Brookhaven National Laboratory

U.S. DEPARTMENT of ENERGY

NSR Modernization

David Brown, Ben Shu, Ramon Arcilla, Boris Pritychenko
National Nuclear Data Center,
Brookhaven National Laboratory, Upton NY, USA

NRDC, Madrid Spain,

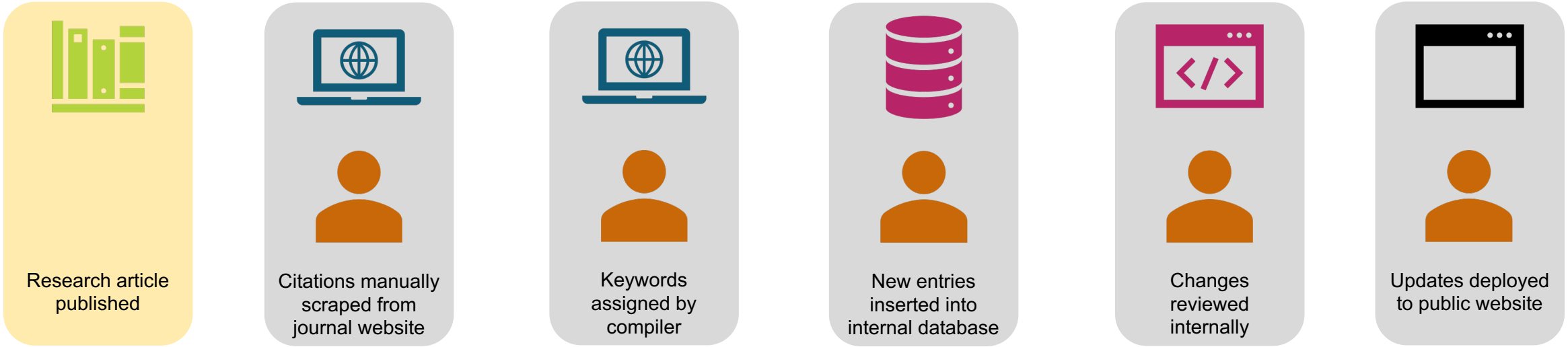
19 June 2025

@BrookhavenLab

NSR Workflow

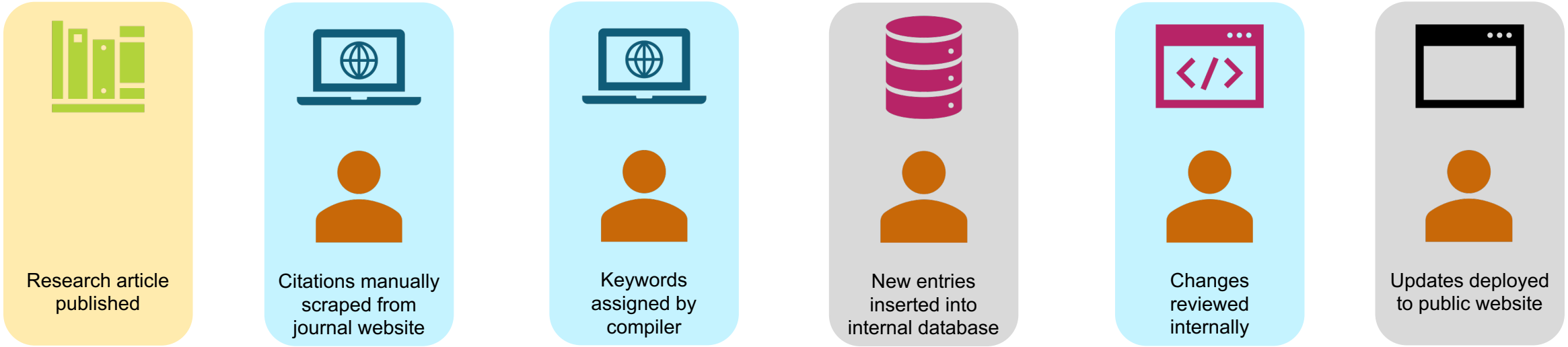
Ben Shu

Previous Workflow



- All steps handled by human beings
  - But we have no more contractors as of April 2025

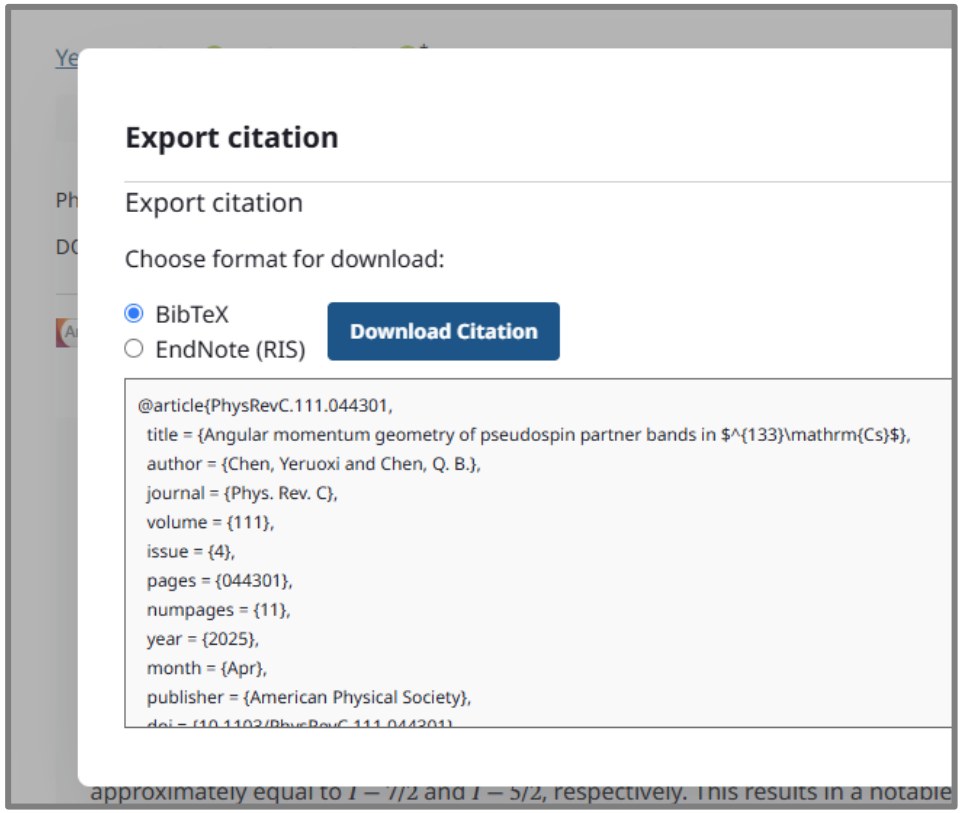
Previous Workflow



- Blue steps are rate determining; are the target for improvements
  - Legacy process must be carefully documented

Web scraping

- Article citations formerly obtained by people searching in websites
  - **Ex. Physical Review C (PRC)**
    - Click on one article, download BibTeX
    - Rinse and repeat for an entire issue
    - Vol 11, Issue 3 – 85 articles added
- This can be automated by web scraping tools

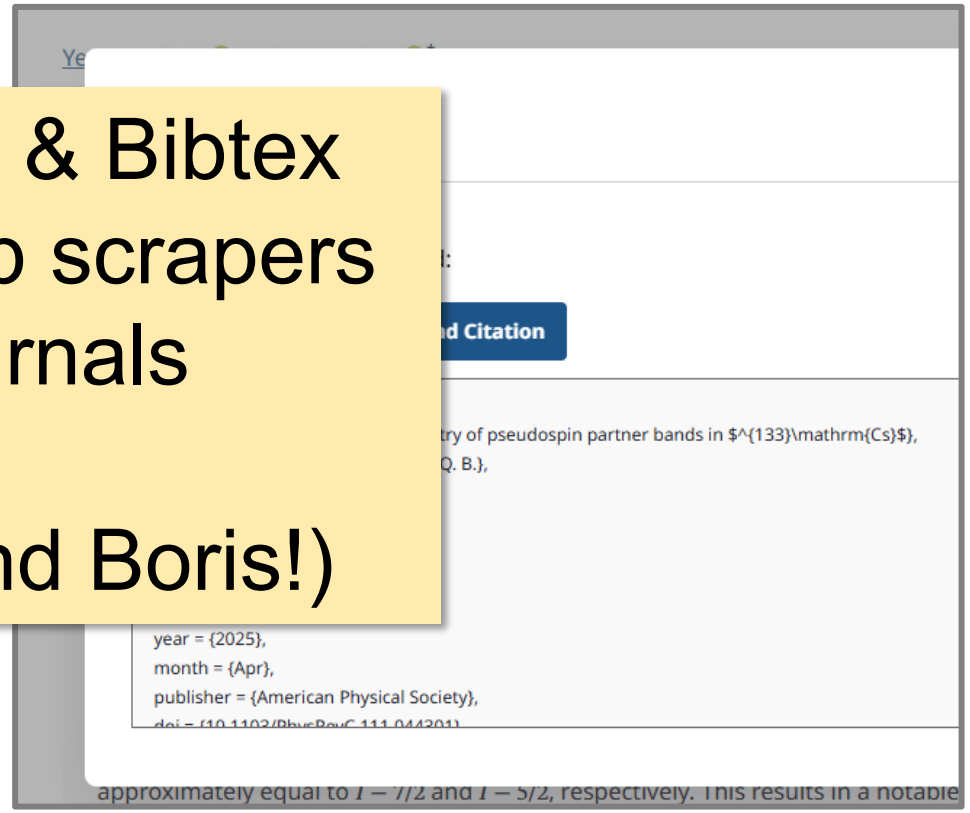


Web scraping

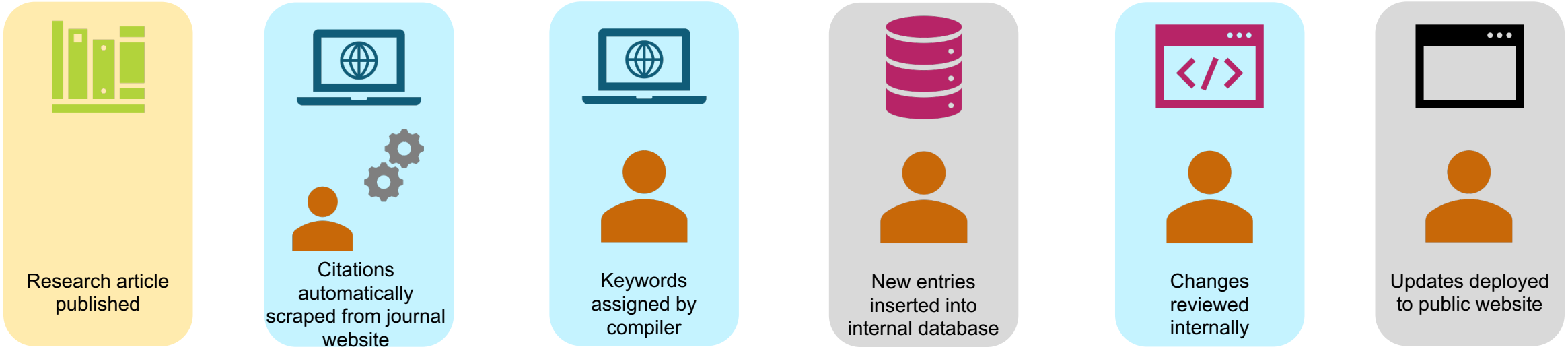
- Article citations formerly obtained by people sea
  - **Ex. Physical**
    - Click on one
    - Rinse and r
    - Vol 11, Issu
- This can be automated by web scraping tools

We now have RIS & Bibtex translators and web scrapers for several journals

(thank you Ben and Boris!)



Slightly Improved Workflow



- Simple scrapers have sped up ingest from 1 week to 1 hour
- We've sped up article ingest, but we can go faster!

Development Workflow

- **scrape-from-prc**
  - Automatically scrapes and processes PRC articles
  - Connects to PRC website to obtain articles
  - Downloads BibTeX, converts to NSR "exchange"

```
@article{PhysRevC.111.044301,
  title = {Angular momentum geometry of pseudospin partner bands in $^{133}\mathrm{Cs}$},
  author = {Chen, Yeruoxi and Chen, Q. B.},
  journal = {Phys. Rev. C},
  volume = {111},
  issue = {4},
  pages = {044301},
  numpages = {11},
  year = {2025},
  month = {Apr},
  publisher = {American Physical Society},
  doi = {10.1103/PhysRevC.111.044301},
  url = {https://link.aps.org/doi/10.1103/PhysRevC.111.044301}
}
```

```
<KEYNO    >2025CHP
<HISTORY  >C20250409
<CODEN    >JOUR PRVCA 111 044301
<REFRENCE>Phys.Rev. C 111, 044301 (2025)
<AUTHORS  >Y.Chen, Q.B.Chen
<TITLE    >Angular momentum geometry of pseudospin partner bands in {+133}Cs
<DOI      >10.1103/PhysRevC.111.044301
```

Development Workflow (contd.)

- **update-prc-errata**
  - Checks for updates to existing NSR entries
  - Preserves key number, title, keywords, selectors, etc.

```
<KEYNO    >2025WAP
<HISTORY >C20250409
<CODEN    >JOUR PRVCA 111 049901
<REFRENCE>Phys. Rev. C 81, 037302 (2010); Erratum Phys.Rev. C 111, 049901 (2025)
<AUTHORS >E.H.Wang, S.H.Liu, J.H.Hamilton, A.V.Ramayya, Y.X.Luo, J.O.Rasmussen,
J.K.Hwang, A.V.Daniel, G.M.Ter-Akopian, S.J.Zhu, W.C.Ma
<TITLE    >Erratum: Nuclear structure of the odd-odd N=85
140}Cs [Phys. Rev. C 81, 037302 (2010)]
<DOI      >10.1103/PhysRevC.111.049901
```

Erratum scraped from PRC

Updated NSR entry

```
<KEYNO    >2010LI10
<HISTORY >A20100426 M20250409
<CODEN    >JOUR PRVCA 81 037302
<REFRENCE>Phys. Rev. C 81, 037302 (2010); Erratum Phys.Rev. C 111, 049901 (2025)
<AUTHORS >E.H.Wang, S.H.Liu, J.H.Hamilton, A.V.Ramayya, Y.X.Luo, J.O.Rasmussen,
J.K.Hwang, A.V.Daniel, G.M.Ter-Akopian, S.J.Zhu, W.C.Ma
<TITLE    >Nuclear structure of the odd-odd N=85 neutron-rich nucleus {+140}Cs
<DOI      >10.1103/PhysRevC.111.049901
<KEYWORDS>RADIOACTIVITY {+252}Cf(SF); measured E|g,I|g,|g|g-coin, |g|g(|q) using
```

Development Workflow (contd.)

- **save-sql-backup**
  - Saves a copy of NSR using **mariadb-dump**
  - Runs *before* any database changes
  - Can be used as a restore point, just in case

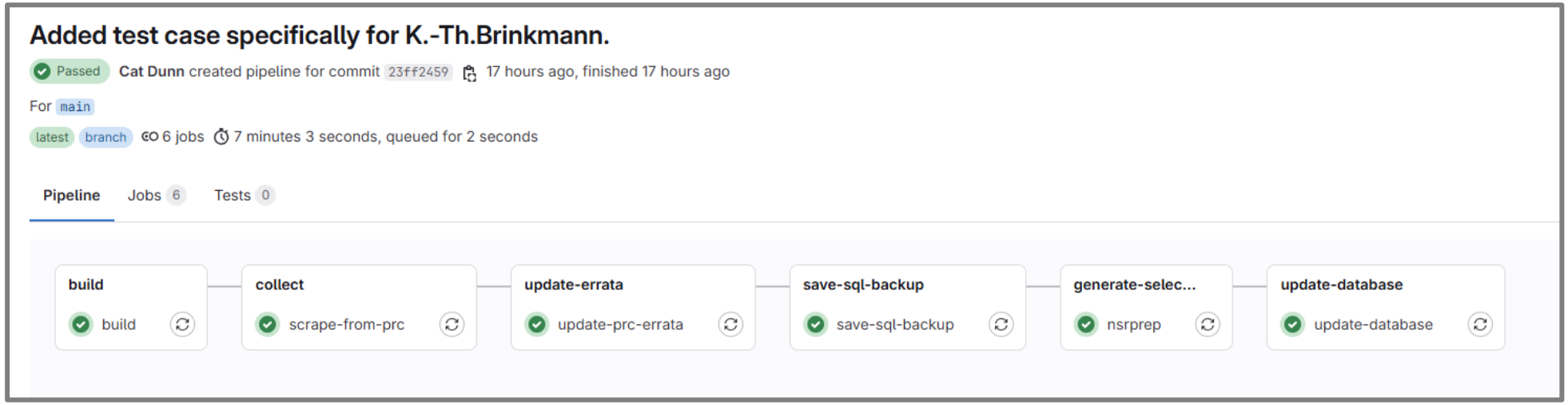
Development Workflow (contd.)

- **nsrprep**
  - Based on legacy web app used by compilers
    - Performs format/validation checks
    - Looks for unknown author names
  - Generates selectors used for NSR website searching
    - Only generates these for **keyworded** entries
    - If an entry is not keyworded, it cannot be found with:
      - **Quick Search:** Nuclide, Reaction
      - **Indexed Search:** Nuclide, Subject, Reaction, Z, A

Development Workflow (contd.)

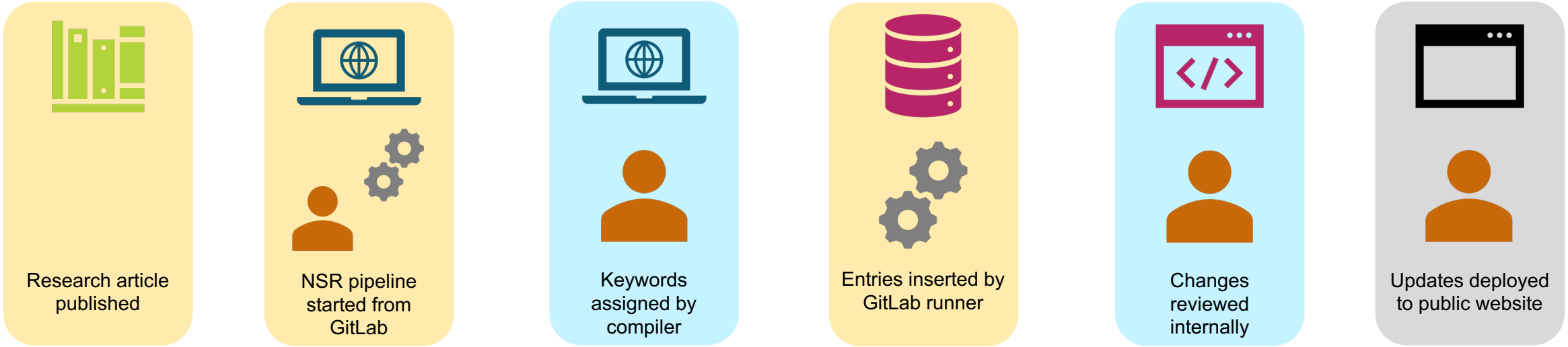
- **update-database**
  - Makes changes to **nsr_automated**
    - Internal only – not deployed to NSR website
  - Inserts new entries for standard PRC articles
  - Updates existing entries for PRC errata
- **Issue:** Currently no way to fix errors before insertion
  - i.e. format errors, PRC typos
  - Fixes require direct SQL changes or restarting pipeline

Development Workflow (contd.)



- Pipeline functional as of April 3rd, 2025
  - No setup/installation - only GitLab access required
  - Click a few buttons, go do something else

A Much Faster Workflow



- Full pipeline takes minutes to run
- We are not taking advantage of the full power of gitlab

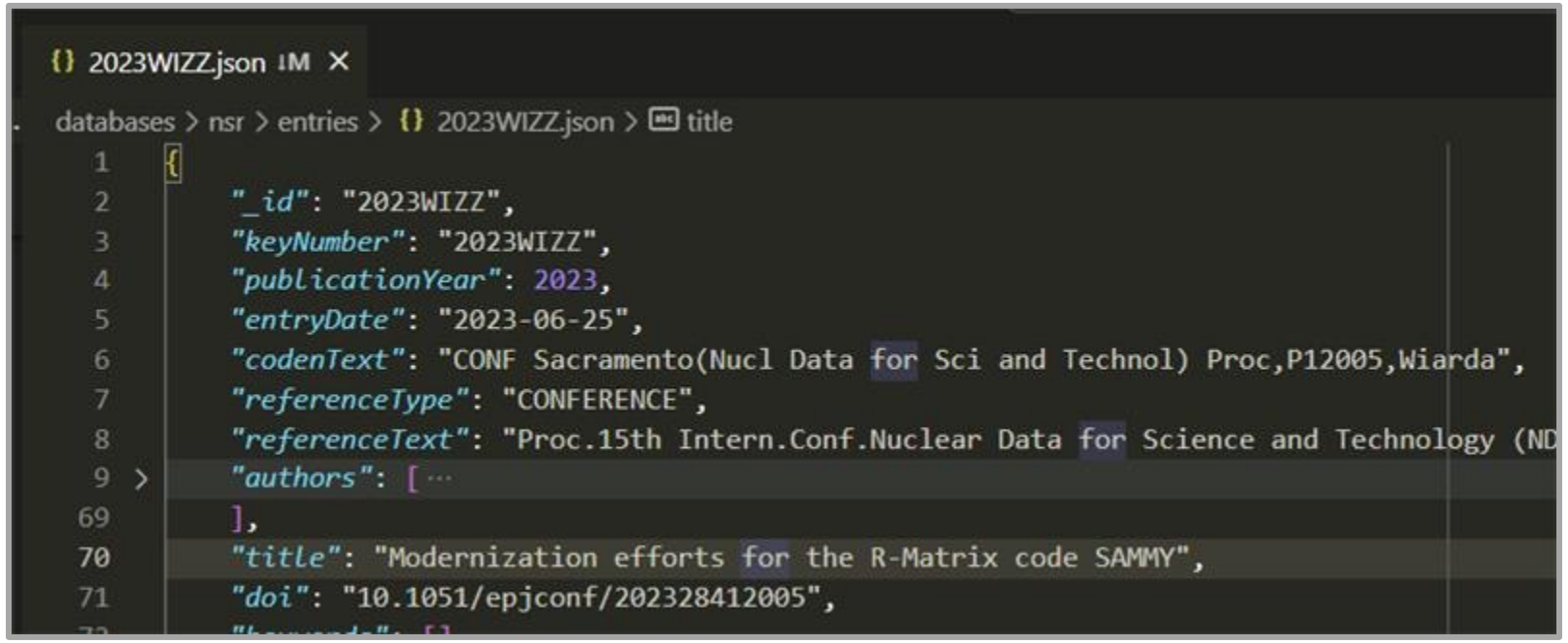
NSR Exchange format can be modernized

- The current NSR "exchange" format...

```
<KEYNO   > 2023WIZZ          Unique ID
<HISTORY > A20230605         Entry/modification dates
<CODEN   > CONF Sacramento...  Source type and information
<REFRENCE> Proc.15th.Intern...  Citation information
<AUTHORS > D.Wiarda, ...     List of authors
<TITLE   > Modernization...  Formatted title
<KEYWORDS> [N/A]             Sentences about content
<SELECTRS> [N/A]             Structured search keys
```

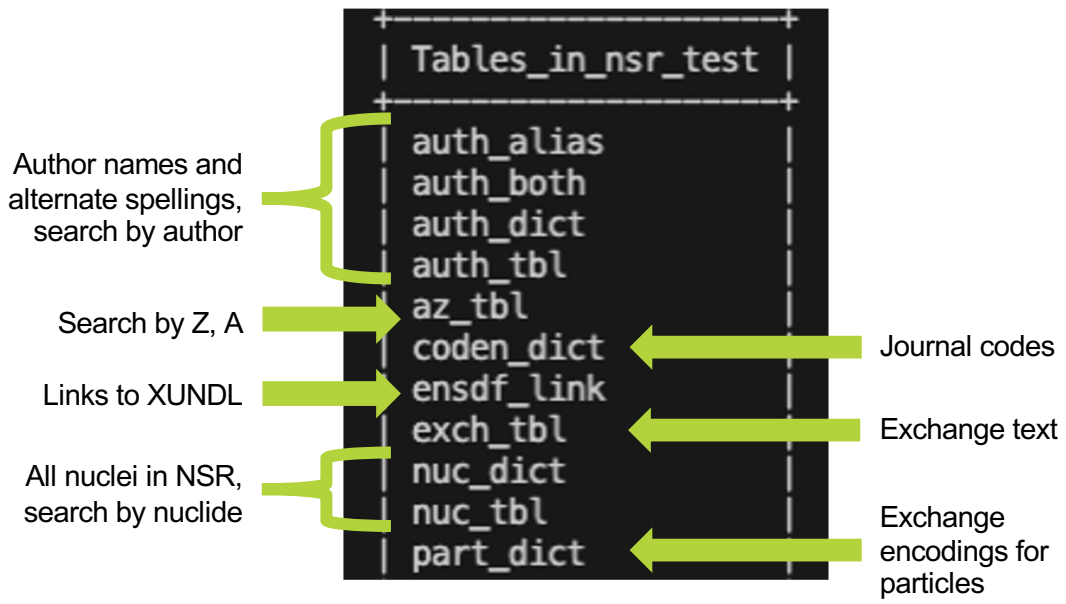
NSR Exchange format can be modernized

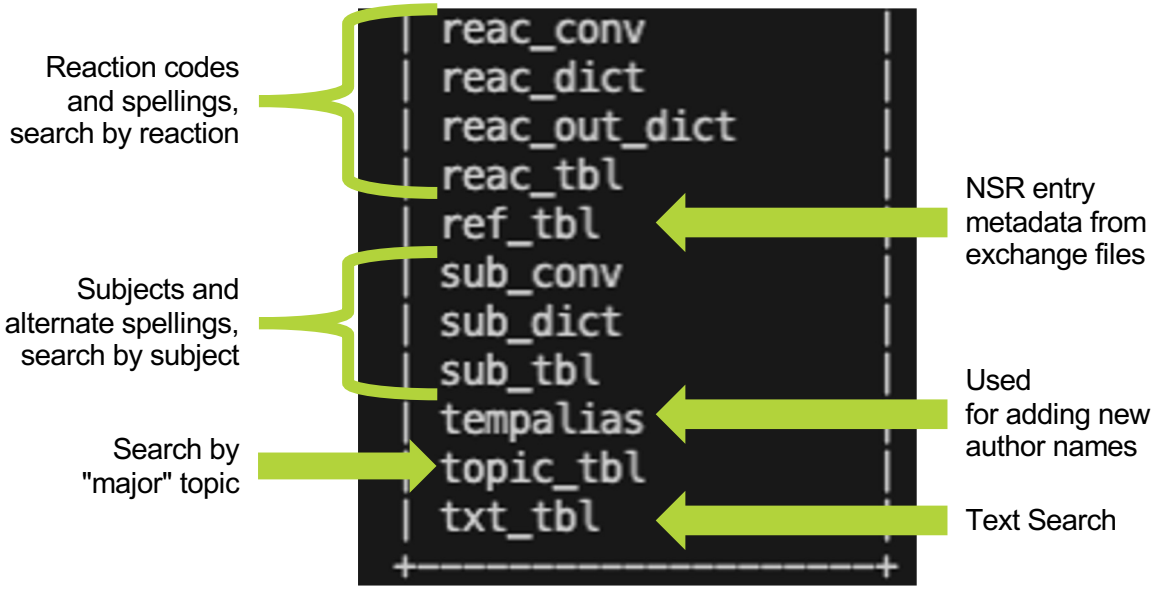
- ...can be translated into JSON...



NSR Exchange format can be modernized

- ...but the exchange files don't tell the full story
- NSR's SQL database has supporting tables:



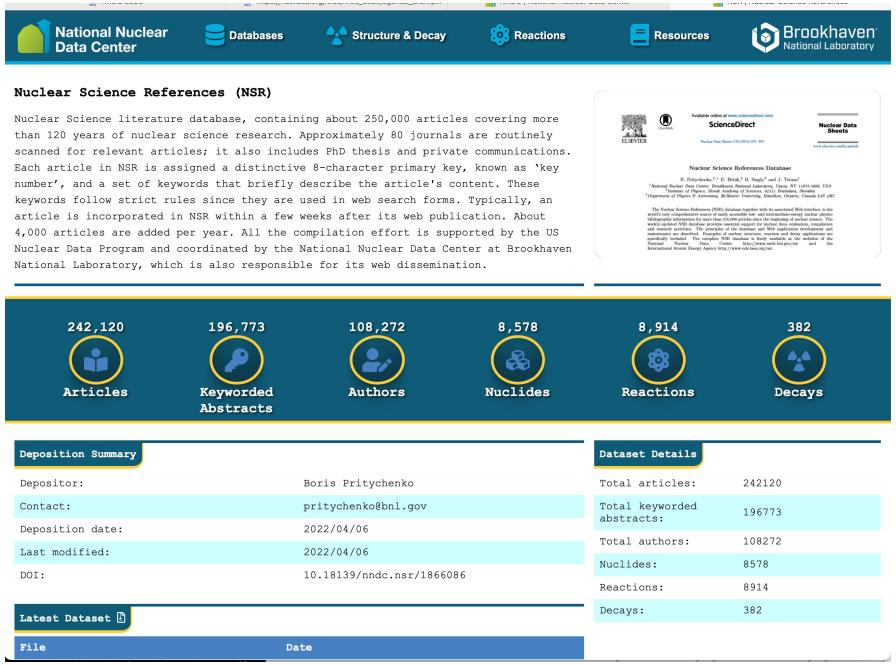


What does this get us?

A JSON format enables:

- A JSON-schema
- Format and correctness checkers
- An editor
- Better NSR archive option
- An AI/ML-friendly format!

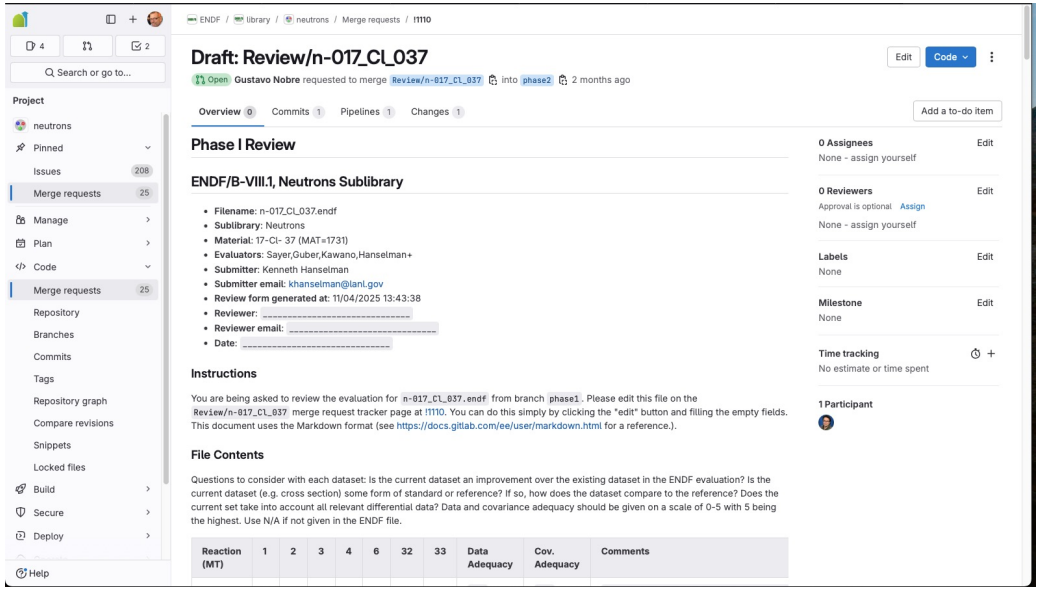
Future work for a summer student?



https://www.nndc.bnl.gov/nsrarchivals/

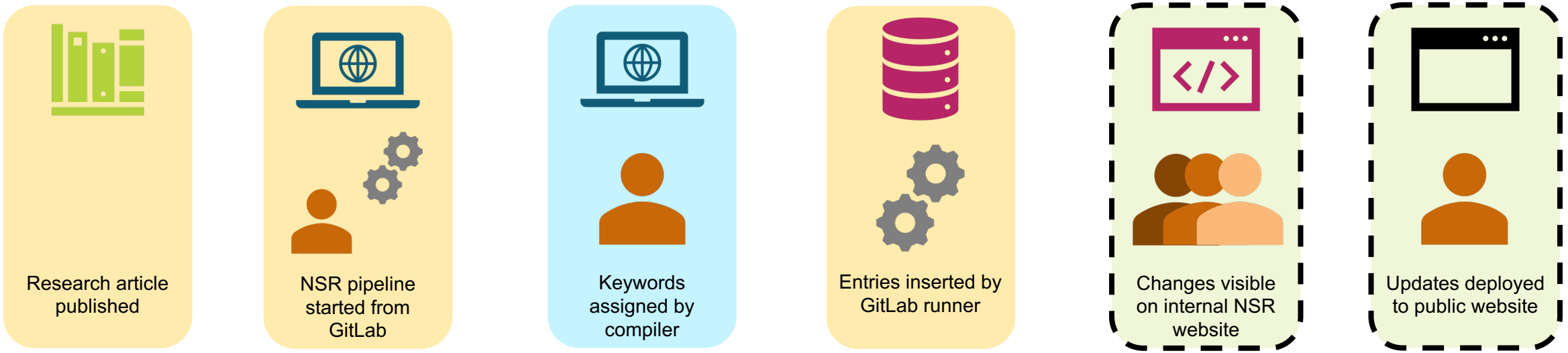
Even without these, with gitlab we have

- Version control
- Built in review step (branch merge process)
- We use this to great effect in ENDF & JEFF



Brookhaven National Laboratory

With Gitlab, we can ensure quality while getting speed



- With a QA process in hand, and an automated workflow, now we can target the hardest step

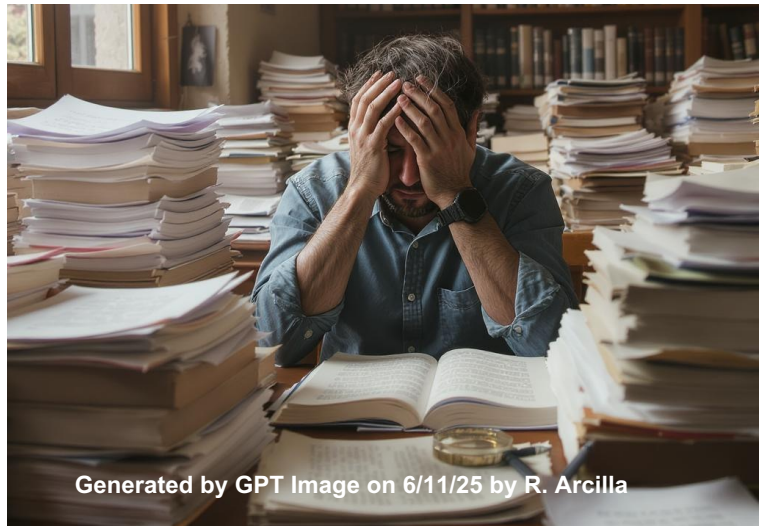
Keywording

Currently a compiler must produce keyworded abstract – time consuming and requires somewhat special skills

As of FY25, NNDC has no contractors

Options:

- Teach more people to do it (ENSDF/ENDF evaluators, students, NNDC staff?)
- Automate it?



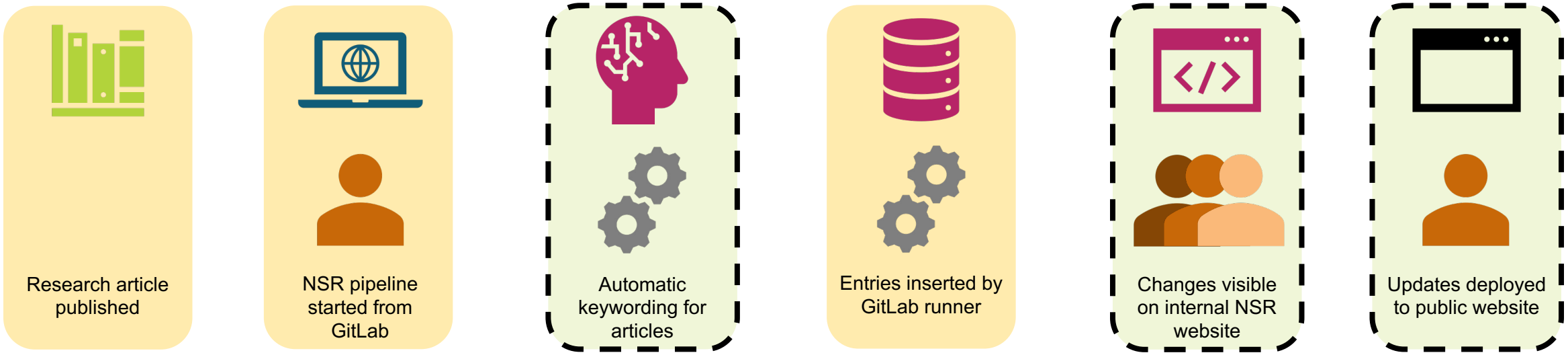
Nuclear Science References Coding Manual

D.F. Winchell
National Nuclear Data Center
Brookhaven National Laboratory
Upton, New York, USA

May 24, 2007

https://www.nndc.bnl.gov/nsr/docs/nsr-coding-manual.pdf

Future Workflow



- Need to fill in missing steps
  - Keyword assignment via chatNSR
  - Tools for human reviewers to make edits

chatNSR: An AI-Enhanced Nuclear Science References Knowledge Base

Ramon Arcilla

Motivation

Why automate keyword abstract generation?

Volume Challenge

- Growing nuclear science literature (~5,000+ annually) strains current manual keywording capacity.

Time Intensive

- Manual keywording requires ~3.5 hours (average) per article.

Limited Coverage

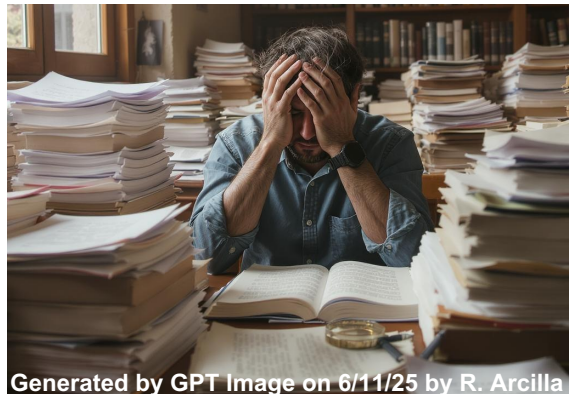
- Manual keywording limits timeliness and breadth of NSR knowledge base content.

Discovery Issues

- Inconsistent keywording limits paper discoverability.

Cost Inefficiency

- Manual keywording demands significant resources.



The chatNSR Project

Current Focus:

- Keywording, the most tedious and costly activity

Key AI Technologies in Use

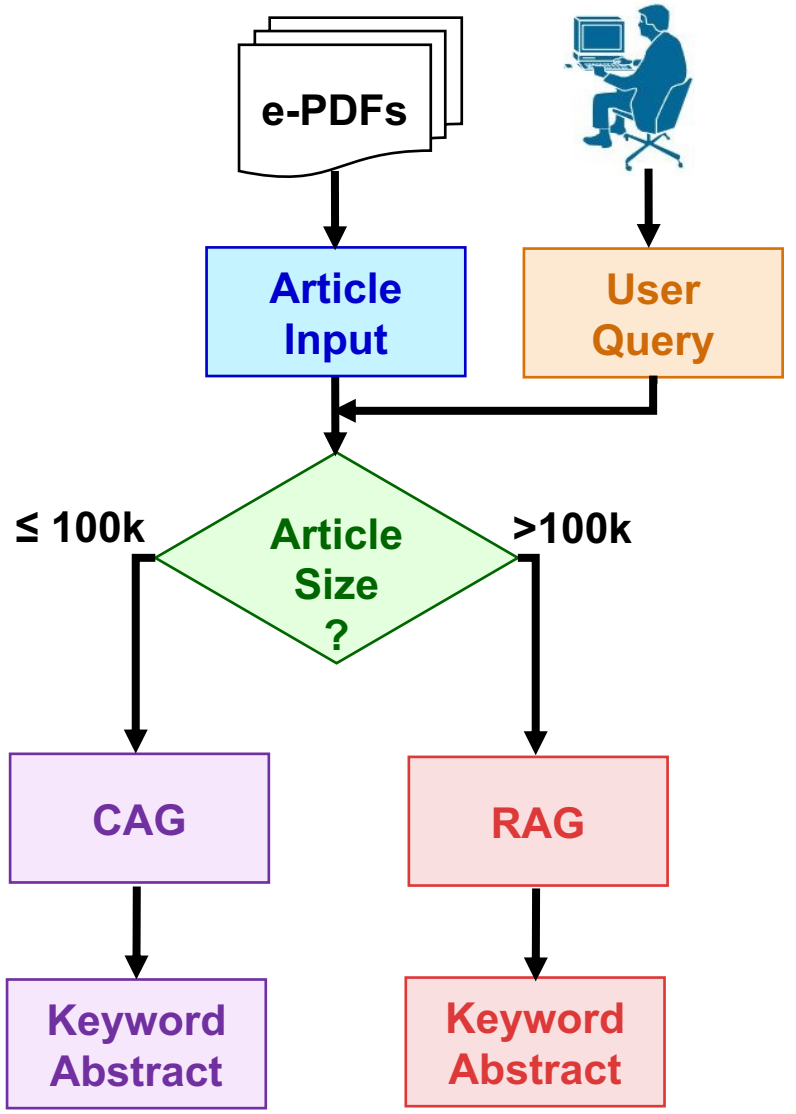
- Open-Source Large Language Models (LLMs)
  - Understand user queries/instructions and generate keyword abstract
- Cache-Augmented Generation (CAG)
  - Loads entire article into LLM's extended context window (Knowledge Cache) for efficiency, more accurate keywording
- Retrieval-Augmented Generation (RAG)
  - Finds contextually relevant parts of article in vector DB for LLM’s use in keywording
- System Prompt Engineering (SYSPROMPT)
  - User instructions to guide LLMs for consistent, accurate keywording


Generated by GPT Image on 6/10/25 by R. Arcilla

The chatNSR Project

Architectural Overview: Two-Path Automation Workflow

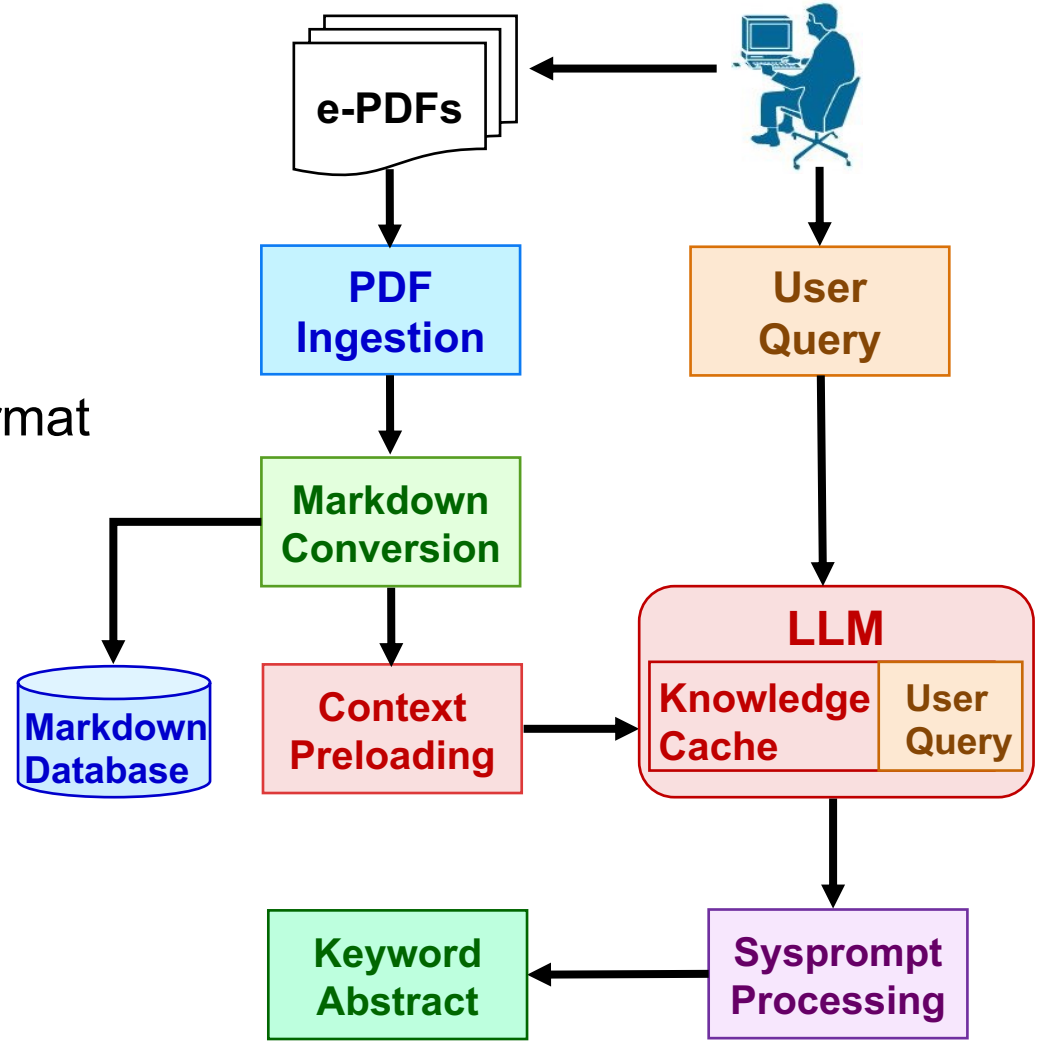
- Document Input
  - Article enters the automation pipeline
- Size Assessment
  - System determines article's token count
- Path Selection
  - CAG for smaller articles (≤100k tokens = 100 pages), RAG for larger ones (>100k tokens)
- Keyword Output
  - Keyword abstracts generated via optimal path



Cache-Augmented Generation (CAG)

Workflow

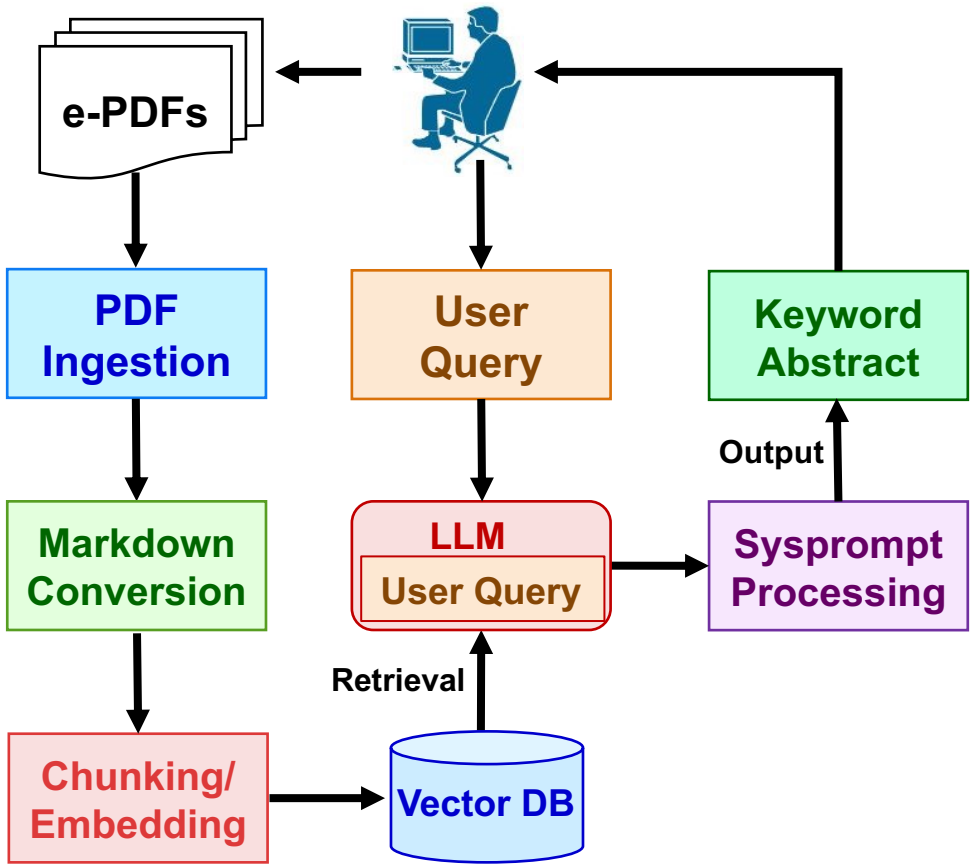
- PDF Ingestion
  - Article enters system and undergoes initial parsing
- Markdown Conversion
  - Content structure preserved in clean Markdown format
- Context Preloading
  - Full article loaded with Knowledge Cache creation
- System Prompt Processing
  - LLM guided by engineered instructions
- Response Output
  - Keyword abstracts delivered to user



Retrieval-Augmented Generation (RAG)

Workflow

- Document Processing
  - PDF ingestion and conversion to markdown format.
- Chunking & Embedding
  - Document segmentation and vector representation creation.
- Retrieval Mechanism
  - Query-based selection of relevant document sections.
- Generation with Prompting
  - LLM processes retrieved chunks with specialized instructions.
- Output Delivery
  - Keyword abstract returned to requesting user.



Advantages of the Two-Path Automation Approach

Universal Scalability

- System handles articles of any length.
- Processing adapts to content size automatically.

Consistent Quality

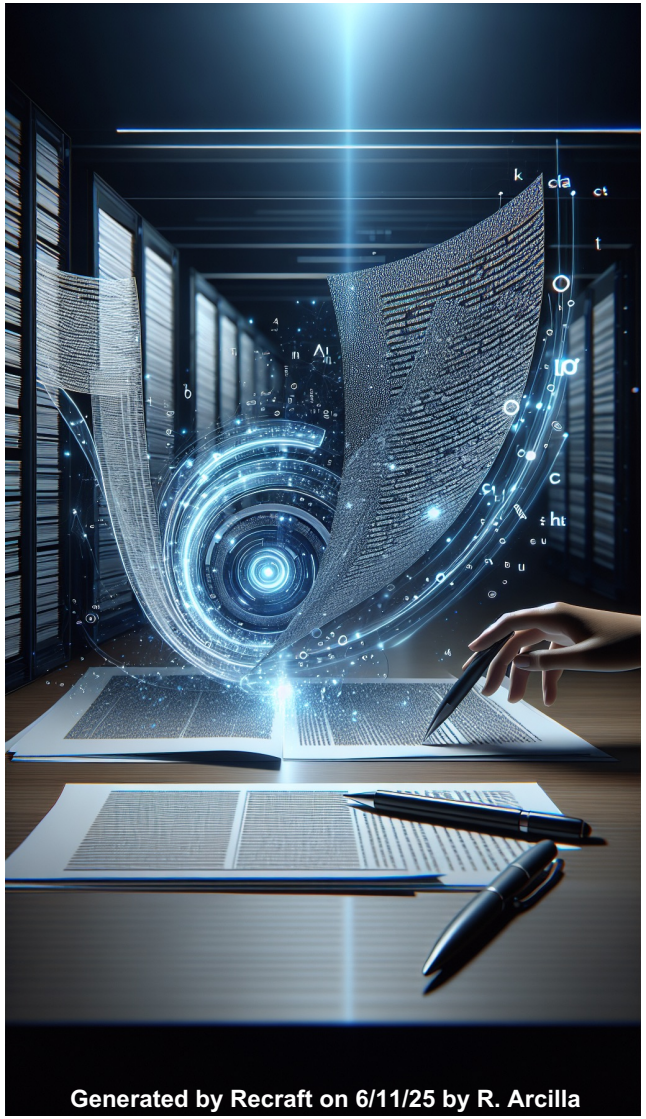
- Keyword abstract coherence maintained across both workflows.
- Keyword abstract remains relevant regardless of article size.

Resource Efficiency

- Computational resources allocated based on article needs.
- Simpler approach (CAG) used when possible.

Workflow Integration

- Unified API endpoint despite dual backend systems.
- Seamless integration with existing document processes.



Generated by Recraft on 6/11/25 by R. Arcilla

EXAMPLE: Keywording D.A. Brown's article "Zirconium Evaluations for ENDF/ B-VII.2 for the Fast Region"

Human-Generated

COMPILATION $^{90, 91, 92, 93, 94, 95, 96}$Zr(n, x), E=0=20 MeV; calculated, evaluated σ, σ(θ) including resonance region using EMPIRE-3.1 code. Compared with available data and evaluated databases.

chatNSR-Generated

COMPILATION $^{90\ 91\ 92\ 93\ 94\ 95\ 96}$Zr(n,x), E=0-20 MeV; calculated, evaluated σ, σ(θ) including resonance region using EMPIRE-3.1 code. Compared with available experimental data and evaluated databases (ENDF/B-VII.1, JENDL-4. 0).

Current Status

Testing

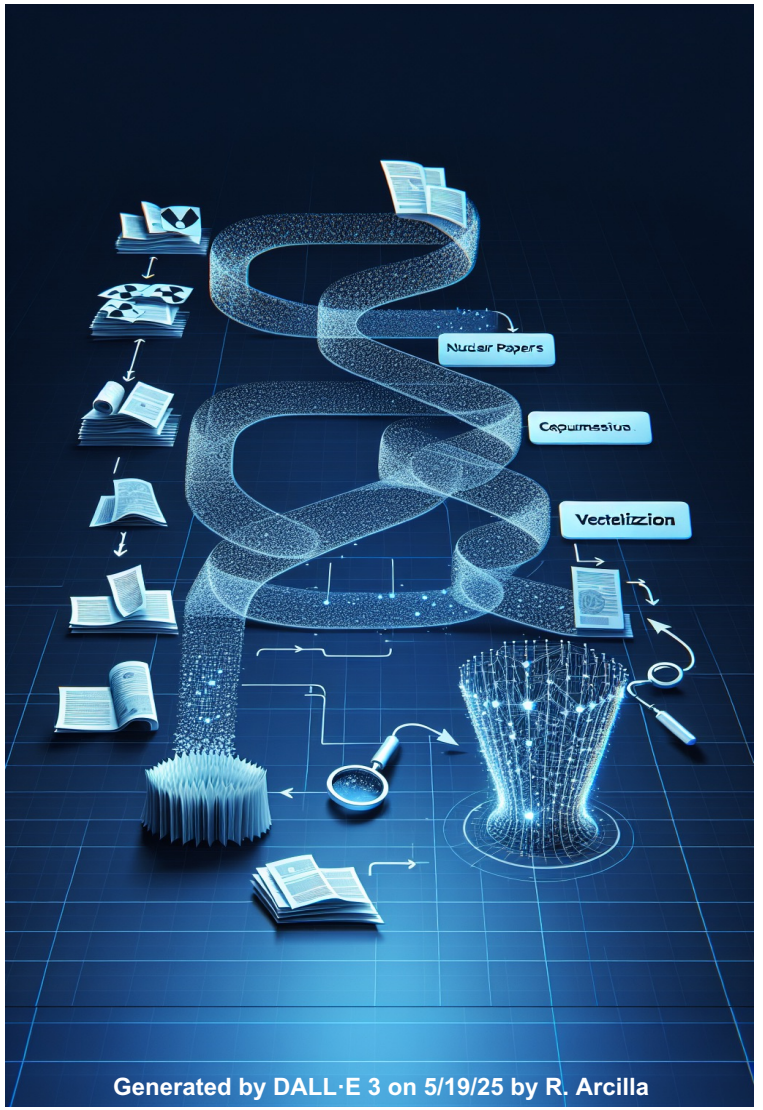
- Dozen open-source LLM's tested
- Dozens of articles (different topics) for LLM training
- Perplexity AI's ***r1-1776:70b*** LLM has highest accuracy

Accuracy

- 80-85% accuracy for articles ≤ 10 pages
- More work needed for articles > 10 pages

Methods and Tools in Use

- Continuous human feedback to LLM
- AI-assisted coding and innovative prompt engineering
- PhysBERT (LBNL) model for embedding and retrieval
- IBM Docling software for PDF-to-Markdown conversion
- Perplexity AI's ***r1-1776*** LLM for keyword generation



Generated by DALL·E 3 on 5/19/25 by R. Arcilla

Summary and Conclusion

Keywording Automation

- Continuous human feedback improves LLM's accuracy.
- Article (varied topics) high-volume ingestion and training enhances LLM's performance.
- CAG, RAG, System Prompts minimize/prevent LLM hallucination.
- 90% time reduction in keyword abstract generation
- PDF-to-Markdown conversion improves LLM's article understanding

On-premise Keywording

- Open-source LLMs and AI tools demonstrate cost-free, flexible, adaptable keywording
- More secure environment behind BNL FireWall
- Compliant with DOE rules and copyright laws

Generated by DALL·E 3 on 5/19/25 by R. Arcilla

Future Direction

LLM capability and size

- Move to more advanced LLMs with > 70 billion parameters for higher accuracy.

Article Ingestion

- Massive article ingestion with one line command:

```
>>> genabs  articles-directory/  keywords-directory/
```

Multi-modal support

- Processing of article's tables, figures, and equations

Computational resources

- Upgrade of computational resources to meet increased processing needs.

If we can (nearly) fully automate NSR, then both XUNDL and EXFOR should be possible

Thank you so much for your attention!

This work is sponsored by the Office of Nuclear Physics, Office of Science of the U.S. Department of Energy under Contract No. DE-SC0012704 with Brookhaven Science Associates, LLC.