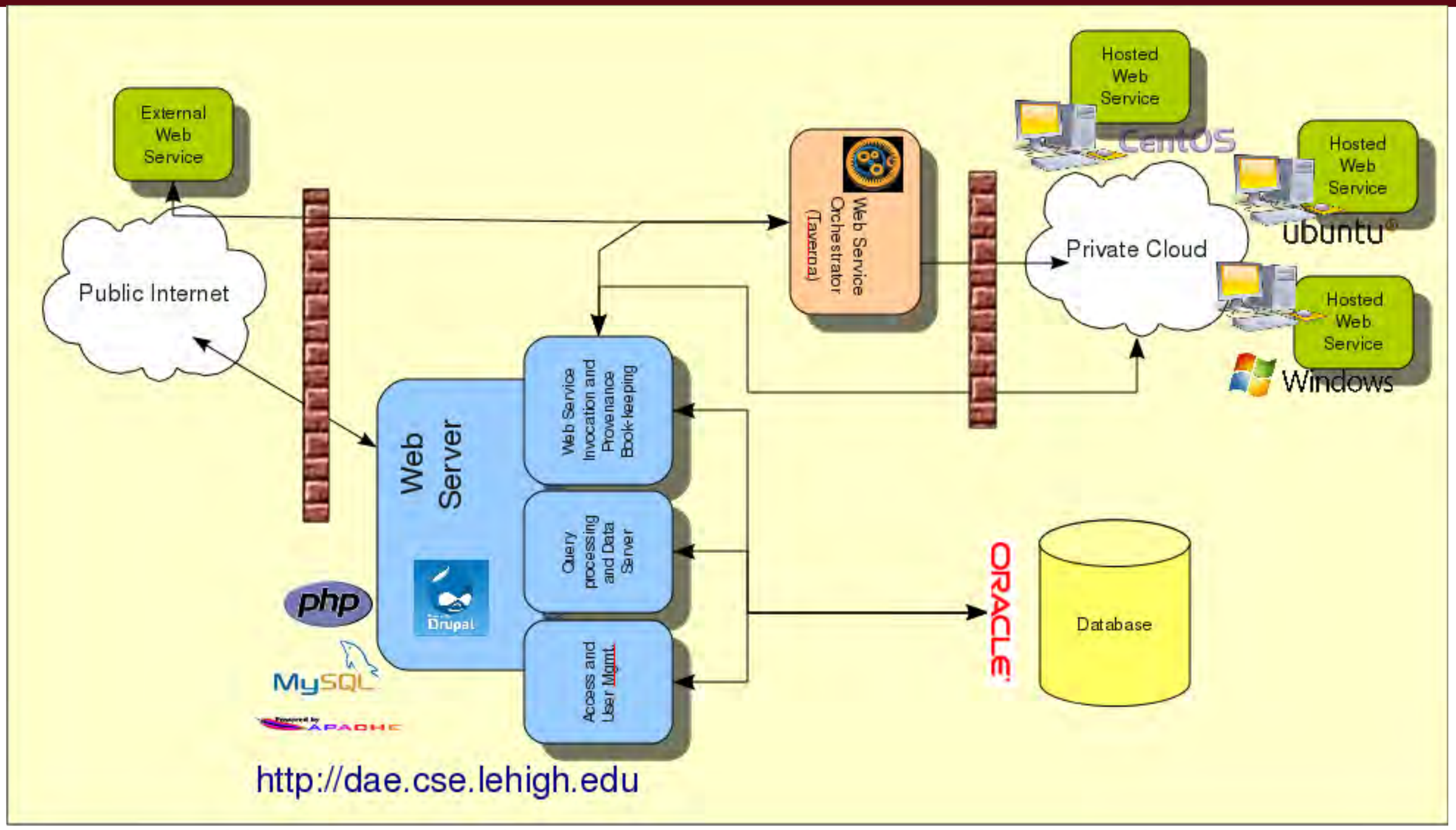
Structure Restoration

- Structure restoration decides which pages belong to separate physical/logical units, e.g., notebooks or topics.
- In real life, it is important for machines to employ customized techniques, e.g., style-based OCR/HWR.
- It is natural to use Lehigh Notebook dataset for such tasks. We have provided notebook IDs, pre-printed ruling line specifications, etc.





The DAE Platform



This slide is from the DRR 2011 talk by Lopresti and Lamiroy.

Document Analysis and Exploitation

Home Goals People Wiki

Document part annotation

Document tagging

Rating & commenting

Newsletters and discussion groups

This slide is from the DRR 2011 talk by Lopresti and Lamiroy.

XML Markup

```
<metadata>
    <page_image>
        <id>page294</id>
        <path>/TIFF/lehigh1003_nb2009_page294.tiff</path>
        <hdpi>600</hdpi>
        <vdpi>600</vdpi>
        <page_element>
            <value_list>
                <value_list_item>
                    <id>author</id>
                    <value>subject1003</value>
                </value_list_item>
                <value_list_item>
                    <id>notebookID</id>
                    <value>nb2009</value>
                </value_list_item>
                <value_list_item>
                    <id>creation date</id>
                    <value>2010/09/15</value>
                </value_list_item>
                <value_list_item>
                    <id>Subject</id>
                    <value>mathematics, statistics, linear models</value>
                </value_list_item>
                ...
            </value_list>
        </page_element>
    </page_image>
</metadata>
```

Authorship

Notebook ID

Subject Tags

Ruling Line Specifications, etc.

Conclusions

- We are motivated by the fact that most existing handwriting datasets are either elicited or curated.
- We aim for collecting 100 notebooks, 3k pages from 100 students. So far we have collected 18 notebooks from nine college students, in a total of 499 pages.
- Currently, the bitonal version if available via: *http://dae.cse.lehigh.edu/DAE/*. The full-color version will be uploaded soon.
- We also call for discussions on its usage.