

System Development

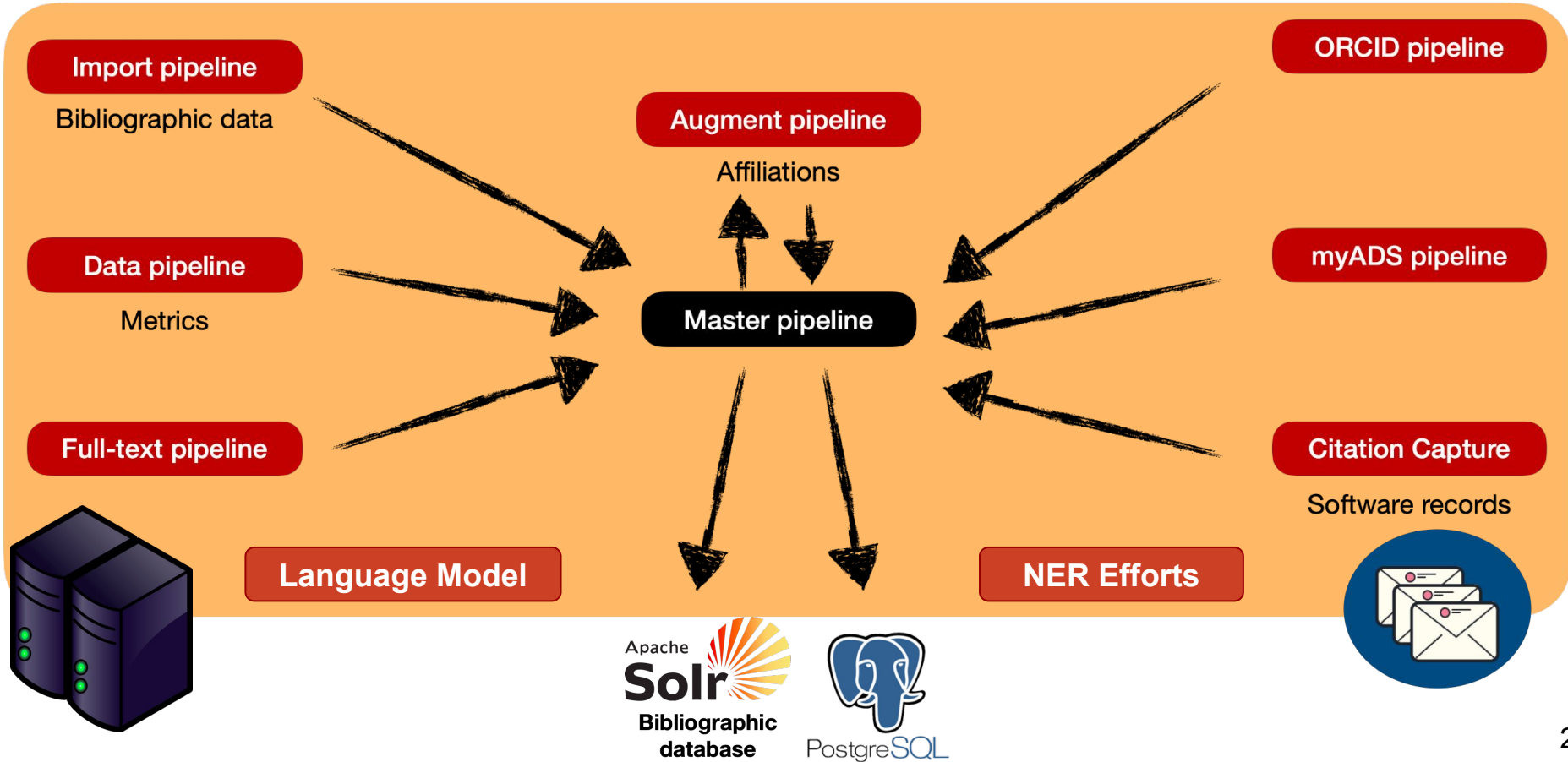
Backoffice Classic Replacement Progress

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

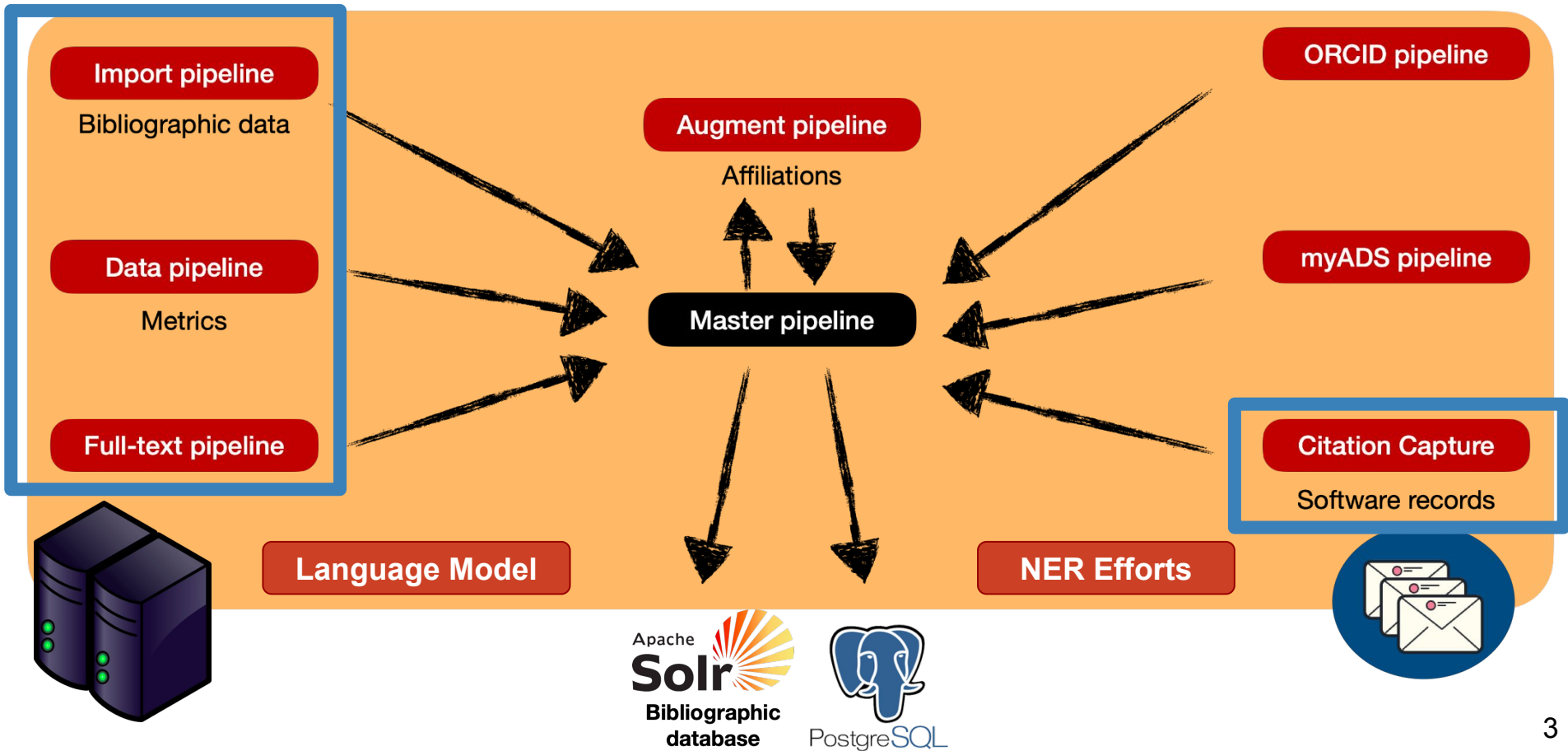
ADS Users Group Meeting, 9-10 Nov. 2022



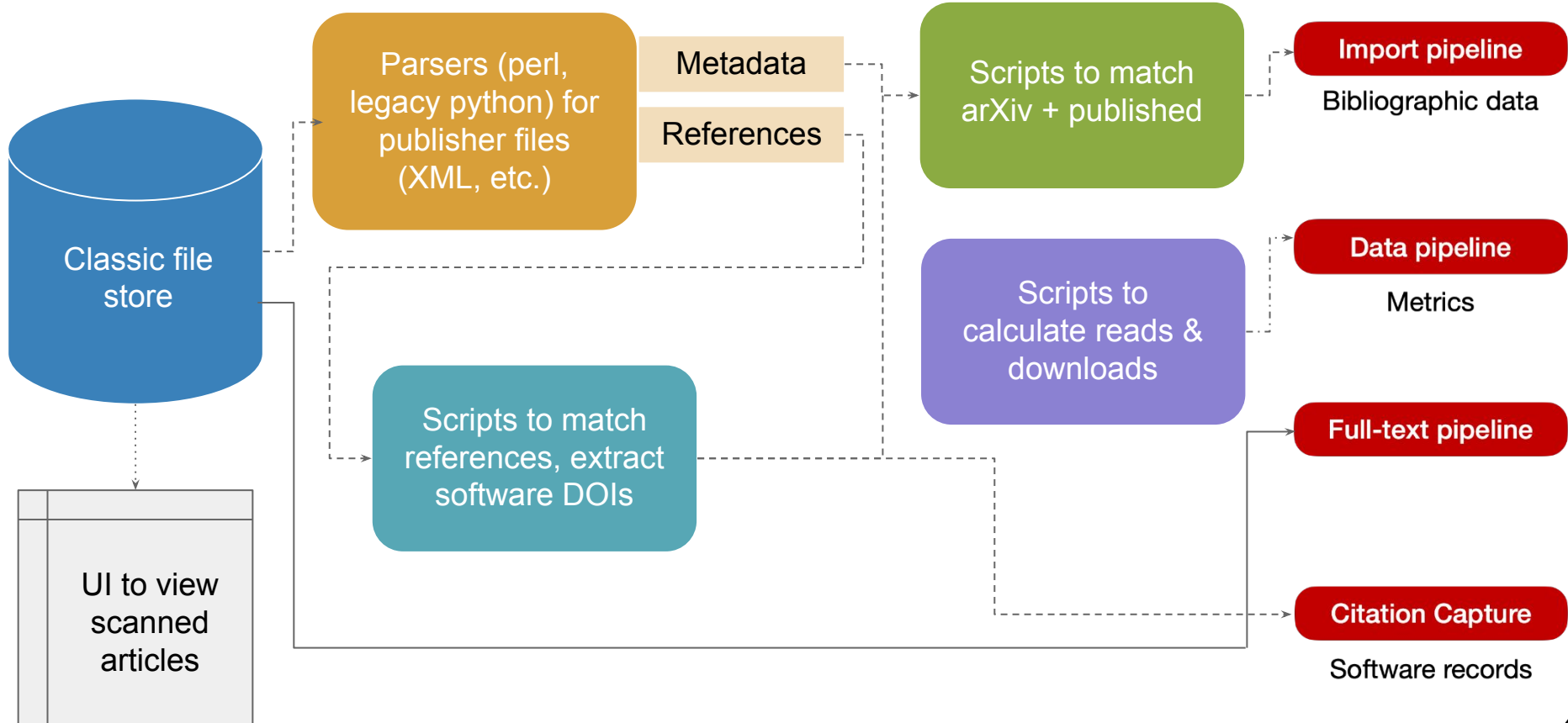
Current pipelines



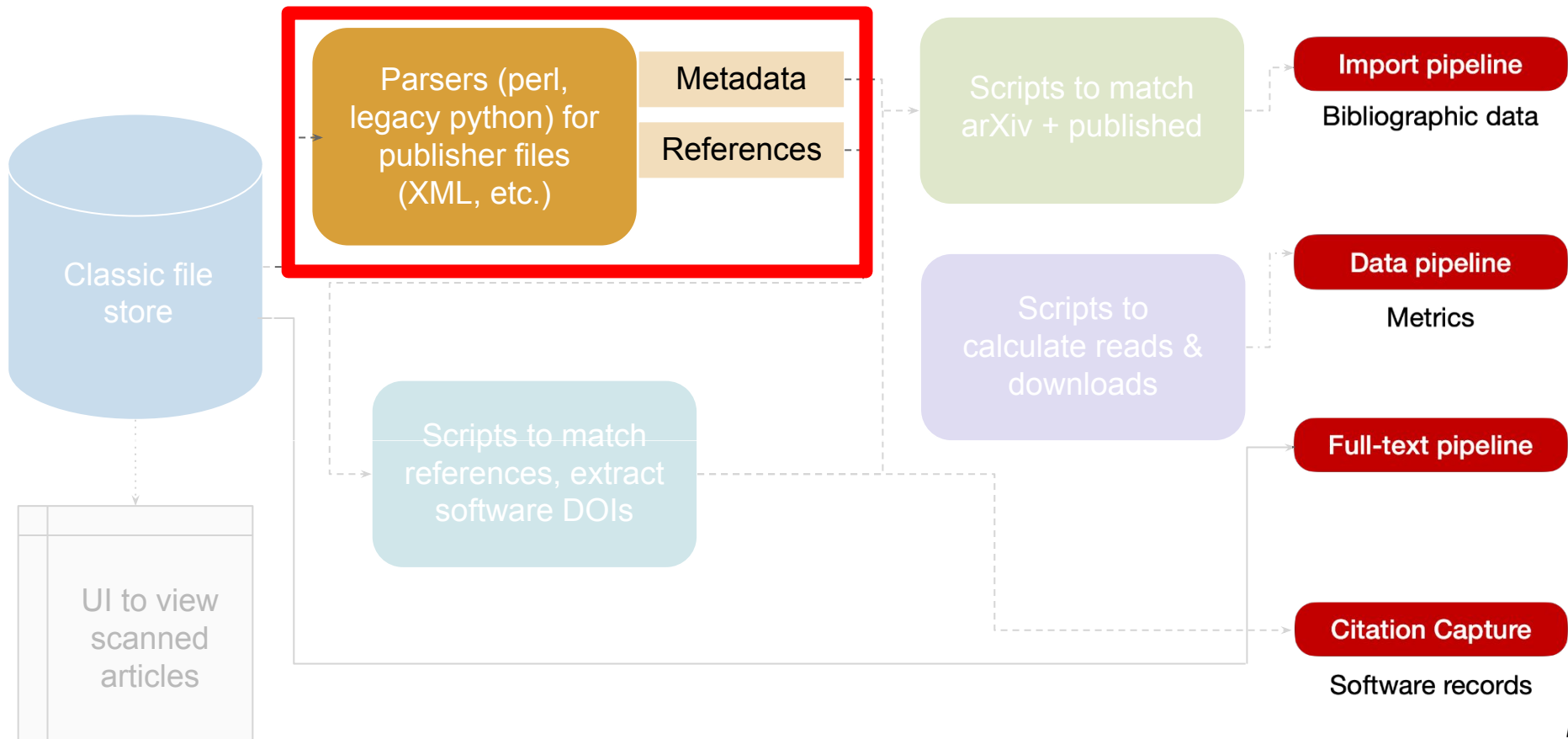
Current pipelines with legacy dependencies



Legacy dependencies



Parsers



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

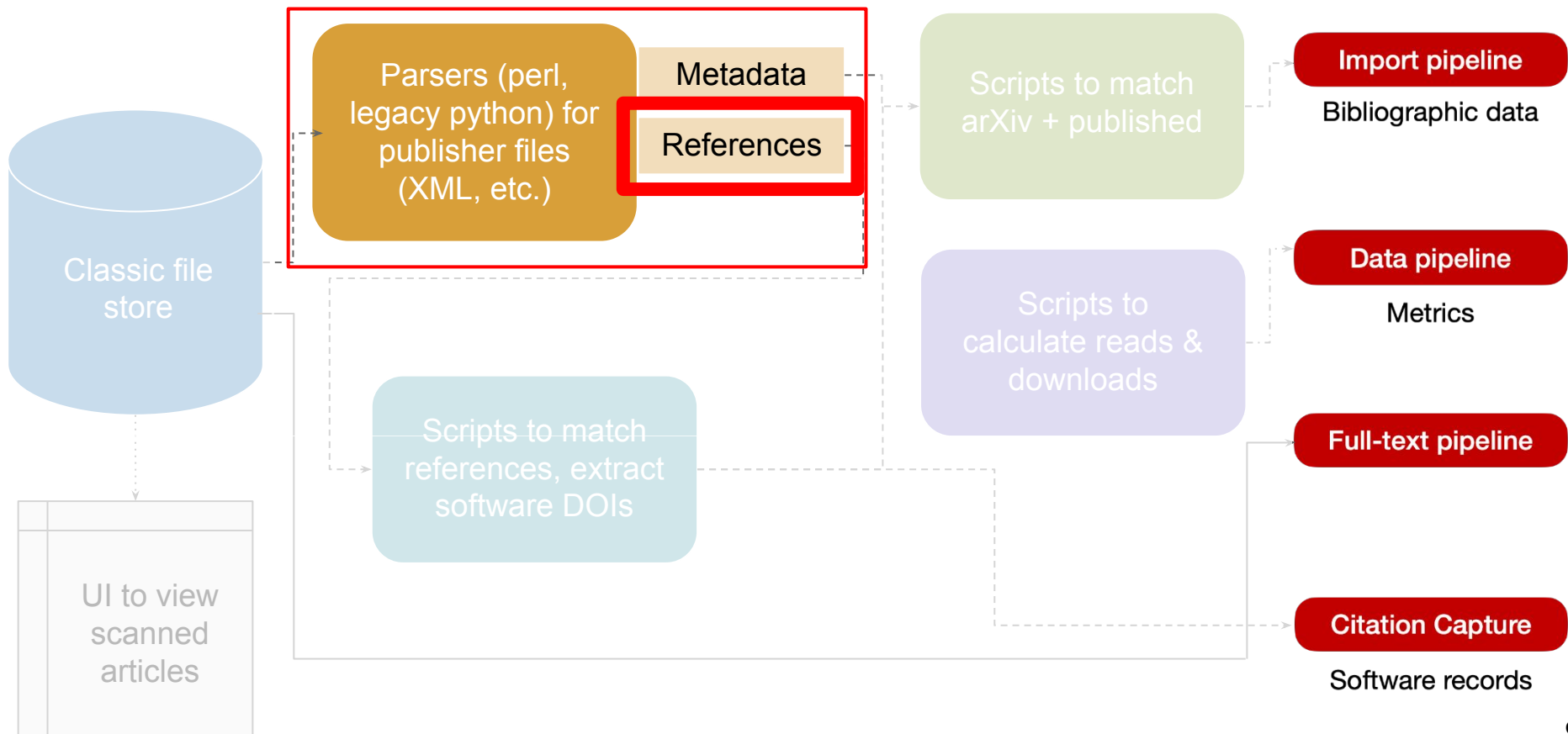
New parsers

- 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 $\sim\frac{1}{4}$ of total parsers, but cover $\sim\frac{2}{3}$ 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



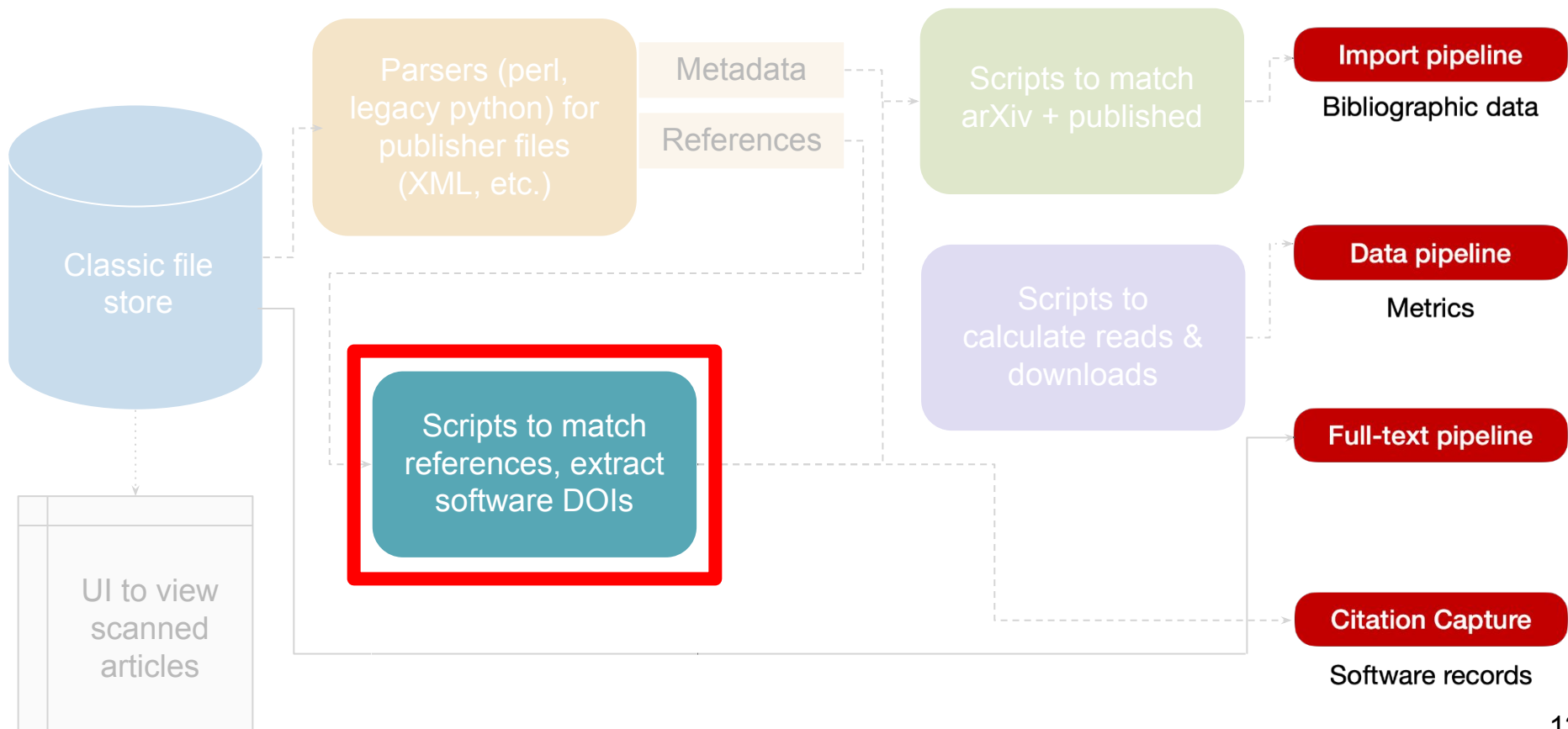
Reference Extraction: Context

- 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



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

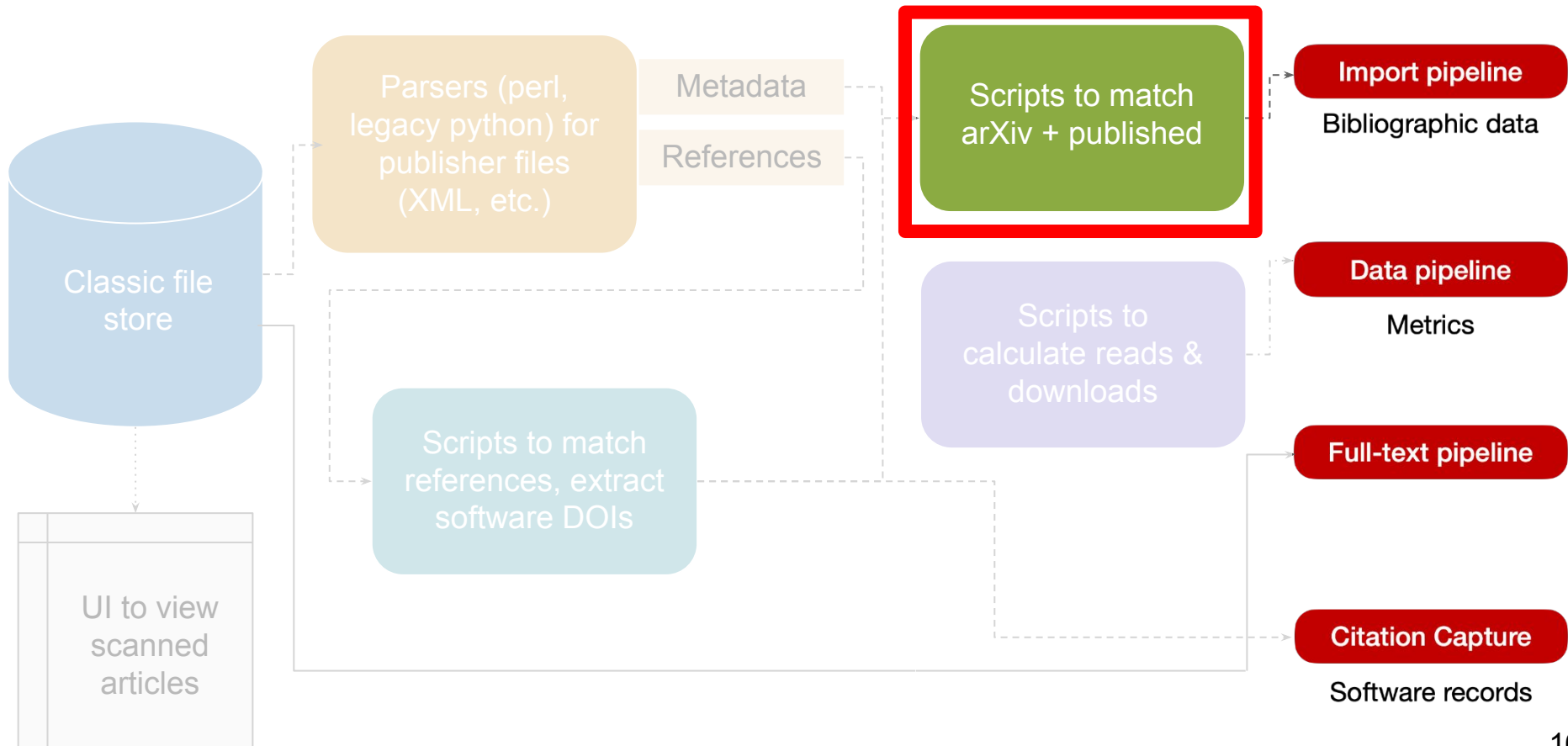
Reference Service for Content & Curation

- 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

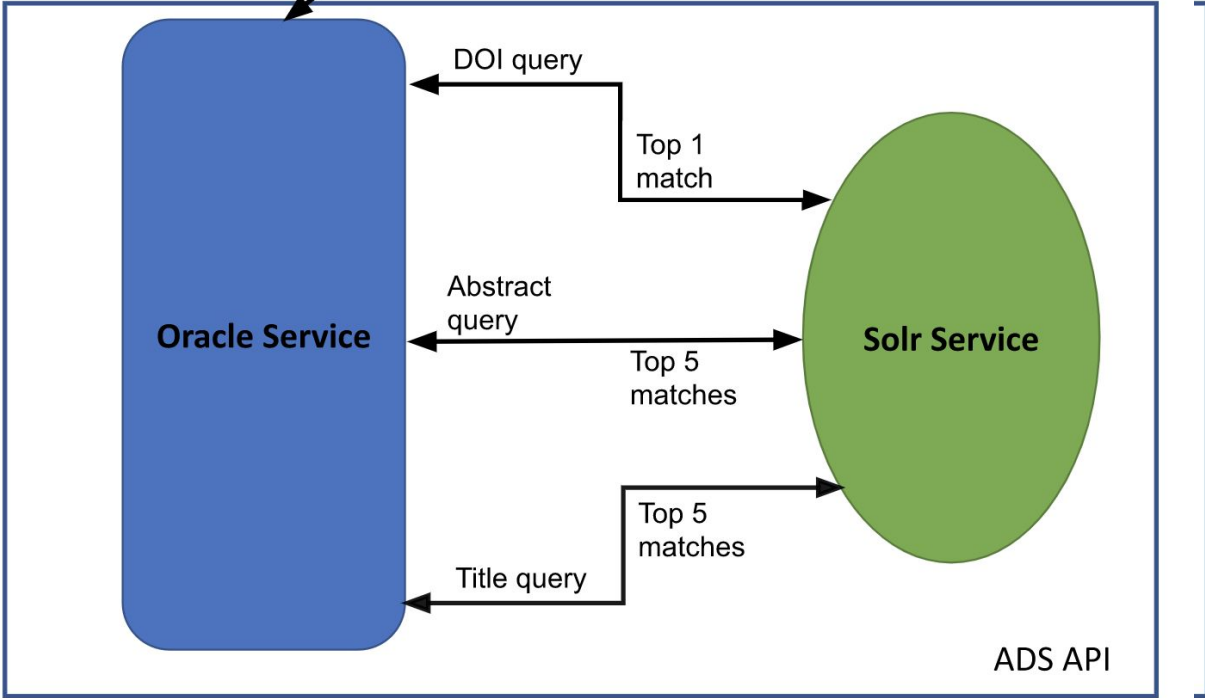
- Reference service is complete and its accuracy exceeds ADS Classic's resolver
- Future work:
 - Parsers for some smaller publishers' references (e.g. conferences, etc.)
 - Integrate into pipeline infrastructure

Docmatcher



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

Title query

Top 5 matches

Solr Service

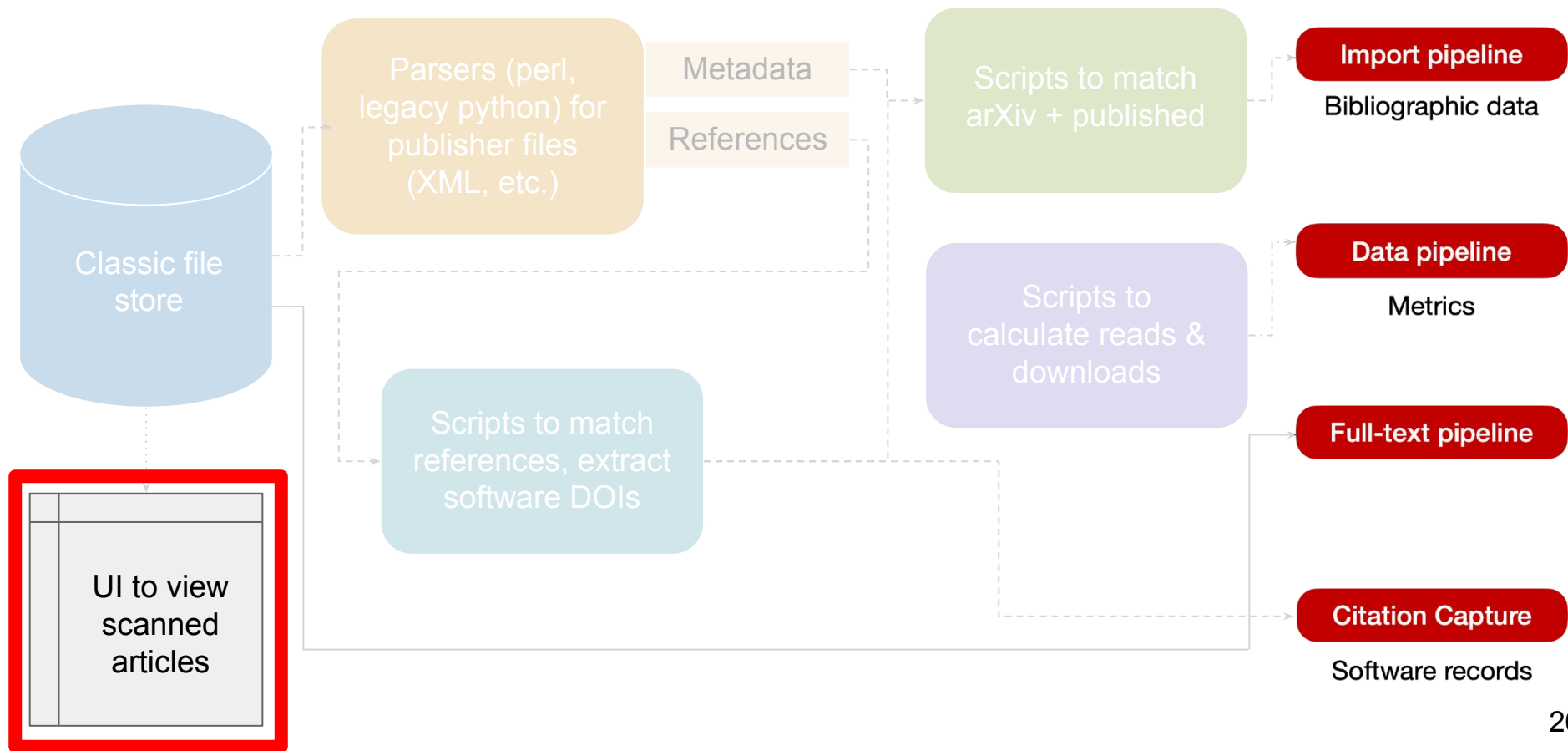
ADS API

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



ADS Digitization Efforts

QUICK FIELD: [Author](#) [First Author](#) [Abstract](#) [Year](#) [Fulltext](#) [All Search Terms](#)

[← Back to results](#)

VIEW

- Abstract**
- Citations (620)
- References
- Co-Reads
- Similar Papers
- Volume Content
- Graphics
- Metrics
- Export Citation

FEEDBACK

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

[Show affiliations](#)

[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](#)

Feedback/Corrections?

FULL TEXT SOURCES

- My Institution
- ADS

[Add paper to library](#)

PDF served from AWS

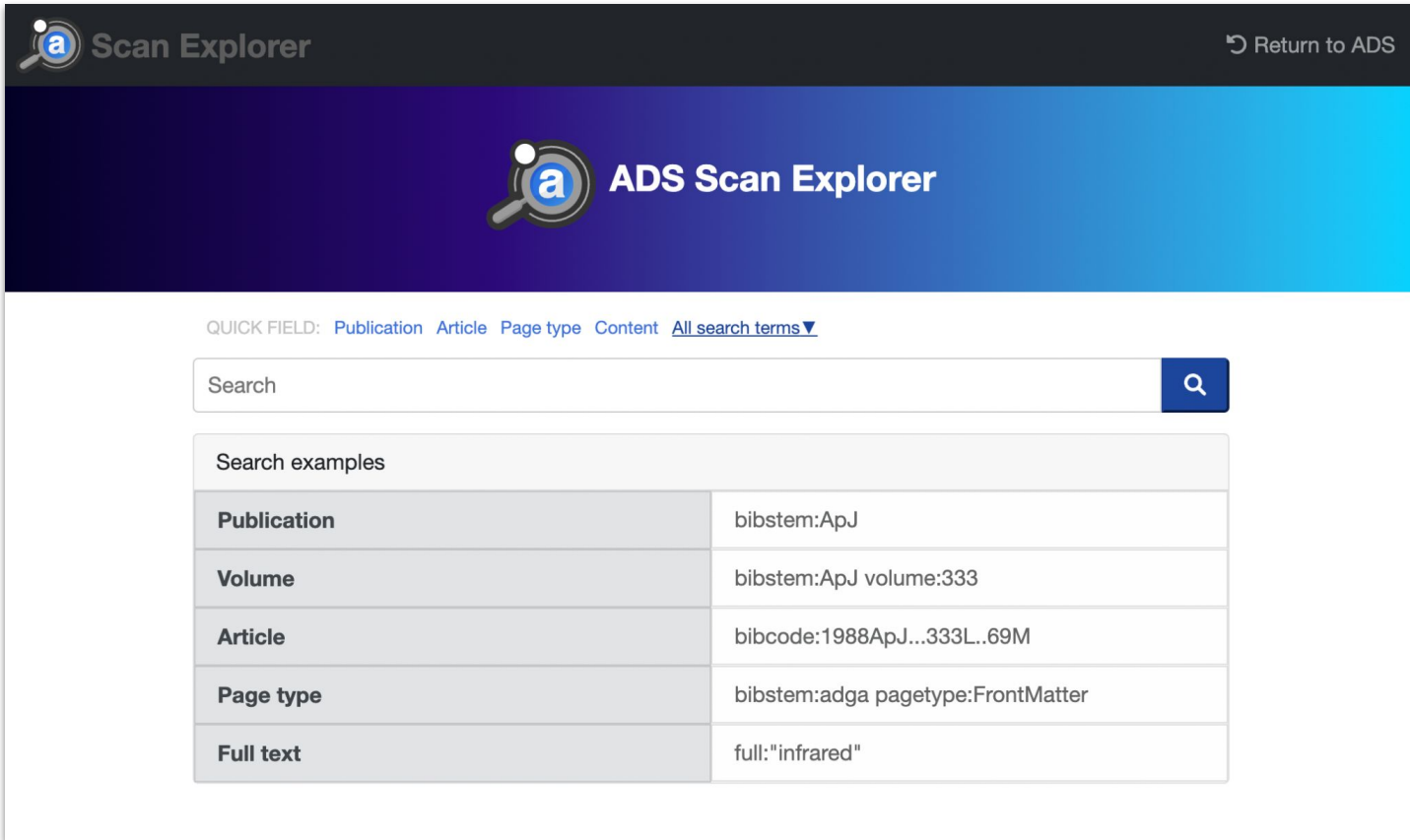
Scans served from ADS Classic

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

Return to ADS

QUICK FIELD: [Publication](#) [Article](#) [Page type](#) [Content](#) [All search terms](#)

bibstem:ApJ

Your search returned **47093** articles

Articles Collections 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 as rapidly as if it is during an eclipse. The essential features of this device are the instantaneous movements of the multi-mirror-plate 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 phenomena of interest which emit radiations of approximately constant wavelength; and the efficiency of the method depends very largely

On the Conditions which Aff...

1895ApJ.....1....1M

10 pages

Michelson, A. A.

PRELIMINARY TABLE OF SOLAR SPECTRUM
WAVELENGTHS. I.

By HENRY A. ROWLAND.

The following table of the lines of the solar spectrum has been in course of preparation at the Johns Hopkins University for many years in connection with an investigation of the spectra of the elements. The spectrum of every known element, except gallium (of which I have no spectrum), has been photographed in connection with the solar spectrum, and some of these plates have been measured.

The whole solar spectrum has now been measured except the extreme ends, and the wavelengths have been mostly reduced to my table of standards. Many of the lines, especially the stronger ones, have been identified with respect to the substance producing them, but this must be a labor of years. Hence I have determined to publish the work so far as I have now proceeded, expecting to add to it and correct it for a term of years, until I can publish a standard list of the lines of the solar spectrum with all the elements to which they belong.

The wavelength measurements have been made from my photographs of the solar spectrum, extending at present down to about wave-length 7000, and they will probably not be much changed in the future. The figures in the table refer to the wave-lengths in air at 20°C and 760^{mm} of mercury, as they are based upon the table of standards.

The intensities of the solar lines go from 1, a line just clearly visible on my map, up to 1000 for the H and K lines. Below 1 the lines in the order of faintness proceed from 0 to 1000, indicating lines more and more difficult to see.

The ordinary scale from 1 to 10 or from 1 to 6 is far too limited for the spectral lines, especially for the metallic spectra.

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.

By F. L. O. WATSON.

The special requirements which an astronomical spectroscopy has to fulfil in addition to the usual ones presented by the best laboratory instruments, are:

1. The greatest possible degree of comparison consistent with a given degree of resolving power, which is usually determined in advance by the character of the work which is to be done with the instrument.

2. The highest possible efficiency as regards loss of light by reflection and absorption in the passage of the beam through the spectroscopic slit. See, however, on spectrum

THE
ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY
AND ASTRONOMICAL PHYSICS

VOLUME 1 FEBRUARY 1895 NUMBER 2

ON A LENS FOR ADAPTING A VISUALLY CORRECTED REFRACTING TELESCOPE TO PHOTOGRAPHIC

ADS Scan Explorer - Article View

The screenshot displays the ADS Scan Explorer interface. At the top, there is a header with the 'Scan Explorer' logo and a 'Return to ADS' link. Below the header, a toolbar contains several icons for document manipulation. The main content area shows a scanned page from 'THE ASTROPHYSICAL JOURNAL', dated 'JANUARY 1895', with the article title 'ON THE CONDITIONS WHICH AFFECT THE SPECTRO-PHOTOGRAPHY OF THE SUN.' by Albert A. Michelson. The page text discusses solar spectro-photography and its applications.

Annotations on the interface include:

- download scans**: A red arrow points to the download icon in the top toolbar.
- download OCR**: A purple arrow points to the OCR download icon in the top toolbar.
- image manipulation**: An orange oval highlights a set of icons (refresh, undo, settings, crop, zoom, rotate, close) in the top toolbar.
- page navigation**: A blue arrow points to the navigation icons (back, forward, search) at the bottom of the page.
- image zooming**: A purple oval highlights the zoom in (+), zoom out (-), and zoom reset (↻) icons at the bottom of the page.

© American Astronomical Society 1895ApJ.....1.....1M

ADS Scan Explorer - Collections View

Scan Explorer Return to ADS

QUICK FIELD: [Publication](#) [Article](#) [Page type](#) [Content](#) [All search terms](#)

Search:

Your search returned 474 collections

Articles **Collections** Pages ↓ bibcode

THE ASTROPHYSICAL JOURNAL
AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

VOLUME: JANUARY 1895 NUMBER 1

ON THE CONDITIONS WHICH AFFECT THE SPECTRO-PHOTOGRAPHY OF THE SUN.
By ALBERT A. MICHAELSON.

This recent development in solar spectro-photography are in great measure due to the device originally suggested by Janssen and perfected by Hale and Dawkins, 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 the device are the substitutional movement of the photographic plate 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 engraved by light of this particular wave-length.

Evidently the process is not limited to the photography of the prominences, but extends to all other peculiarities of structure which reach radiations of approximately constant wave-length; 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 F. HALE, JAMES E. KEELER, HENRY CREW, E. B. FROST, F. L. O. WARDWORTH

ASSOCIATE EDITORS: H. A. CORNU, M. S. BUNDA, WILLIAM MOODS, P. FACHINCHI, H. C. HOYLE


AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

EDITORS: GEORGE F. HALE, JAMES E. KEELER, HENRY CREW, E. B. FROST, F. L. O. WARDWORTH

ASSOCIATE EDITORS: H. A. CORNU, M. S. BUNDA, WILLIAM MOODS, P. FACHINCHI, H. C. HOYLE

<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](#)

Your search returned 433812 pages

Articles Collections **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 ANASTAS MICHELSON.

This recent development in solar spectro-photography is a great measure due to the device originally suggested by Janssen and Perrotin, and perfected by Jahn and DeWahlund, by means of which a photograph of the Sun's prominences may be obtained as very faint as readily as it is during an eclipse. The essential features of the device are the simultaneous movements of the collimator-slits 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 wave-length.

Obviously the process is not limited to the photography of the prominences, but extends to all other peculiarities of structure which reach relations of approximately constant wave-length, and the efficiency of the method depends very largely

importance can hardly be overestimated, consists in the size of the height 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 statement, it is of course necessary that the actual intensity of the light should be sufficient to affect the photographic plate, and also that the routine 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: a = width of collimator slit,
 β = length of the image of a prominence,
 α and α' = their optical images,
 $\theta = \frac{a}{\beta}$ = aperture of objective,
 $\theta' = \frac{\alpha}{\beta'}$ = aperture of collimator,
 $\theta'' = \frac{\alpha'}{\beta''}$ = aperture of photographic lens,
 ω = angle of incidence,
 ω' = angle of diffraction,
 $\omega'' = \omega$ = grazing angle.

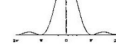
ω = angle of spectrum.

If the whole spectrum of the radiometer is filled with light by the cone of rays from the objective, then $\theta = \theta'$. The magnification of the spectroscopic image is a further enlargement due to diffraction by the aperture of the photographic lens (to which should be added the effects of imperfections in the lens and grating or prism, of vibrations and of irregularity of temperature) and, finally, the enlargement by dispersion.

Page 1
Page 1 in ApJ..0001

Page 2
Page 2 in ApJ..0001

This is shown graphically in the figure, in which the ordinates represent intensity and the abscissae the angle θ . The actual width of the line is that for which $\theta = \theta'$; that is, $\omega = \omega'$; but the apparent width will vary with the sensitiveness of the eye, or the photographic plate.



In the case of a source of finite width, the effect would be found by integrating this expression between values corresponding to the width of the source. It may also be generalized: the image is broadened by a quantity $\frac{1}{2}\theta''$, in which (in view of the uncertainty just noted, and considering also that the various diffracted elements previously noted, make the actual value still

<https://dev.adsabs.harvard.edu/scan/search?q=bibstem%3AApJ+%8apage=1&limit=10&t=page>

27

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