System Development
Backoffice Classic Replacement Progress

Kelly Lockhart, Peter Williams, Jenny Koch, Carolyn Grant, Edwin Henneken and the ADS Team

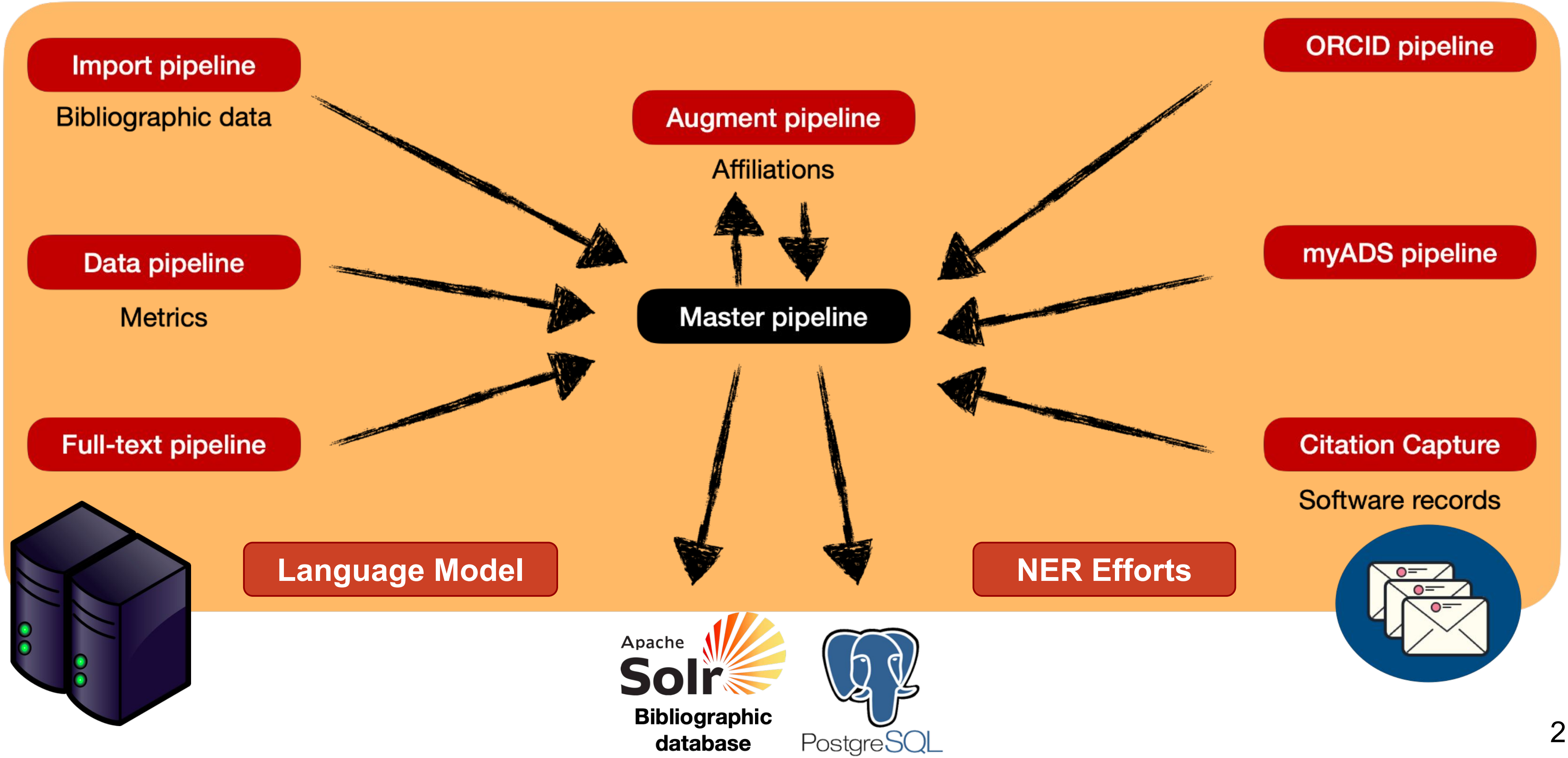
ADS Users Group Meeting, 9-10 Nov. 2022

Smithsonian Institution

astrophysics data system

NASA

Current pipelines



Import pipeline

Bibliographic data

Data pipeline

Metrics

Full-text pipeline

Augment pipeline

Affiliations

Master pipeline

ORCID pipeline

myADS pipeline

Citation Capture

Software records

Language Model

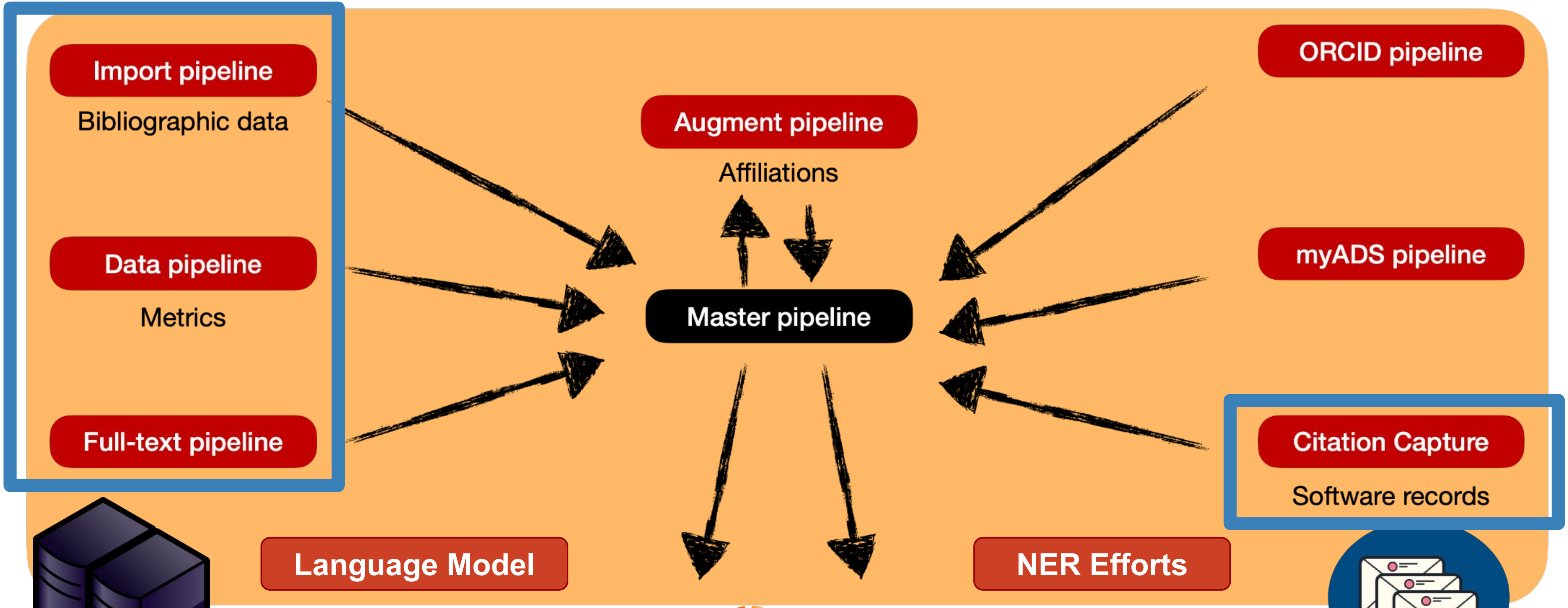
NER Efforts

Apache Solr

Bibliographic database

PostgreSQL

Current pipelines with legacy dependencies



Import pipeline

Bibliographic data

Data pipeline

Metrics

Full-text pipeline

Augment pipeline

Affiliations

Master pipeline

ORCID pipeline

myADS pipeline

Citation Capture

Software records

Language Model

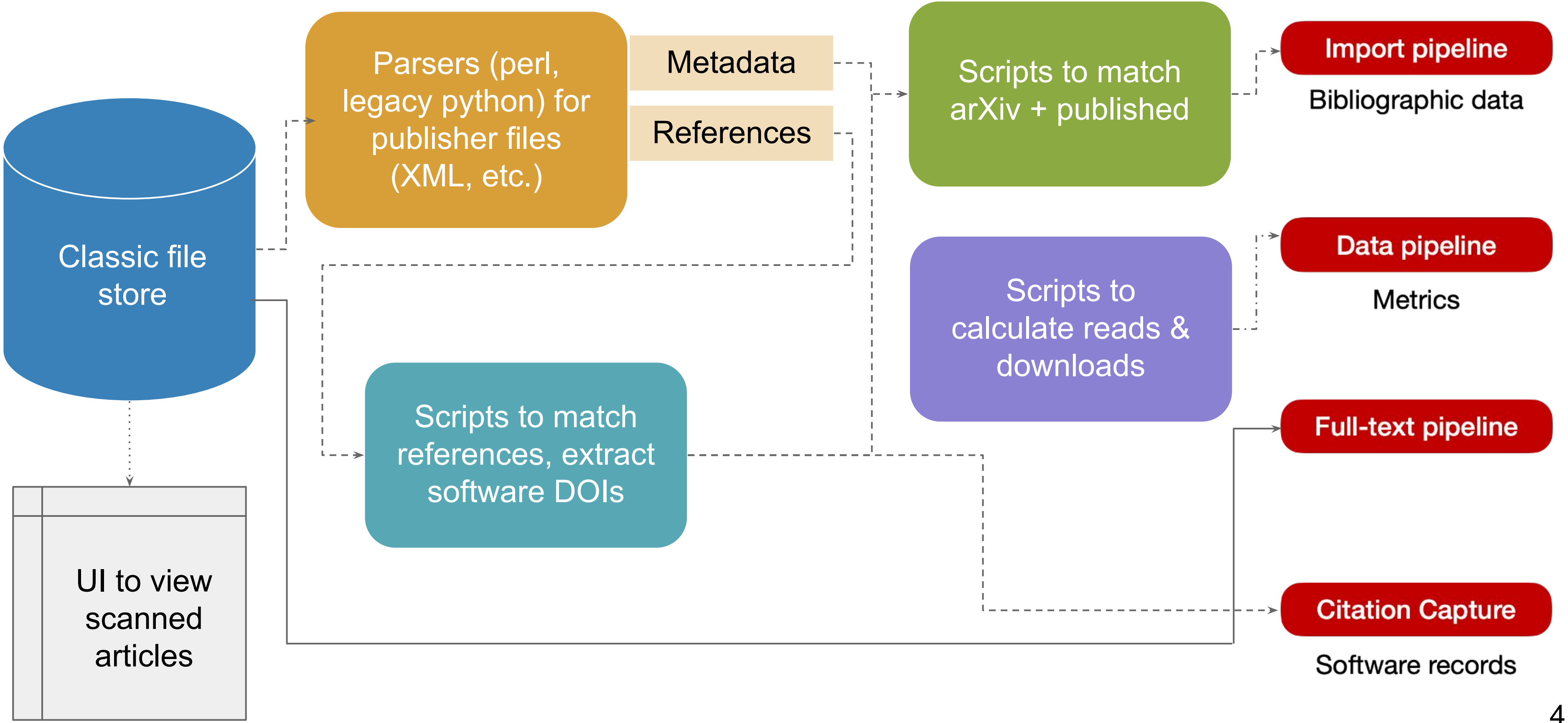
NER Efforts

Apache Solr

Bibliographic database

PostgreSQL

Legacy dependencies



Classic file store

Parsers (perl, legacy python) for publisher files (XML, etc.)

Metadata

References

Scripts to match arXiv + published

Scripts to calculate reads & downloads

Scripts to match references, extract software DOIs

UI to view scanned articles

Import pipeline

Bibliographic data

Data pipeline

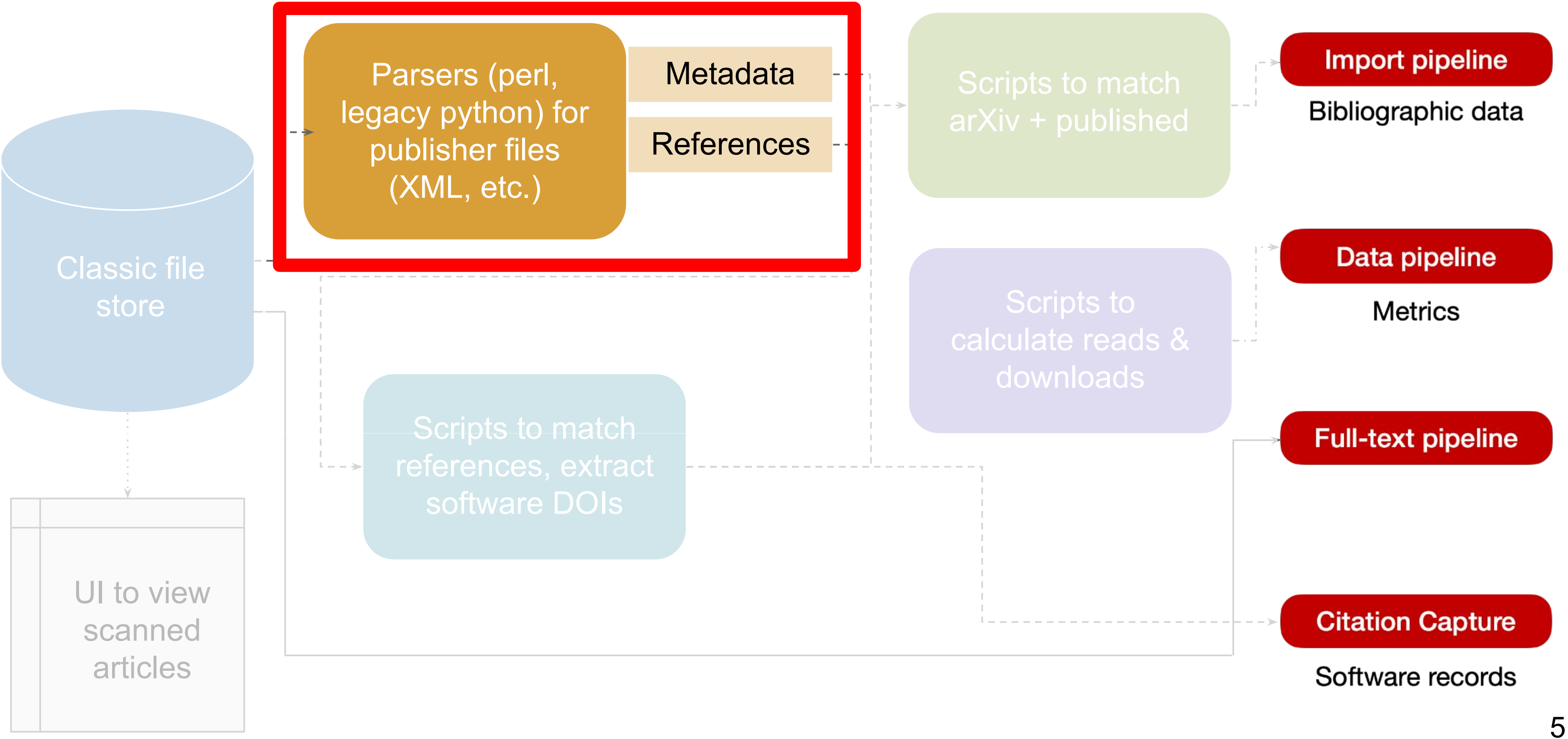
Metrics

Full-text pipeline

Citation Capture

Software records

Parsers


Parsers (perl, legacy python) for publisher files (XML, etc.)
Metadata
References
Scripts to match arXiv + published
Import pipeline
Bibliographic data
Classic file store
Scripts to calculate reads & downloads
Data pipeline
Metrics
Scripts to match references, extract software DOIs
Full-text pipeline
UI to view scanned articles
Citation Capture
Software records

Problems with legacy parsers

- Technical issues:
    - Inconsistencies in style, call syntax, and input/output
    - Old, not easily maintainable code: some in perl, some in legacy python
- Project management issues:
    - Parsing code is spread across multiple libraries/pipelines: code to parse metadata is separate from code to parse fulltext, for example, even though they parse the same input XML

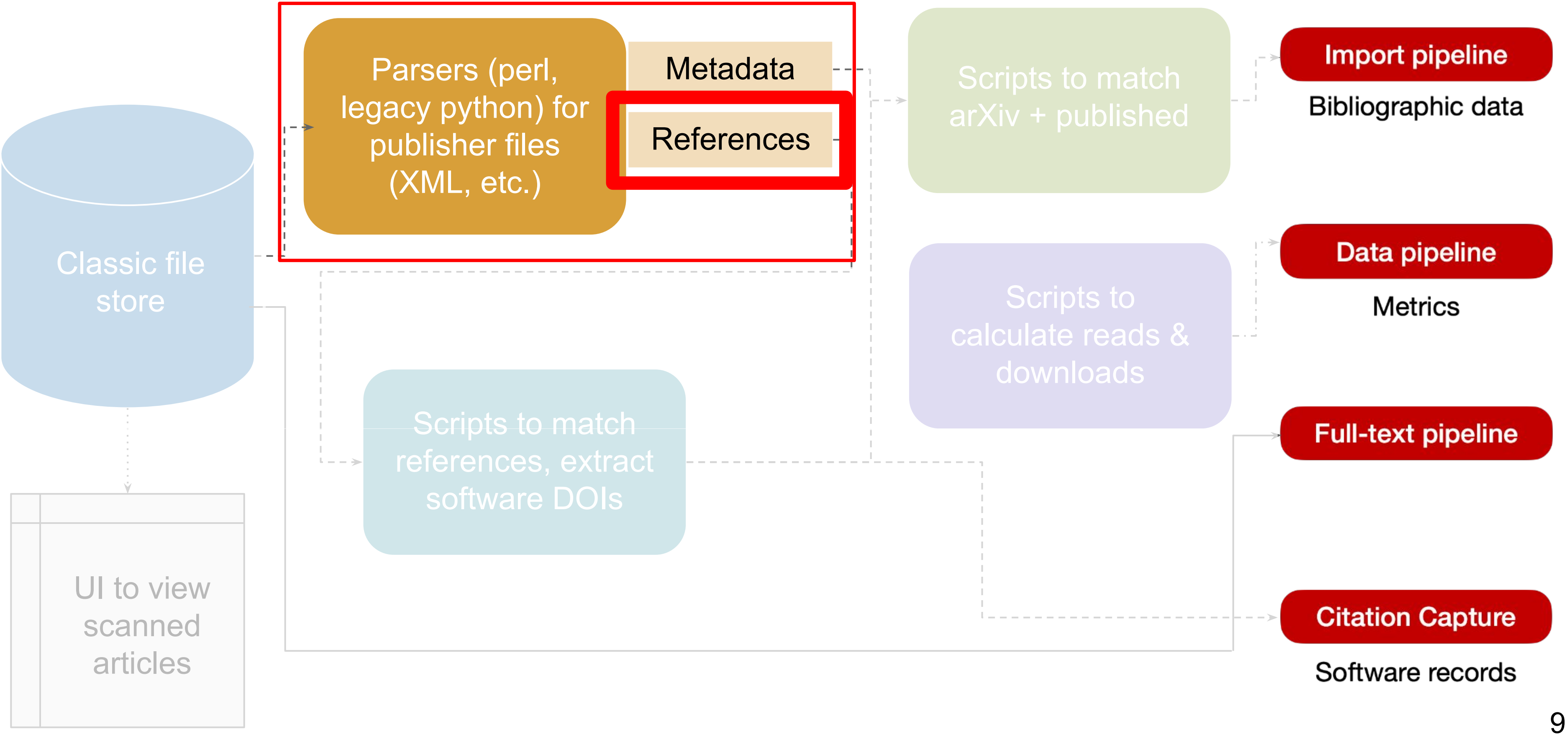
New parsers

- New parsing library is unified in style, syntax, input, and output (the new data model)
- Parsers completed and in testing
    - JATS, arXiv, Elsevier, DataCite, CrossRef, Wiley
    - These are ~¼ of total parsers, but cover ~⅔ of records ingested

Next steps

- Finish remaining parsers
  - 10 more in Perl, plus several more in legacy Python
- Start using new parsers in production
  - DataCite parser: ingestion script using it now, will be added to Citation Capture pipeline soon
  - Harvester pipeline, in progress, will be needed to fully make use of these, likely starting with one publisher at a time
- Long term goals
  - Port parsing code from other pipelines/libraries to the parsing library, such as fulltext parsing, and citation context for machine learning efforts

Reference Extraction



Parsers (perl, legacy python) for publisher files (XML, etc.)
Metadata
References
Scripts to match arXiv + published
Import pipeline
Bibliographic data
Classic file store
Data pipeline
Metrics
Scripts to calculate reads & downloads
Scripts to match references, extract software DOIs
Full-text pipeline
UI to view scanned articles
Citation Capture
Software records

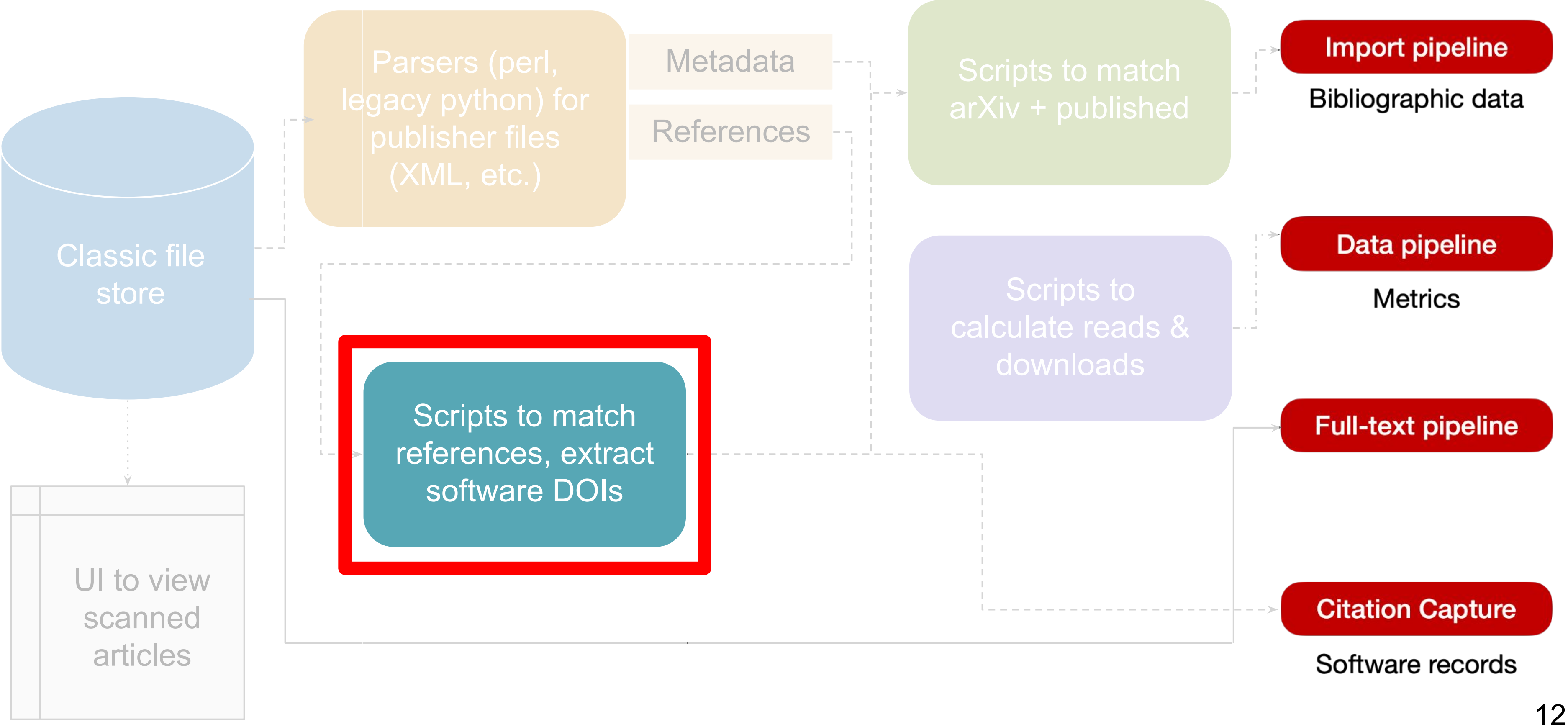
Reference Extraction: Context

- The ADS citation network doesn't build itself
    - Many data sources do not provide structured reference information
    - Notably ArXiv; a typical day adds tens of thousands of references
        - Ex: 2021-11-07 had 31,895
    - Fulltext ⇒ "refstrings" ⇒ bibcodes
    - Fulltext is generally PDF, but ArXiv has lots of TeX
        - 2021-11-07: 821 TeX, 122 PDF, 8 withdrawn, 12 fail processing
- Good reasons to update the current pipeline
    - A "classic" Perl framework
    - Very good at ArXiv TeX, but some known limitations (e.g. Unicode)
    - PDF extractor also good, but lots of new ML tools to bring to bear
    - PDF extraction *could* do lots more than references (abstracts, figures, ...)

Reference Extraction: Status

- ArXiv TeX extraction has been updated
  - Upgraded to ArXiv’s current TeX install
  - New Python implementation in modernized Docker framework
  - Robustness improvements for corner cases
  - Doing any better probably requires much more work (*real* TeX parsing)
- Modernized PDF extraction is nearing deployment
  - Industry-standard tool GROBID offers solid improvements
  - Other options investigated, not competitive
- Exploring future directions for generalized PDF extraction
  - Created ADS-tuned training set for testing/ranking approaches
  - GROBID’s models can be customized for ADS content
  - New tools are becoming available (e.g. ScienceParsePlus)
  - Likely makes more sense to adopt/collaborate than build from scratch

Reference Resolver


Classic file store
Parsers (perl, legacy python) for publisher files (XML, etc.)
Metadata
References
Scripts to match arXiv + published
Scripts to calculate reads & downloads
Scripts to match references, extract software DOIs
UI to view scanned articles
Import pipeline
Bibliographic data
Data pipeline
Metrics
Full-text pipeline
Citation Capture
Software records

Reference Resolver: Service & Pipeline

➢ Presented 2 years ago at ADSUG 2020 (PDF)

- Service (completed):
    - Reference Resolving: find record that matches a reference string input and outputs the bibcode and a computed confidence score
- Pipeline (future development):
    - Framework that will make use of Reference Service
        - Input new document’s bibliography and outputs matching documents as the linked references in ADS (citation graph)
- Goals for use:
    - Replacing classic machinery
    - Content & curation support

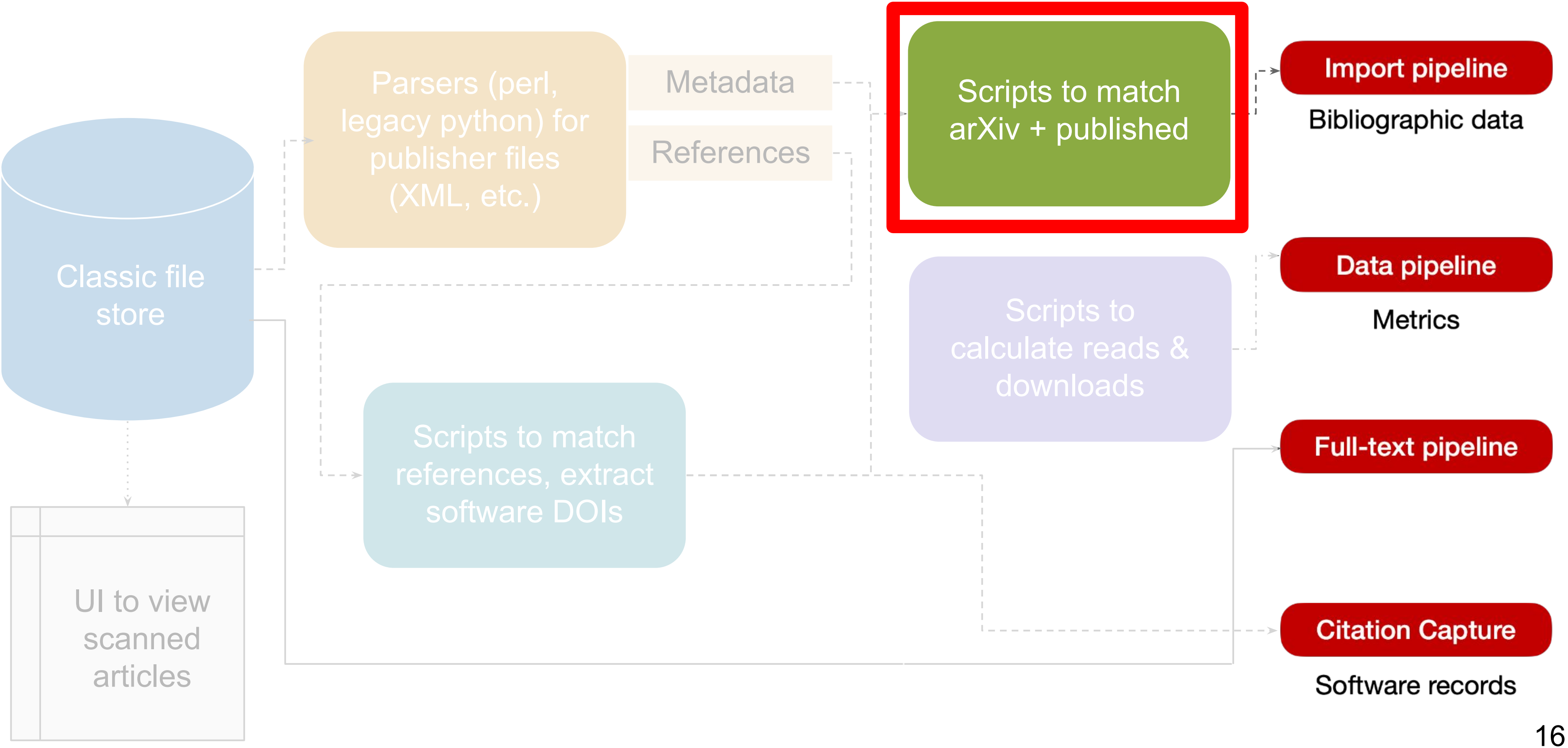
Reference Service for Content & Curation

- Identifying Coverage Gaps
  - ARC/Space Science & Astrobiology Division - Blog post (Nov 2021) details project
    - Matching ARC/SS Division bibliography with ADS content for coverage of NASA papers
    - ADS Ref Service API - match by reference strings (Author + Year + Publication)
      - This found the majority of results
    - ADS search API - match additional by DOI or Title
  - HOLLIS Harvester
    - Project detailed in Jenny's GitHub HOLLIS Harvester documentation
      - Matching HOLLIS monographs with ADS content for coverage of gray literature
    - Searched Ref Service API with reference strings (Author + Title + Pub Year)
    - Match existing items; reviewed unmatched/new items for curation and ingest

Reference Resolver: Development

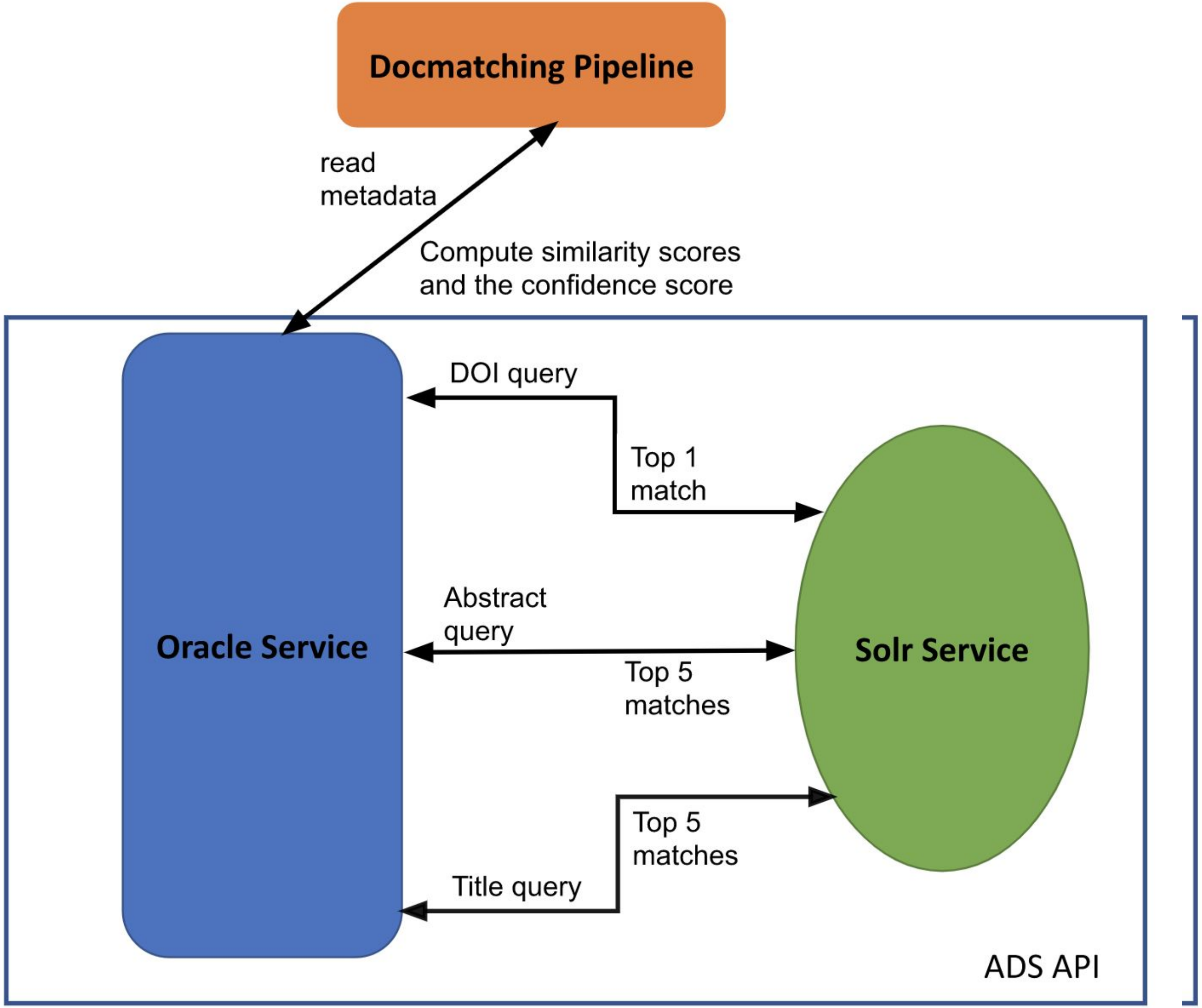
- Reference service is complete and its accuracy exceeds ADS Classic’s resolver
- Future work:
    - Parsers for some smaller publishers’ references (e.g. conferences, etc.)
    - Integrate into pipeline infrastructure

Docmatcher


Classic file store
UI to view scanned articles
Parsers (perl, legacy python) for publisher files (XML, etc.)
Metadata
References
Scripts to match arXiv + published
Scripts to calculate reads & downloads
Scripts to match references, extract software DOIs
Import pipeline
Bibliographic data
Data pipeline
Metrics
Full-text pipeline
Citation Capture
Software records

ArXiv matching with new docmatcher

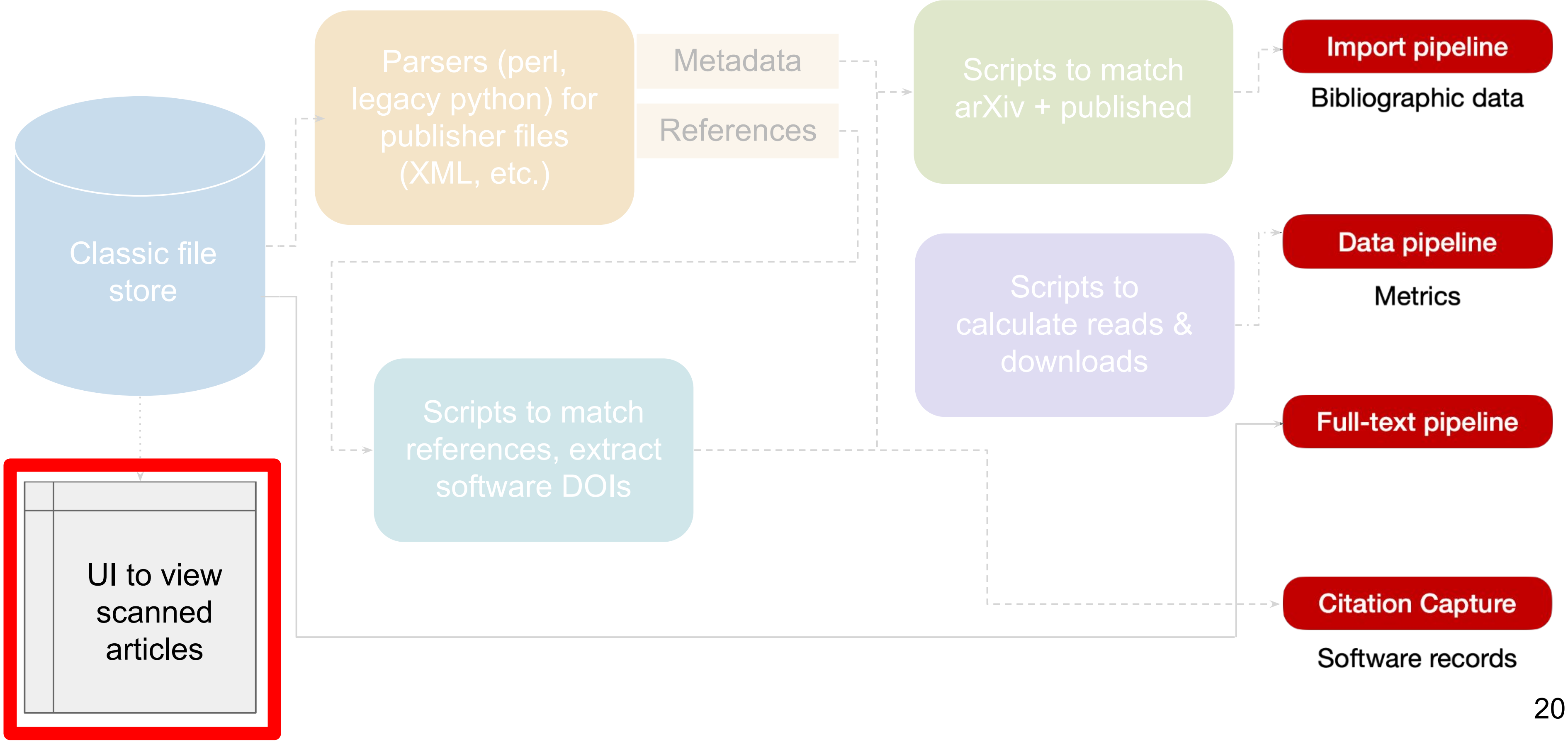
- In production as of October, 2022
    - Running in parallel with classic matching
    - 116,500 arXiv in October, 2022
    - Both classic and arxiv matched on the order of 8% (slightly more with docmatcher) but docmatcher’s are more accurate
    - 20% checked to help determine threshold for accepting
    - Correct matches at > 97%
    - Blog post September, 2022


Docmatching Pipeline
read
metadata
Compute similarity scores
and the confidence score
Oracle Service
DOI query
Top 1
match
Abstract
query
Top 5
matches
Solr Service
Top 5
matches
Title query
ADS API

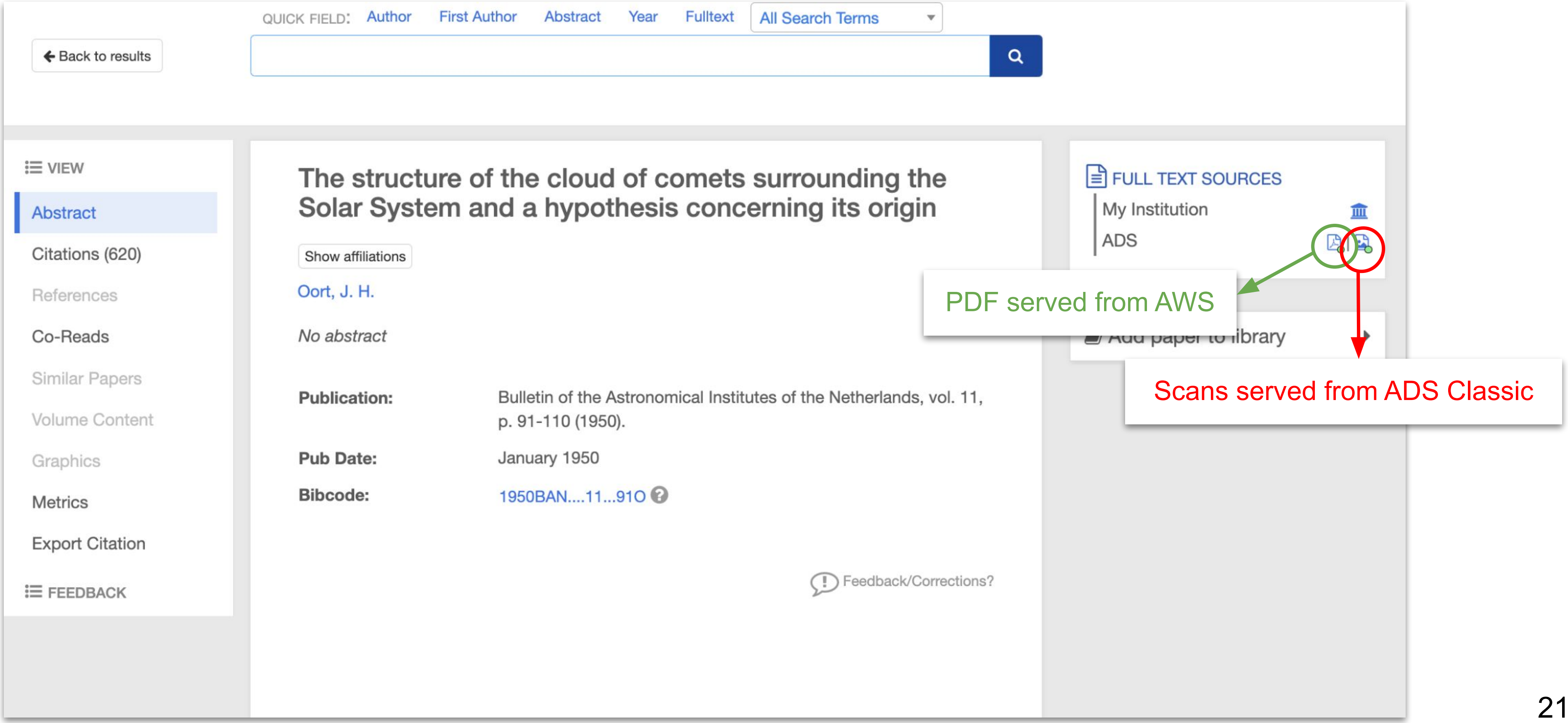
ArXiv matching with new docmatcher

- Next Steps
  - Automate curated matches back into system
  - Remove classic matching from indexing processes
    - Will speed up classic indexing (probably by a factor of 2)
  - Turn off classic matching December, 2022
    - Will reduce number of user submissions
    - Will reduce number of user corrections

Scan Explorer


Parsers (perl, legacy python) for publisher files (XML, etc.)
Metadata
References
Scripts to match arXiv + published
Import pipeline
Bibliographic data
Classic file store
Data pipeline
Metrics
Scripts to calculate reads & downloads
Scripts to match references, extract software DOIs
Full-text pipeline
UI to view scanned articles
Citation Capture
Software records

ADS Digitization Efforts



The structure of the cloud of comets surrounding the Solar System and a hypothesis concerning its origin

Oort, J. H.

No abstract

Publication: Bulletin of the Astronomical Institutes of the Netherlands, vol. 11, p. 91-110 (1950).

Pub Date: January 1950

Bibcode: 1950BAN....11...91O

PDF served from AWS

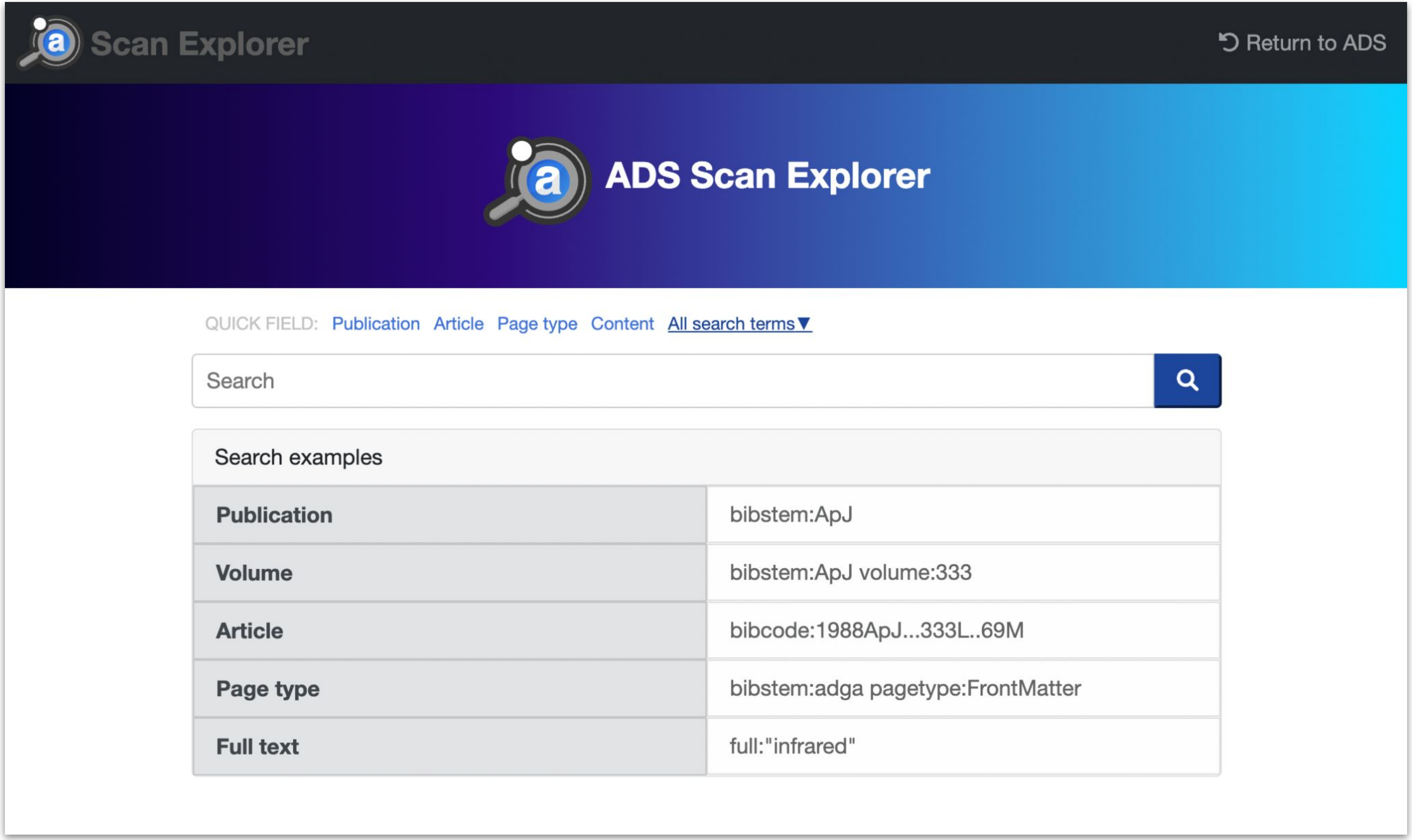
Scans served from ADS Classic

ADS Scan Explorer

- Deployment:
  - The different components of the ADS Scan Explorer should be easily deployable as Docker containers.
- Pipeline/Provisioning:
  - The ADS Scan Explorer is required to have infrastructure that will allow it to be provisioned with new data and which will allow standard data operations on existing content.
- API/Image Server:
  - API architecture for serving images based on IIIF standards
- User Interface/ADS-branded viewer:
  - Image viewer compatible with IIIF standards and current ADS UI software (Mirador, https://projectmirador.org/)

Implementation: outsourced to Winter Way based on a SOW written by the ADS

ADS Scan Explorer (https://dev.adsabs.harvard.edu/scan)

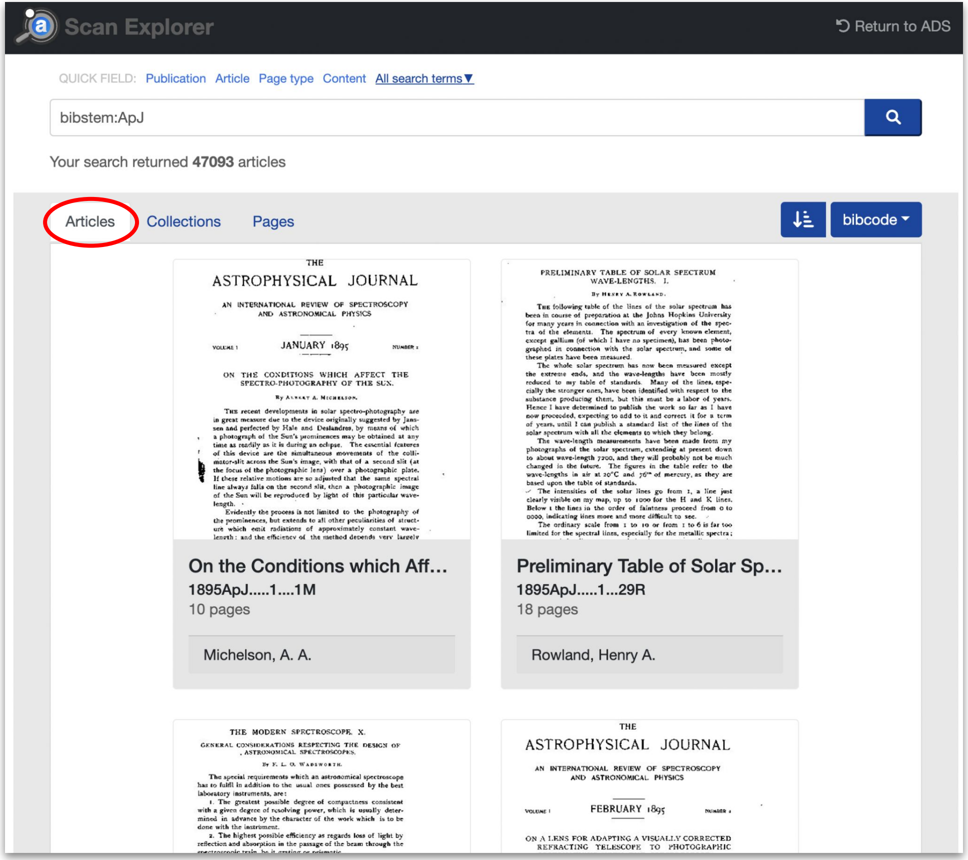


Scan Explorer
Return to ADS
ADS Scan Explorer
QUICK FIELD: Publication Article Page type Content All search terms▼
Search

Search examples

Publication	bibstem:ApJ
Volume	bibstem:ApJ volume:333
Article	bibcode:1988ApJ...333L..69M
Page type	bibstem:adga pagetype:FrontMatter
Full text	full:"infrared"

ADS Scan Explorer - Article View



Scan Explorer
Return to ADS
QUICK FIELD: Publication Article Page type Content All search terms▼

bibstem:ApJ

Your search returned **47093** articles

Articles
Collections
Pages
bibcode

THE ASTROPHYSICAL JOURNAL
AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS
JANUARY 1895
ON THE CONDITIONS WHICH AFFECT THE SPECTRO-PHOTOGRAPHY OF THE SUN.

On the Conditions which Aff…
1895ApJ.....1....1M
10 pages
Michelson, A. A.

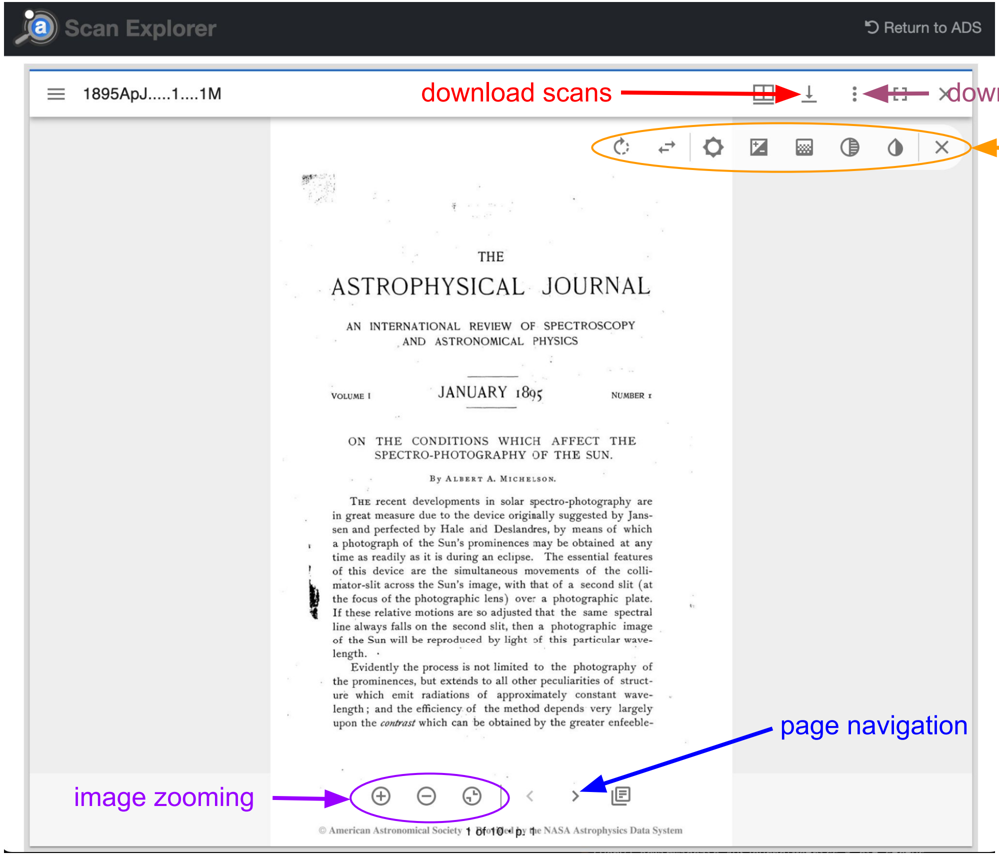
PRELIMINARY TABLE OF SOLAR SPECTRUM WAVE-LENGTHS. I.

Preliminary Table of Solar Sp…
1895ApJ.....1...29R
18 pages
Rowland, Henry A.

THE MODERN SPECTROSCOPE. X.
GENERAL CONSIDERATIONS RESPECTING THE DESIGN OF ASTRONOMICAL SPECTROSCOPES.

THE ASTROPHYSICAL JOURNAL
AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS
FEBRUARY 1895
ON A LENS FOR ADAPTING A VISUALLY CORRECTED REFRACTING TELESCOPE TO PHOTOGRAPHIC

ADS Scan Explorer - Article View



THE

ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY
AND ASTRONOMICAL PHYSICS

VOLUME I — JANUARY 1895 — NUMBER 1

ON THE CONDITIONS WHICH AFFECT THE
SPECTRO-PHOTOGRAPHY OF THE SUN.

By ALBERT A. MICHELSON.

THE recent developments in solar spectro-photography are in great measure due to the device originally suggested by Janssen and perfected by Hale and Deslandres, by means of which a photograph of the Sun's prominences may be obtained at any time as readily as it is during an eclipse. The essential features of this device are the simultaneous movements of the collimator-slit across the Sun's image, with that of a second slit (at the focus of the photographic lens) over a photographic plate. If these relative motions are so adjusted that the same spectral line always falls on the second slit, then a photographic image of the Sun will be reproduced by light of this particular wavelength.

Evidently the process is not limited to the photography of the prominences, but extends to all other peculiarities of structure which emit radiations of approximately constant wavelength; and the efficiency of the method depends very largely upon the *contrast* which can be obtained by the greater enfeeble-

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1 of 10 • p. 1

download scans

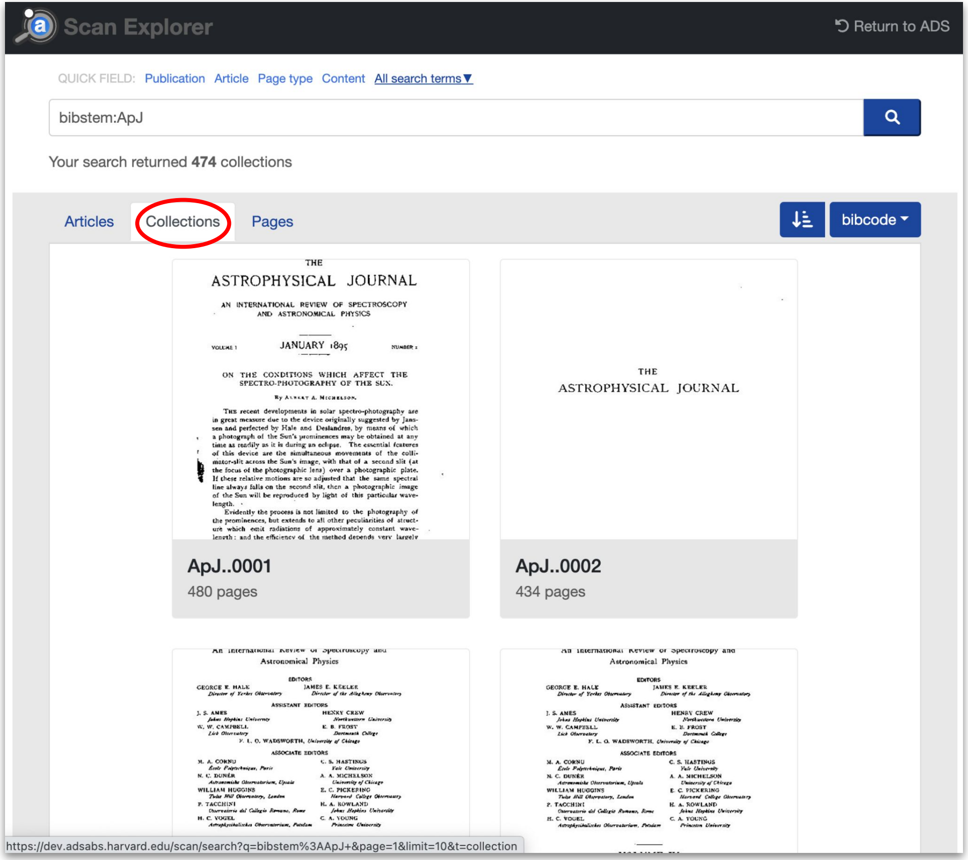
download OCR

image manipulation

image zooming

page navigation

ADS Scan Explorer - Collections View



Scan Explorer
Return to ADS
QUICK FIELD: Publication Article Page type Content All search terms▼
Articles
Collections
Pages
bibcode

bibstem:ApJ

Your search returned **474** collections

THE
ASTROPHYSICAL JOURNAL
AN INTERNATIONAL REVIEW OF SPECTROSCOPY
AND ASTRONOMICAL PHYSICS

VOLUME 1 JANUARY 1895 NUMBER 1

ON THE CONDITIONS WHICH AFFECT THE SPECTRO-PHOTOGRAPHY OF THE SUN.

By ALBERT A. MICHELSON.

THE recent developments in solar spectro-photography are in great measure due to the device originally suggested by Janssen and perfected by Hale and Deslandres, by means of which a photograph of the Sun's prominences may be obtained at any time as readily as it is during an eclipse. The essential features of this device are the simultaneous movements of the collimator-slit across the Sun's image, with that of a second slit (at the focus of the photographic lens) over a photographic plate. If these relative motions are so adjusted that the same spectral line always falls on the second slit, then a photographic image of the Sun will be reproduced by light of this particular wavelength.

Evidently the process is not limited to the photography of the prominences, but extends to all other peculiarities of structure which emit radiations of approximately constant wavelength: and the efficiency of the method depends very largely

ApJ..0001
480 pages

THE
ASTROPHYSICAL JOURNAL

ApJ..0002
434 pages

An International Review of Spectroscopy and
Astronomical Physics

EDITORS

GEORGE E. HALE, *Director of Yerkes Observatory*
JAMES E. KEELER, *Director of the Allegheny Observatory*

ASSISTANT EDITORS

J. S. AMES, *Johns Hopkins University*
HENRY CREW, *Northwestern University*
W. W. CAMPBELL, *Lick Observatory*
E. B. FROST, *Dartmouth College*
F. L. O. WADSWORTH, *University of Chicago*

ASSOCIATE EDITORS

M. A. CORNU, *École Polytechnique, Paris*
C. S. HASTINGS, *Yale University*
N. C. DUNÉR, *Astronomiska Observatorium, Upsala*
A. A. MICHELSON, *University of Chicago*
WILLIAM HUGGINS, *Tulse Hill Observatory, London*
E. C. PICKERING, *Harvard College Observatory*
P. TACCHINI, *Osservatorio del Collegio Romano, Rome*
H. A. ROWLAND, *Johns Hopkins University*
H. C. VOGEL, *Astrophysikalisches Observatorium, Potsdam*
C. A. YOUNG, *Princeton University*

An International Review of Spectroscopy and
Astronomical Physics

EDITORS

GEORGE E. HALE, *Director of Yerkes Observatory*
JAMES E. KEELER, *Director of the Allegheny Observatory*

ASSISTANT EDITORS

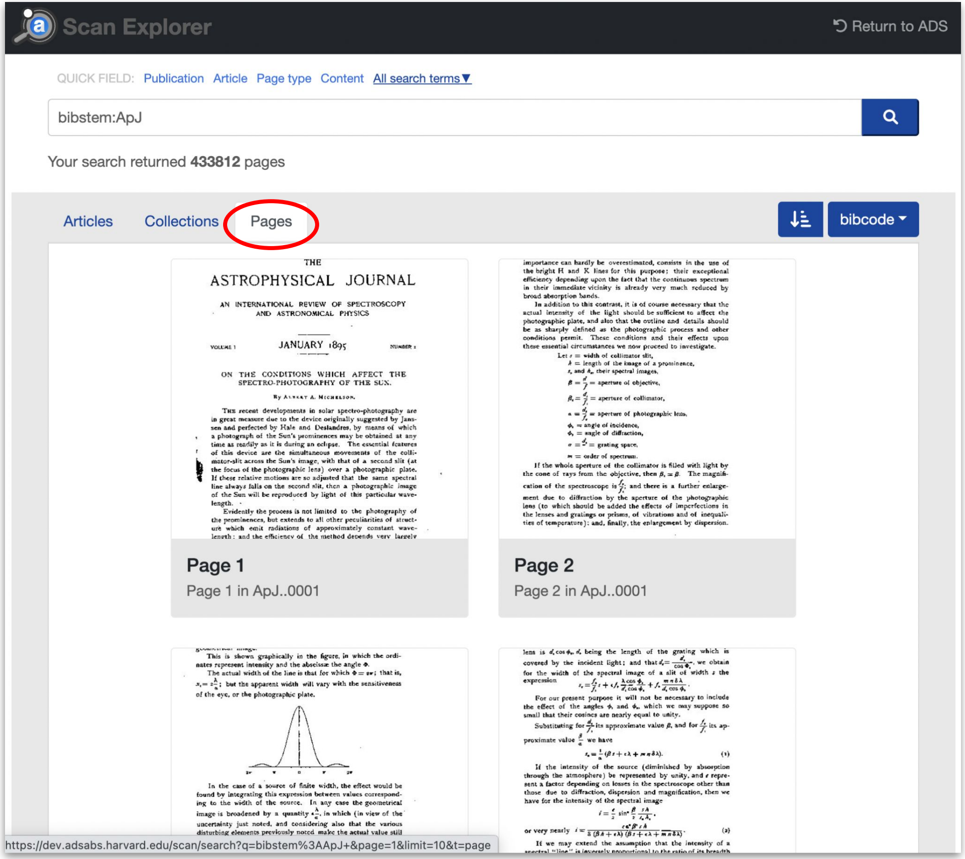
J. S. AMES, *Johns Hopkins University*
HENRY CREW, *Northwestern University*
W. W. CAMPBELL, *Lick Observatory*
E. B. FROST, *Dartmouth College*
F. L. O. WADSWORTH, *University of Chicago*

ASSOCIATE EDITORS

M. A. CORNU, *École Polytechnique, Paris*
C. S. HASTINGS, *Yale University*
N. C. DUNÉR, *Astronomiska Observatorium, Upsala*
A. A. MICHELSON, *University of Chicago*
WILLIAM HUGGINS, *Tulse Hill Observatory, London*
E. C. PICKERING, *Harvard College Observatory*
P. TACCHINI, *Osservatorio del Collegio Romano, Rome*
H. A. ROWLAND, *Johns Hopkins University*
H. C. VOGEL, *Astrophysikalisches Observatorium, Potsdam*
C. A. YOUNG, *Princeton University*

https://dev.adsabs.harvard.edu/scan/search?q=bibstem%3AApJ+&page=1&limit=10&t=collection

ADS Scan Explorer - Pages View


Scan Explorer
Return to ADS
QUICK FIELD: Publication Article Page type Content All search terms▼
Articles
Collections
Pages
bibcode

bibstem:ApJ

Your search returned **433812** pages

THE

ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

VOLUME 1 JANUARY 1895 NUMBER 1

ON THE CONDITIONS WHICH AFFECT THE SPECTRO-PHOTOGRAPHY OF THE SUN.

By ALBERT A. MICHELSON.

THE recent developments in solar spectro-photography are in great measure due to the device originally suggested by Janssen and perfected by Hale and Deslandres, by means of which a photograph of the Sun's prominences may be obtained at any time as readily as it is during an eclipse. The essential features of this device are the simultaneous movements of the collimator-slit across the Sun's image, with that of a second slit (at the focus of the photographic lens) over a photographic plate. If these relative motions are so adjusted that the same spectral line always falls on the second slit, then a photographic image of the Sun will be reproduced by light of this particular wavelength.

Evidently the process is not limited to the photography of the prominences, but extends to all other peculiarities of structure which emit radiations of approximately constant wavelength: and the efficiency of the method depends very largely

Page 1

Page 1 in ApJ..0001

importance can hardly be overestimated, consists in the use of the bright H and K lines for this purpose; their exceptional efficiency depending upon the fact that the continuous spectrum in their immediate vicinity is already very much reduced by broad absorption bands.

In addition to this contrast, it is of course necessary that the actual intensity of the light should be sufficient to affect the photographic plate, and also that the outline and details should be as sharply defined as the photographic process and other conditions permit. These conditions and their effects upon these essential circumstances we now proceed to investigate.

Let s = width of collimator slit,
h = length of the image of a prominence,
s_s and h_s, their spectral images,
$\beta = \frac{d}{f}$ = aperture of objective,
$\beta_1 = \frac{d_1}{f_1}$ = aperture of collimator,

Page 2

Page 2 in ApJ..0001

https://dev.adsabs.harvard.edu/scan/search?q=bibstem%3AApJ+&page=1&limit=10&t=page

Final thoughts

- ADS Scan Explorer functionality
  - Supports all functionality of Classic interface
  - Additional functionality (e.g. search & download OCR text)
  - Mirador image viewer supports plugins (e.g. annotation)
- Goal: in production by end of calendar year
- Outsourcing success story