NLP for Biocuration

Working Group

Evaluating AI & NLP Tools for Agricultural Database Curation

Tanya Berardini & Adam Wright, Co-Chairs

March – December 2025

PAG 33

Working Group Members

Adam Wright
Alyssa Proia
Bob Cottingham
Carson Andorf
Doreen Ware
Edwin Ong Jun Kiat
Elly Poretsky
Gabriel Barrett
Jacqueline Campbell
Jodi Callwood
Kapeel Chougule
Laura Tibbs-Cortes
Laurel Cooper
Marcela Tello-Ruiz

Olivia Haley
Parul Gupta
Pierre Larmande
Qi Li
Rex Nelson
Sook Jung
Srikanth Kumar Karaikal
Sudhansu Dash
Sushma Naithani
Taner Sen
Tanya Berardini
Trish Whetzel
Zhiliang Hu

Working Group Goals

1 Evaluate AI/NLP tools for biocuration use cases

2 Develop recommendations for AgBioData community

3 Plan proof-of-concept implementation (Textpresso)

4 Publish manuscript documenting findings

Tools Evaluated

GENERAL AI

ChatGPT

Google Gemini

Claude

GitHub Copilot

SPECIALIZED

Elicit

ChatPDF

PubTator

NotebookLM

CUSTOM / LOCAL

OntoGPT

PubMedBERT

Custom GPTs

Textpresso

Key Findings

What Works

Newer LLMs significantly outperform older ones

'Chat with a PDF' speeds up document review

Prompt engineering and ongoing conversation improves results

Challenges

Hallucination remains a concern

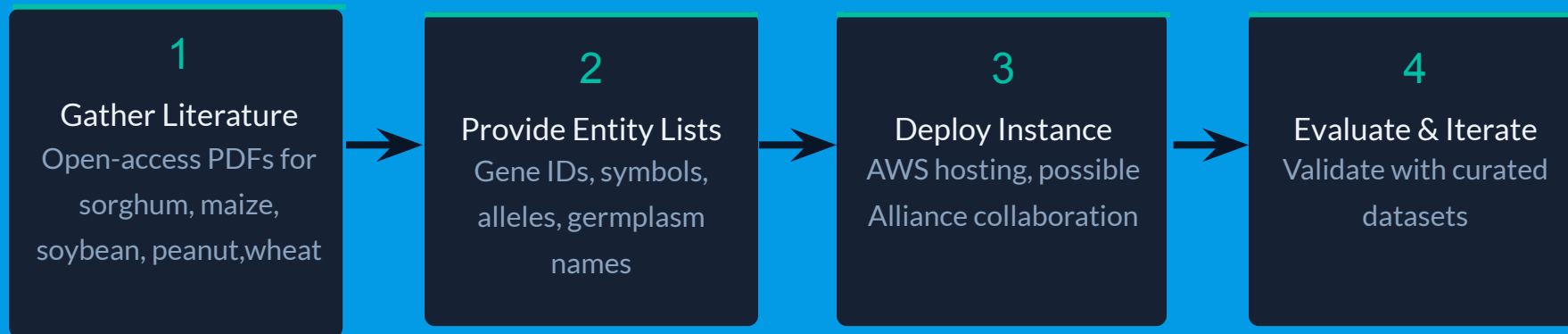Limited coverage of plant-specific entities

Human validation still essential

Textpresso: Proof of Concept Plan

WHAT IS TEXTPRESSO?

Literature mining system developed by Caltech/WormBase, now used by the Alliance of Genome Resources.

Indexes full-text papers with controlled vocabularies and entity recognition for streamlined curation.

1	2	3	4
Gather Literature Open-access PDFs for sorghum, maize, soybean, peanut,wheat	**Provide Entity Lists** Gene IDs, symbols, alleles, germplasm names	**Deploy Instance** AWS hosting, possible Alliance collaboration	**Evaluate & Iterate** Validate with curated datasets

Deliverables

Manuscript

WG learnings and overview of AI/NLP tools for biocuration

Target: Early 2026

Community Survey

AI adoption assessment across 44 databases

12 responses received

Textpresso Proof-of-Concept

Multi-database pilot implementation

Planning phase

Conference

AGBT Agricultural Meeting presentation

April 2026

Next Steps & Looking Ahead

ROADMAP

Jan 2026   ●   Analyze survey results and incorporate into manuscript

Q1 2026   ●   PAG2026 presentation, Finalize manuscript for submission

Q1 2026   ●   Begin Textpresso proof-of-concept setup

Apr 2026   ●   Present at AGBT Agricultural Meeting

2026+   ●   Evaluate Textpresso Proof-of-Concept results (possibly write grant to expand implementation)

Thank You

Questions?