Client: Paul Mather

Virginia Tech CS4624, Blacksburg

March 4, 2014

By Nathanael Bice, Scott Brink & Adam Piorkowski

BTD IMPORTER

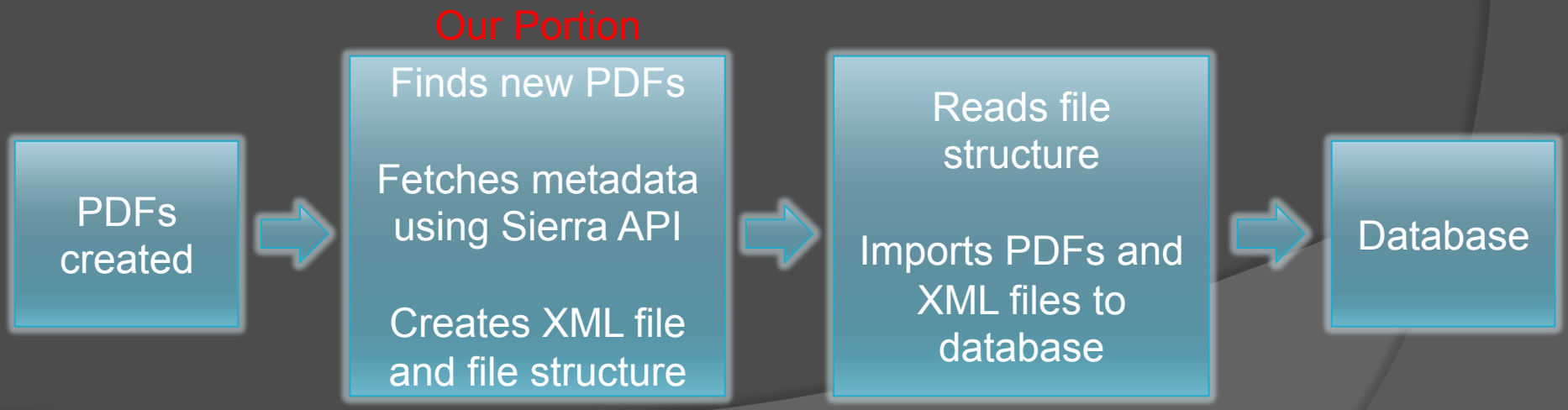
The Current Process

- Hard copy bound theses are scanned into an electronic PDF form
- Script looks for new or updated PDF files to add to database
- PDF file name includes Call Number
- Script fetches metadata for thesis based on Call Number
- PDFs with metadata are uploaded to database of theses

Our Job

- Rewrite the way PDFs have their metadata looked up
- Replace the use of Airpac Classic look up with the new Sierra API for metadata
- Create a file structure based on metadata and place PDF and XML file in correct folder
- A new importer will then take our file structure output and upload the files to a database

Before & After



Necessary Skills

- Java – Importer
- PHP – Current scanner and look up scripts
- SQL – Fetching metadata
- XML – Create an XML file filled with metadata for each thesis

Current Progress & Setbacks

- Written XML file format that will hold metadata of thesis
- Sierra API is not well documented and we do not have access to it
- Using current Addison system in place

End Goal

- Implement new Sierra API and rewrite importer script to find new PDFs
- Create a file directory that is sorted based on thesis metadata
- The file directory will then be read by the new importer to upload PDFs to database
- Import over 13,000 theses using the new system