AGENDA

1. Summary Statistics
2. ETD Architecture
3. ETD Services
   a. Validate Input
   b. Figure/Table Extraction
   c. Chapter Segmentation
   d. Text Extraction
   e. Classification
   f. Metadata Ingestion (ELS)
Summary Statistics

- Collection contains 19,779 theses and 14,691 dissertations.
  - Subsets contain 410 theses and 282 dissertations
- 19% and 32% of the 31 attributes are empty for theses & dissertations respectively
- We add attributes to ETD Metadata and create metadata for chapters
Summary Statistics

- Collection contains 19,779 theses and 14,691 dissertations.
  - Subsets contain 410 theses and 282 dissertations
- 19% and 32% of the 31 attributes are empty for thesis & dissertations respectively
- We add attributes to ETD Metadata and create metadata for chapters
- Fact Check: Issue Date highest in June
Summary Statistics

- Civil engineering
- Electrical engineering
- Molecular biology
- Public administration
- Educational leadership

Keywords:
- Design
- Control
- Space
- Agriculture
- Computation
- Access
- Security
- Method
- Model
- Architecture

AUTHOR

DEPARTMENT

SUBJECTS

Environmental Engineering
Mechanical Engineering
Civil Engineering
Electrical Engineering
Industrial Design
Science
Computer Engineering
Architecture
Political Science
Engineering Electrical

Theses

Dissertations
Validate Input

{handle: 9704} handle
{type: Thesis} doc_type

Creates a new row in the database for the incoming document.

INSERT INTO metadata...

Postgres DB

- contributor: \text{Silber, Joseph Allan}
- date-accessioned: 2011-06-14T14:45:39Z
- date-available: 2011-06-14T14:45:39Z
- date-issued: 2002-11-18
- identifier: \text{etd-12012002-165307}
- identifier-uri: \text{http://hdl.handle.net/10919/9704}
- description-abstract: The vestibular system of vertebrates consists of the utricle...
Validate Input

- **handle**: 9704
  - handle

- **type**: Thesis
  - doc_type

Creates a new row in the database for the incoming document.

**Future Works**

- Add support for more forms of ETD metadata (e.g. JSON, XML)
First we convert pdfs to an image format that is amenable to inference.
Inference is accomplished via the best performing model trained by Sampanna and others.

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the 3-D axis. Tapering at the base of stereocilia was

**Figure 2.2:** Six ciliate cells - electron micrograph and 3-D rendering.

**Figure 2.3:** Six ciliate cells - electron micrograph and 3-D rendering.
The model outputs bounding boxes that can be used to crop images.
Table/Figure Extraction

- The model outputs bounding boxes that can be used to crop images

Future Works

- Adapt the full deepfigures-open pipeline to Airflow
- Expose services to train and evaluate new figure extraction models
Chapter Segmentation

Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allen Silver

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of


in

Engineering Mechanics

J. Wallace Grant, Chair

Ethelene H. Peterson

John R. Cotton

November 18, 2002

Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silver

Median 7 Chapters
80% under 20 chapters
60% coverage of test data
Chapter Segmentation

Further testing
★
Improve coverage and accuracy
★
PyMuPDF

Future Works

Median 7 Chapters
80% under 20 chapters
60% coverage of test data
Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silver

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J. Wallace Grant, Chair
Ellenwenna H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silver

{handle: 9704} handle

{type: Thesis} doc_type
Text Extraction

- Get one page from PDF file.
Text Extraction

- Extract text from this PDF page. And then, write the text into a .txt file. Repeat this steps until finish all PDF pages.
Classification

Extracted Text

Dublin Core XML

Subject: ["Biomedical Engineering"]

Labels for ETD
Classification Architecture

- **Training Classifier**
  - Vectorizers trained on each label
  - Focused Text uses: title, abstract, and keywords
  - Used Logistic Regression to generate label probabilities per \( n \) word segment

- **Implementing Classifier on data**
  - Segment ETD into \( n \) word blocks
  - Classify each segment
  - Cumulate per label probability over segments
  - Pick top \( k \) labels based on top 5 labels’ variance
Ingestion

Read metadata from Postgres and index into Elasticsearch

```json
{
    "contributor_author": "Silber, Joseph Allan",
    "contributor_committeechair": "Grant, John Wallace",
    "contributor_committeemember": [
        "Peterson, Ellengone H.",
        "Cotton, John R."
    ],
    "contributor_department": "Engineering Science and Mechanics",
    "date_available": "2011-08-06T14:45:39Z",
    "subject": [
        "Finite Element",
        "Vestibular System",
        "Hair Cell",
        "Biomedical Engineering"
    ],
    "degree_grantor": "Virginia Polytechnic Institute and State University",
    "degree_level": "masters",
    "degree_name": "Master of Science",
    "description_abstract": "The vestibular system of vertebrates consists of the utricle, ...
    ...
}
```
Classification Metrics

- Test classifier with varied set of hyper-parameters
- Implement zones for metadata weighting
- Try different baseline models
Lessons Learned

- Understand Intended Structure Early
- Microservices Workflow
- Cross-team Communication

Next Steps

- Expose more Airflow goals to provide further customization
- Expose new Airflow goals to allow for separate collections in Elasticsearch