Ingest System

Steve McDonald and the ADS Team

ADS Users Group Meeting, 20-21 Nov. 2019

Ingest System Overview

- **8 Individual Pipelines**
  - Bibliographic, Non-bibliographic
  - Affiliation, Citation Capture, Fulltext, myADS, Orcid
  - Master
- **Each Sends Processed Data To Master Pipeline**
  - Master merges and sends to persistent stores
    - Solr, SQL Database, API Endpoint
- **Tech Stack**
  - Python, Cron, Docker, RabbitMQ, Protobufs
- **Mostly Reuses Classic Files**
  - Needs to change!

Ingest Operations Overview

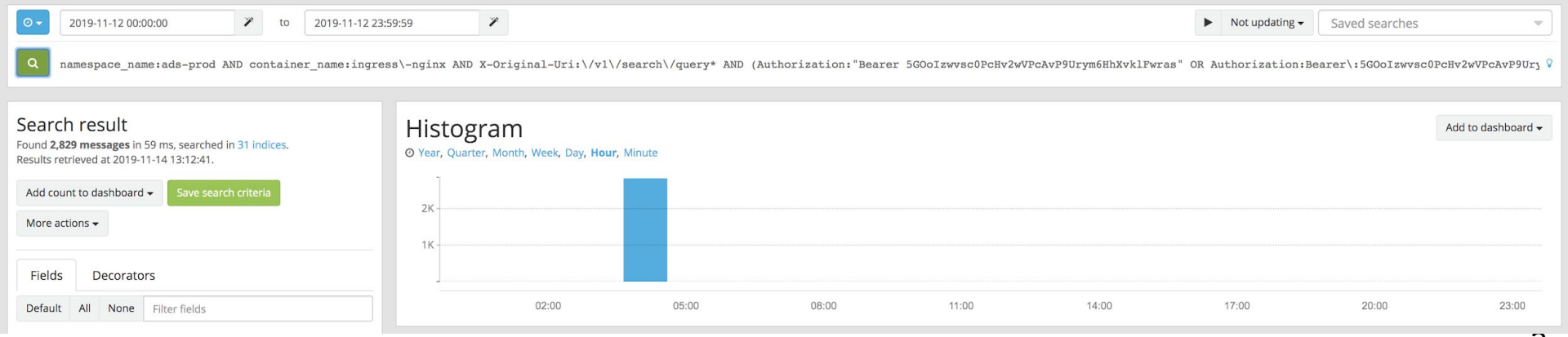
- **635,989 More Bibcodes**
- **8,676,103 More Citations**
- **Persistent Vigilance**
  - Immediate Action By Curators
  - Daily and Monthly Reports
- **Number Of Failed Records In Daily Ingest**
  - None: 38%
  - 1 to 5 Records: 42%
  - These Are Conservative Estimates!
  - Publisher And Non Curation Issues

Affiliation Pipeline

- **Affiliations Data Debuted @ 2019 Winter AAS**
- **Pipeline In Daily Operation Since ~ Spring 2019**
  - New records have aff_ids assigned daily
  - Existing records are updated as the dictionary of aff/aff_id is updated (~monthly)
- **Machine-Learning Based Curation**
  - Curation of a learning model is ongoing (slow, mainly hand-work)
  - Not yet in automated production, but can be used to assist curation
- **Statistics: See CSG Presentation (Thursday) On Affiliations**

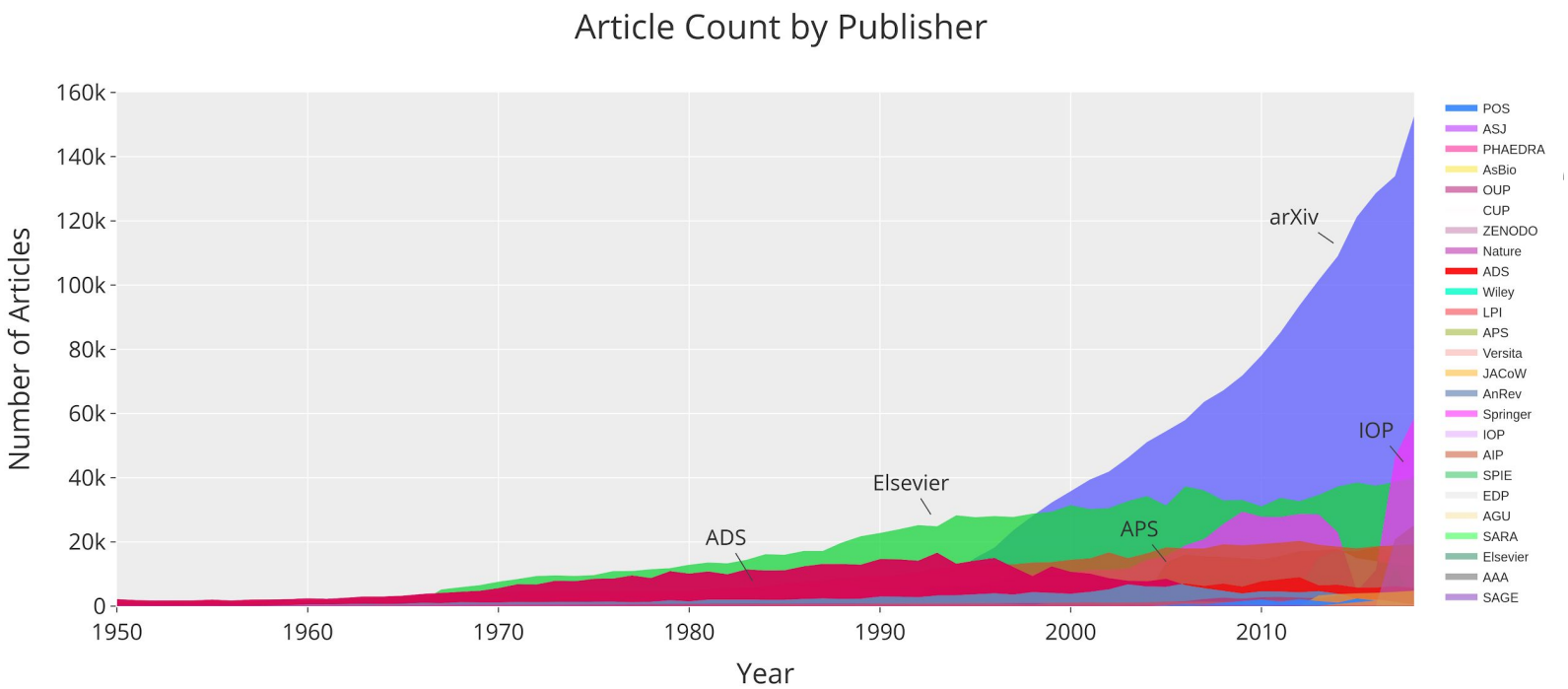
ORCID Pipeline

- **Restructure To Improve Performance**
  - Reduced Pipeline's API Search Calls
    - Requests from the ORCID pipeline were ~80% of search traffic
    - Reduced to 1% of search traffic
- **Changes To Support ORCID Microservice/UI Restructure**
  - Functionality migrated from UI to ORCID microservice



Fulltext Pipeline Overview

- **Over 5 Million Articles With Fulltext**
  - Exponential growth

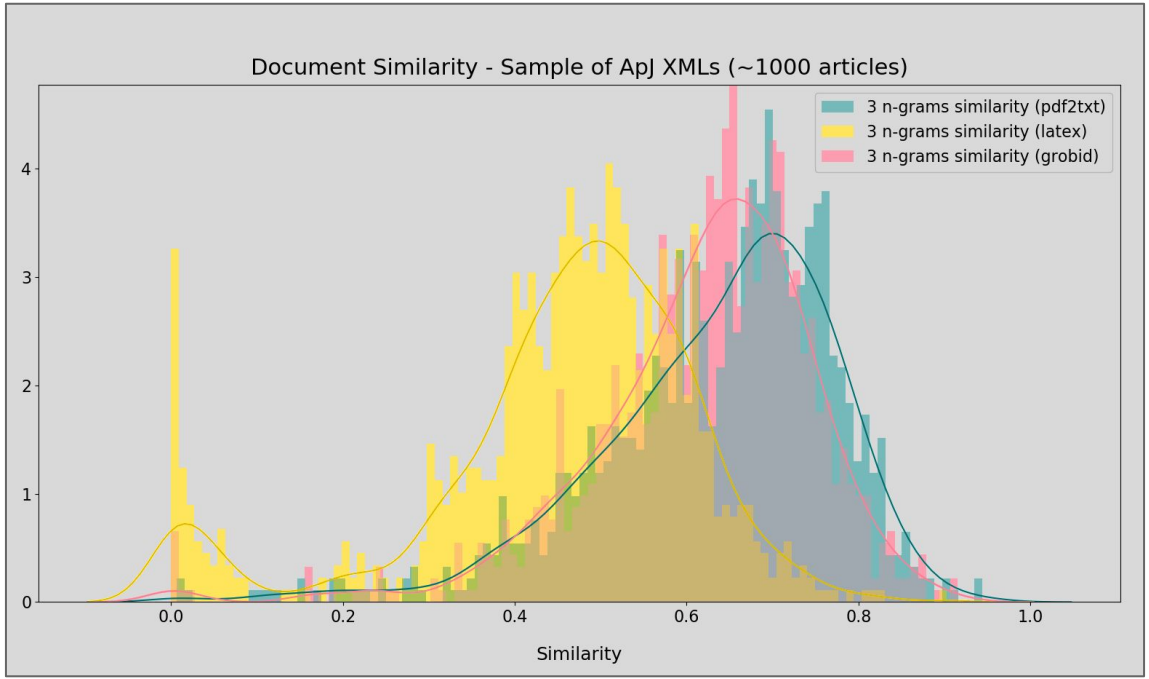


Fulltext Pipeline Updates

- Upgraded To Python 3 Compatible External Libraries
- More Robust XML Parsing
- Added Performance Monitoring

Fulltext Pipeline Updates

- **PDF Extraction Analysis**
  - **Validated Current Technique**



Ingest System Future

- **Reimplement Non-bibliographic Pipeline**
  - Non-metadata (reads, citations, metrics)
  - Changes .4M to 2M records daily
  - Investigations underway
  - Behind schedule
- **Create New Ingest System Plan**
  - Remove dependency on classic pipeline
    - Large effort
  - Consider framework like Kafka
- **Implement Plan**
  - Long running project

Ingest System

Steve McDonald and the ADS Team

ADS Users Group Meeting, 20-21 Nov. 2019

Data Timeliness

- **Server Room Issues**
  - Twice cooling system failed
  - Once network upgrade went poorly
- **Nonbib Failed With Bad Import Once Every 2 months**
  - Delays data reaching prod servers by ~10 hours
- **Changes To Data Links Field Not Automatically Detected**
- **End To End System Performance**
- **Misc**
  - Replication fail, disks full

Data Quality

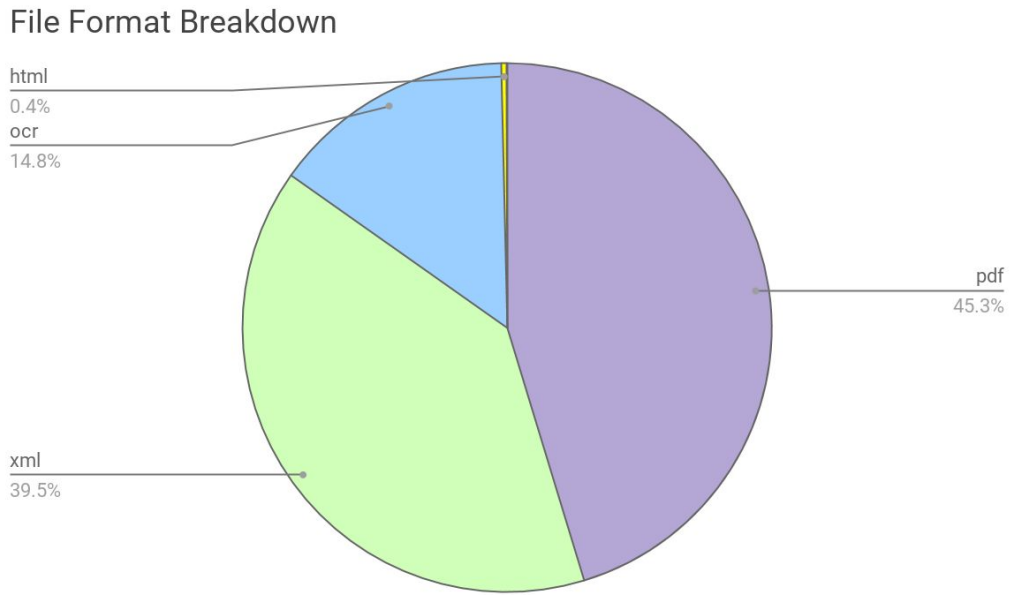
- **Generally Good**
- **We Avoid Repeated Bugs**
  - Root Cause Analysis
- **Any Outstanding Issues**
  - ?

Ingest System 2019 Highlights

- **New Affiliation Pipeline**
- **Enhancements To Orcid Pipeline**
- **Improvements To Fulltext Pipeline**
- **Bug Fixes To Improve Data Quality**

Fulltext Pipeline

File Formats



Fulltext Pipeline

Extraction Analysis