J. Mark Locklear, IT Project Manager, Extension Foundation
David Warren, Senior Director Integrated Digital Strategies, Oklahoma State University
Aaron Weibe, Interim Director of Technology Services and Communications, Extension Foundation
Debbie Brown, Director of Learning Engineering and Special Projects, Thia Corporation

Title

MERLIN: A National Data Management Platform for Extension Resources and Research

Abstract:

MERLIN (Machine-driven Extension Research and Learning Innovation Network) is a purpose-built data management platform designed to unify and standardize Cooperative Extension data across the United States. Designed to overcome long standing challenges of fragmentation, inconsistency, and limited interoperability in Extension data systems, MERLIN aggregates structure and unstructured content from Land-grant institutions using automated feeds, web crawling, and API-based integration. Its architecture facilitates the ingestion of data into ExtensionBot, an AI-powered chatbot that employs Retrieval-Augmented Generation (RAG) models to provide accurate, contextual responses using Extension-only content. However, because MERLIN operates as an independent platform, it can also be integrated with other AI-driven tools and data analytics systems. By transforming Extension resources into sanitized, AI-ready, machine-readable formats, MERLIN enables the consolidation of disparate institutional data into a single, unified platform—supporting national-scale access and integration. The platform also facilitates institutional dashboards that offer insights into user behavior, content utilization, and system performance. In doing so, MERLIN empowers educators, researchers, and practitioners to collaborate, discover, and act on knowledge more efficiently. Beyond serving as a national knowledge repository, MERLIN provides foundational infrastructure for predictive analytics, program evaluation, and integration with agronomic research datasets. Its flexible, privacy-conscious design accommodates varying levels of institutional readiness while supporting the long-term vision of Cooperative Extension as a data-informed, nationally connected ecosystem. MERLIN represents a transformative leap in how Extension data is managed, accessed, and mobilized to support research, education, and outreach at a national scale..

Keywords (5-7 keywords)

- MERLIN, Cooperative Extension, Data Management, Agronomy, AI in Agriculture, Knowledge Dissemination, RAG Models

---

Introduction

Agricultural data management and extension research face a number of challenges, including the fragmentation of data across diverse sources, the variability in data formats, and the need for timely, actionable insights around that data. Additionally, agricultural data are commonly fragmented, distributed, and incompatible, making it challenging to structure data such that it can be readily analyzed with AI.[1] Historically, Cooperative Extension services have struggled with sharing data across states and regions, which hinder efforts to integrate research findings and share on-the-ground expertise effectively. This fragmentation has limited the capacity for comprehensive analysis and rapid dissemination of knowledge to farmers, educators, and policy-makers. As agricultural practices evolve and data volumes grow exponentially, overcoming these challenges becomes critical for ensuring that Extension research remains relevant and impactful in addressing contemporary agronomic issues.

MERLIN is a purpose-built data management platform designed to address the fragmented nature of Extension data. Its primary objective is to bridge the gap between disparate data topics and sources by aggregating and standardizing information from various land-grant institutions and Extension programs. By employing a combination of automated data feeds, web crawling services, and robust data integration protocols, MERLIN creates a unified data environment that enhances accessibility and usability. The platform is engineered to support advanced analytics and AI-driven insights, ensuring that Extension professionals have the necessary tools to interpret complex datasets and translate them into actionable strategies for modern agronomy.

MERLIN is designed to close the gap between data generation and data utilization in the agricultural Extension ecosystem. Unlike traditional Extension repositories, which often require manual searches and lack interoperability, MERLIN employs automated data gathering and retrieval processes, enabling real-time access to high-quality Extension content. Its common format allows for efficient indexing, retrieval, and delivery, making it easier for AI-driven tools like ExtensionBot to deliver accurate and contextually relevant information. Additionally, MERLIN's API-based architecture facilitates the opportunity for seamless integration with existing Extension platforms or tools, ensuring that institutions can leverage its capabilities without overhauling their existing digital infrastructure. Through these innovations, MERLIN empowers Cooperative Extension to meet the evolving needs of the agricultural community more effectively.

Agricultural Data Management Systems: Complementary Approaches

No known system replicates MERLIN's function to bring together fragmented Extension data and content, however there are some complementary and related platforms.

DIRECT4AG connects farms to research insights through real-time data collection. Developed by the National Center for Supercomputing Applications, this platform brings immediate agricultural intelligence directly to producers through Extension channels. Its modular, open-source framework uses Clowder for data management, Geostreams for sensor data processing, and containerized applications for scalability. The system handles diverse datasets—soil moisture readings, satellite imagery, genetic traits, and forecasting models—and is currently being tested with farmers, researchers, and Extension personnel in Illinois and Alabama. By delivering timely, data-driven recommendations to farmers, DIRECT4AG aims to transform traditional Extension services.

Ag Data Commons, operated by USDA's National Agricultural Library, serves as the central repository for agricultural datasets from USDA-funded research. Built on a Drupal-based open data platform (DKAN), it follows FAIR principles—making data Findable, Accessible, Interoperable, and Reusable. The platform offers standardized metadata and seamlessly integrates with federal systems like Data.gov. Researchers nationwide, particularly at land-grant universities, use Ag Data Commons to fulfill USDA data-sharing requirements, enhancing transparency and collaboration opportunities.

While these systems all tackle agricultural data management challenges, they serve distinct purposes:

MERLIN specializes in unifying fragmented Extension data from multiple sources through automated feeds, web crawling, and standardized formatting. Its API-based architecture enables AI-driven analytics and real-time information sharing, making it uniquely suited for Cooperative Extension needs.

DIRECT4AG prioritizes farm-level decision support through real-time data integration, combining sensor readings, satellite imagery, and predictive modeling. Unlike MERLIN's broad Extension focus, DIRECT4AG emphasizes data generated at the farm level, fostering direct communication between researchers, Extension staff, and producers.

Ag Data Commons functions primarily as a stable repository for USDA research datasets, emphasizing long-term storage, accessibility, and interoperability. While MERLIN actively integrates diverse Extension resources and DIRECT4AG delivers real-time analytics, Ag Data Commons provides structured, persistent access to research data for secondary analysis and public transparency.

Together, these complementary systems address different aspects of agricultural data management: MERLIN tackles information fragmentation across Extension services, DIRECT4AG delivers timely insights for on-farm decisions, and Ag Data Commons ensures research data remains accessible and usable for future applications.

MERLIN: System Design and Architecture

Core Functionality

Data Ingestion from Land-Grant Institutions

MERLIN consolidates ingestion of datasets from land-grant institutions through robust, automated methods collection methods based on the data structures that exist at each institution. For institutions providing JSON data endpoints, MERLIN directly consumes structured data feeds. These are nice because the data is already in the desired JSON format that is ready to be consumed by ExtensionBot. Alternatively, for institutions lacking such endpoints, MERLIN employs web crawling scripts to systematically extract and aggregate relevant content from institutional websites. These scripts run seamlessly on the MERLIN platform, automating data collection processes and significantly reducing manual interventions, thereby ensuring a consistent flow of current and comprehensive data. These scripts might either crawl sitemaps (this is ideal) or if sitemaps do not exist then the scripts will crawl the institution's html pages.

Structured and Unstructured Data Integration

One useful way to conceptualize the variety of datasets MERLIN ingests is to view them along a spectrum—from highly structured to unstructured. At the most structured end are institutions that provide data through well-formed JSON endpoints. These feeds are already machine-readable and conform to the format required by downstream tools like ExtensionBot. For example:

```
{
    "title": "Maintaining a Mean, Green Lawn using Integrated Pest
Management - Pests In The Home",
    "link":
"https://pestsinthehome.extension.org/pest-prevention/maintaining-a-mean-gr
een-lawn-using-ipm/",
    "state": "NA",
    "modified_date": "2021-10-05T01:34:16+00:00",
    "content": [
        {
            "content_text": " Most homeowners enjoy a thick, green lawn,
and as the weather gets warmer, keeping that lawn green and clean will get
more challenging. Too much water can leave a lawn susceptible to disease,
while too little water makes it dry and brown. Soil-borne insects such as
grubs can eat the roots, and weeds can make the lawn look unkempt. ...
        }
    ]
}
```

This represents a single resource from a national website—pestsinthehome.extension.org. Since the site is not affiliated with a specific state institution, no state metadata is included. For readability, the content text here is truncated, but in practice, the full document body would be included.

Moving along the spectrum toward less structured sources, a common intermediate case involves websites with XML sitemaps. Sitemaps are machine-readable indexes that list all URLs on a site and often include metadata such as last-modified dates and crawl priorities. These are valuable for data collection because they consolidate discoverable content in one location. However, each listed URL still needs to be fetched and parsed individually, and the success of extraction depends on the consistency of the underlying page structure. Crawling scripts written for these sites are often "brittle," in that any change to the website's layout or structure—such as the addition of modals, new navigation systems, or dynamic content—can break the crawling script.

A slightly less structured but still manageable scenario involves paginated lists of publications or factsheets. In these cases, a crawler can follow a logical sequence of pages and collect data one item at a time. While more labor-intensive than sitemaps, these patterns are relatively easy to accommodate so long as the site and page structure remains consistent.

At the far end of the spectrum are institutions where content is highly decentralized—spread across departments, colleges, or affiliated programs, each with its own subdomain or organizational identity. These environments present challenges that are as much organizational as they are technical. Such cases demand more technical attention and engineering effort, often requiring the development and ongoing maintenance of multiple crawling scripts to accommodate varying site structures and formats. In particularly complex environments, it is often worthwhile to engage directly with institutional IT staff to explore options for consolidating Extension-related resources into a centralized location or directory. When feasible, the implementation of a sitemap can significantly streamline the data collection process by providing a single, structured index of web-accessible resources.

Finally, its with saying a few words about PDF parsing in data collection. PDFs remain a widely used format within the Cooperative Extension system and across many academic and institutional contexts. Any robust data ingestion pipeline must be capable of efficiently accessing and converting PDF content into machine-readable text. In our implementation, we have found success using open-source Python libraries such as **PyPDF2** and **pdfplumber**, which provide reliable tools for extracting structured text from PDF documents. Notably, the field of AI-assisted PDF parsing is evolving rapidly, with new tools and models emerging that significantly enhance accuracy and extraction capabilities. Given the pace of advancement, implement technologies and workflows with flexibility in mind—allowing teams to quickly adapt and integrate improved solutions as they become available.

API-based Data Retrieval for ExtensionBot

While MERLIN's primary function is to serve as a data source for ExtensionBot—the Extension Foundation's AI-powered chatbot—it is not limited to this use case. As outlined in the *Future Use Cases* section, the platform is designed with broader use cases in mind. Within the ExtensionBot ecosystem, MERLIN supplies curated, Extension-specific datasets that power the chatbot's Retrieval-Augmented Generation (RAG) capabilities. MERLIN exposes RESTful API endpoints that allow ExtensionBot to programmatically pull Extension resources into its environment for ingestion. In this context, “ingestion” refers to the process of storing and indexing the retrieved documents in a vector database optimized for semantic search. This infrastructure enables ExtensionBot to rapidly identify and retrieve relevant content in response to user queries, forming the foundation for accurate, context-aware chatbot responses. The communication between MERLIN and ExtensionBot is bidirectional. While MERLIN delivers curated data for ingestion, ExtensionBot can return system updates, ingestion status, and usage logs back to MERLIN. This feedback loop provides real-time visibility into system performance and content readiness, allowing MERLIN users—particularly institutional administrators—to monitor ingestion progress and track the availability of their resources within ExtensionBot.

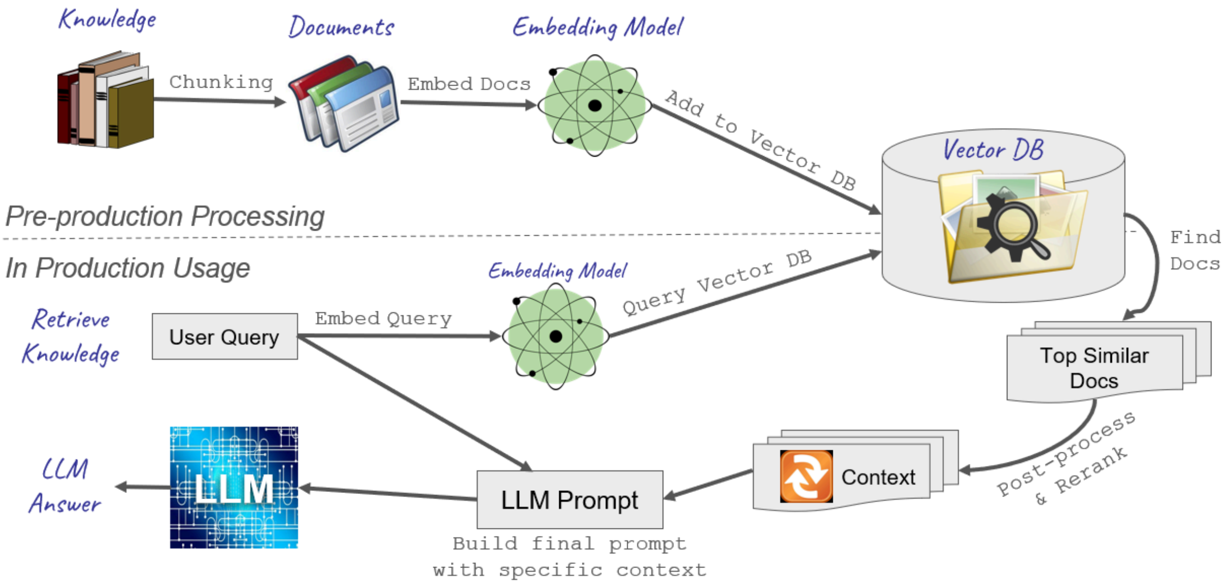
Dashboard Analytics for User Insights

MERLIN provides institutional stakeholders with visibility into ExtensionBot’s usage on their respective websites. When ExtensionBot is deployed, all user interactions—including submitted queries and generated responses—are logged and stored within the ExtensionBot platform. Through an API, MERLIN can retrieve this interaction data, enabling universities and Extension administrators to access detailed, actionable insights on system performance and user engagement. Administrators gain access to key metrics such as popular questions and queries, engagement rates, content utilization trends, and click-through rates to institutional resources. These analytics not only highlight how users are interacting with ExtensionBot, but also help institutions identify content gaps and opportunities for improvement. Additionally, MERLIN provides tools for monitoring and optimizing the data ingestion and retrieval processes, ensuring that the underlying content remains timely, accurate, and relevant. By supporting data-informed decision-making, these capabilities enable institutions to refine their digital strategy, improve user experience, and align Extension programming more closely with evolving stakeholder needs.

3.3. AI and RAG Model Integration

Retrieval-Augmented Generation (RAG) is an AI architecture that enhances the performance of large language models (LLMs) by integrating an external retrieval component into the response generation process. Rather than relying solely on the model’s pre-trained knowledge, RAG models dynamically query a curated knowledge base—such as a vector database of [in the case of ExtensionBot] Extension resources—to retrieve relevant information in real time. The retrieved content is then used as contextual input for the language model, enabling it to generate responses that are both grounded in up-to-date, relevant, domain-specific content and aligned with the user’s query. This architecture significantly improves factual accuracy, domain relevance, and the traceability of generated responses, in addition to significantly reducing hallucinations.[2]

Another primary benefit for institutional users of MERLIN is that it provides a very organized and accessible method for digitizing knowledge documents to be both human and machine-consumable. As mentioned above, in the case of institutionally generated data dumps, MERLIN allows institutions to selectively curate resources that are representative of subject matter expertise, and they can monitor the transformation of their knowledge documents as well as download them as data for consumption by other tools. MERLIN additionally provides API access to ExtensionBot's document processing pipeline for training both the embedding model and large language model. In the diagram below, we briefly examine what ExtensionBot does with the resources obtained through MERLIN.



As depicted in the Pre-production Processing portion of the diagram, the first step that must take place with the digitized knowledge documents is referred to as "chunking". Every base AI model has a limit to the maximum number of tokens that it can effectively handle when processing information. This maximum is used as a criterion for breaking large resources into smaller related chunks that fit within this context window, each continuing to retain the metadata about the original source. Then, the pipeline vectorizes these document chunks by generating embeddings for them. Vectorized representations of the document chunks are stored in a Vector database, and then the information is fully indexed within the resulting vector space.

Once the resources have all been indexed, the information can be retrieved and utilized by ExtensionBot in production usage. ExtensionBot receives the user query as text and vectorizes it so that it can be used to search the Vector DB. The vector DB search results include the top similar documents based on distance in the vector space from the query to the identified resources. These resources undergo post-processing that includes localized re-ranking of the resources based on the chatbot instance's configuration. This context is then combined with the original user query and submitted to the Large Language Model to produce a generalized summary based on the highest ranked results with citations and suggested follow-up questions. All of the conversations that take place within ExtensionBot are also logged for review and evaluation, and institutions can use their dashboards via the MERLIN API to download or query the logs when reporting on their utilization of any instances they have of ExtensionBot running on their institutions domain.

During maintenance cycles, ExtensionBot updates RAG resources through the MERLIN API. Institutions are able to schedule these AI ingestion cycles in their dashboard view, and model updates primarily take place in non-peak hours. Our recommendation will be weekly updates of content.

The Vector DB enables the efficient retrieval of relevant research documents when a user submits a query to ExtensionBot. During the indexing phase, each document chunk is converted into a high-dimensional vector embedding using a trained embedding model, which captures semantic relationships between words and phrases. These embeddings are stored in the vector database and organized in a way that optimizes similarity searches—typically using approximate nearest neighbor (ANN) algorithms. When a query is submitted, it is similarly vectorized, and the database performs a fast similarity search to retrieve the most relevant document chunks based on cosine similarity in the vector space. To further refine results, a re-ranking step is applied, which accounts for institution-specific relevance criteria (e.g., document freshness, authority, or domain-specific keywords). This dual-phase approach—initial broad retrieval via vector similarity followed by localized re-ranking—ensures that ExtensionBot delivers highly accurate and contextually appropriate research materials in response to user queries while maintaining low latency.

4. Applications in Agricultural Extension and Research

4.1. Enhancing Knowledge Access for Farmers, Educators and Researchers

MERLIN will improve how Cooperative Extension knowledge is accessed and applied by key stakeholder groups, including farmers, educators, and researchers. Historically, Extension data and research have been dispersed across dozens of independently maintained institutional websites and repositories, making it difficult to locate, compare, or synthesize information across state lines. MERLIN addresses this challenge by creating a unified, national platform for aggregating, standardizing, and indexing Extension content from land-grant institutions across the country. This unified platform enables broader dissemination of research-based information while preserving the local context and institutional identity of its sources.

For farmers and educators, MERLIN supports more equitable and efficient access to trusted resources—factsheets, bulletins, production guides, and best practices—regardless of where those materials were originally published. It empowers Extension professionals to quickly find peer-reviewed content that can be reused, adapted, or shared in their own programming. MERLIN also powers downstream applications like ExtensionBot, which uses AI to deliver conversational, context-aware answers drawn from this national knowledge base.

Importantly, MERLIN also opens new possibilities for researchers who study agriculture, education, climate resilience, or rural development. By providing a programmatic interface to a curated, well-tagged corpus of Extension outputs, MERLIN facilitates meta-analyses, trend tracking, and the integration of Extension materials into broader interdisciplinary studies. Researchers can now ask new questions—such as how climate adaptation advice has evolved across regions over time—that were previously difficult or impossible to pursue at scale. In this way, MERLIN supports not only the dissemination of knowledge, but also its advancement.

While MERLIN represents a unified national platform for Cooperative Extension data, it also reflects the decentralized nature of the Extension system. Individual datasets remain associated with their originating institutions and states, allowing users to filter, search, and curate content at the state or regional level as needed. This structure preserves the value of locally developed knowledge—a founding principle of Extensions mission—while still enabling cross-institutional discovery and collaboration. Whether users are seeking resources tailored to specific geographic conditions or comparing program strategies across multiple states, MERLIN provides the potential to navigate both local and national contexts within a single interface.

For Extension educators, MERLIN offers a transformative shift in how program development and content discovery can occur. Rather than searching across dozens of separate institutional websites—or starting from scratch when creating new educational materials—educators can now explore a national corpus of curated, research-based content in one place. This reduces duplication of effort and empowers educators to build on existing work, adapt proven materials for local contexts, and accelerate the delivery of high-quality programming. By making it easier to find and repurpose relevant resources, MERLIN helps educators focus more time on engagement and impact, rather than content creation from the ground up.

4.2. Supporting Decision-Making in Agronomy

Integration with agronomic research datasets.

As agricultural challenges grow in complexity and require increasingly interdisciplinary approaches, the ability to connect Extension content with agronomic research datasets becomes essential. MERLIN provides a foundational infrastructure for this integration by offering structured, metadata-rich Extension data that can potentially be cross-referenced with other scientific datasets, including those related to soil health, crop performance, climate variability, pest surveillance, and precision agriculture. By aligning Extension publications with these complementary data sources, the platform can enable a more holistic set of insights that bridge applied research with on-the-ground practice.

The platform’s API-based architecture and standardized json format make it particularly well-suited for interoperability with external agronomic databases, such as USDA’s Ag Data Commons, university-led research repositories, and emerging climate-smart agriculture networks. As MERLIN evolves, additional tagging and taxonomy layers will allow users to link Extension outputs directly to underlying experimental trials, modeling frameworks, or region-specific agronomic parameters. This opens the door for a wide range of applications—from enabling AI-driven synthesis of crop management recommendations, to supporting meta-analyses of Extension data across regions.

Ultimately, the integration of Extension and agronomic research data within MERLIN will strengthen the evidence base behind outreach and education efforts, while creating new opportunities for data-informed decision-making across the agricultural value chain. This capability positions MERLIN not only as a repository of knowledge, but as a dynamic interface between research, practice, and policy.

4.3. Future Use Cases

- Predictive analytics for Extension services.
- Expansion into more granular farm-level data.

5. Challenges and Lessons Learned

Maintaining Data Quality and Accuracy

Ensuring the quality, accuracy, and non-duplication of data across a distributed network of sources remains one of the most significant challenges in implementing a platform like MERLIN. A key strategy to address this issue involves assigning primary responsibility for data curation to the contributing institutions themselves. This approach aligns with institutional priorities, as each organization has a vested interest in ensuring that the data shared is both accurate and representative of its Extension activities. Furthermore, the process of preparing data for MERLIN often serves as a valuable opportunity for institutions to review, refine, and retire outdated or redundant content. This preparatory work not only improves the integrity of MERLIN's dataset but also provides long-term benefits to the institutions by encouraging the development of well-organized, machine-readable data repositories. Whether institutions continue working with MERLIN or pursue alternative AI-enabled platforms in the future, engaging in this data curation process builds internal capacity and positions them to integrate more effectively with a wide range of digital tools and knowledge systems.

Conversely, for Extension websites where MERLIN employs custom crawling scripts, the platform relies entirely on the logic embedded in those scripts to both identify relevant content and extract essential metadata—such as publication dates, categories, and other contextual indicators required for accurate indexing and retrieval. One of the central challenges in this context is managing updates to content and identifying stale or outdated resources. Unlike institutions that provide structured data dumps—where data freshness and deduplication are handled upstream—crawling scripts must infer updates through mechanisms like `last-modified` timestamps in sitemaps or page headers. Detecting and removing stale data presents an even greater challenge. To address this, the platform must either implement delta-based logic that compares existing resources by URL and content signature or periodically conduct full-site crawls followed by complete reprocessing and reindexing of the dataset. While resource-intensive, these strategies are necessary to maintain the accuracy and relevance of the information accessible through the MERLIN platform.

Security and Privacy Considerations

MERLIN is designed to ingest only publicly accessible data from the open web. All content processed by the platform is sourced from pages that are not protected by authentication mechanisms or paywalls. However, it is acknowledged that personally identifiable information (PII) or other sensitive content may occasionally be exposed on public websites—whether inadvertently or by design. The existence of a centralized platform like MERLIN provides an important opportunity for sanitizing and validating such data before it is made available to downstream applications like ExtensionBot.

A key component of MERLIN's privacy strategy lies in its intelligent handling of URL structures during web crawling. RESTful URL patterns often provide meaningful indicators of the type of content hosted at a given path. For example, many institutional websites use standardized paths such as `/people` or `/directory` to organize faculty or staff listings. By explicitly excluding these directories from crawling, MERLIN reduces the likelihood of ingesting sensitive information such as names, phone numbers, email addresses, or physical addresses—content that ExtensionBot is explicitly configured not to surface in its responses.

Similarly, URLs containing paths like `/news` or `/events` are also excluded, as this type of ephemeral or person-centric content falls outside the scope of MERLIN's primary mission. The platform is purpose-built to prioritize research-based Extension outputs—publications, factsheets, and curated knowledge products that reflect the evidence-based ethos of the Cooperative Extension system. By implementing URL parsing logic and content exclusion at the crawler level, MERLIN balances comprehensive data gathering with a strong commitment to privacy, relevance, and institutional trust. These same standards are also applied when working with institutions that provide structured data dumps: we collaborate closely to ensure that only relevant, research-backed content is included, and that personally identifiable or time-sensitive information is excluded.

Authentication and Access Control

To ensure secure and authorized access to data, MERLIN employs a multi-layered approach to authentication and access control. This includes the use of institution-specific UUID bearer tokens, which act as lightweight credentials to facilitate secure communication between MERLIN and downstream systems such as ExtensionBot. These tokens are used to identify and authenticate institutions during API interactions, allowing for fine-grained control over access to datasets. MERLIN will employ IP whitelisting to restrict access to known, trusted servers or systems, further minimizing the risk of unauthorized data requests.

While these methods are currently sufficient for the public and low-risk nature of the data, the architecture is designed with flexibility to incorporate more robust authentication models if needed—such as JWT (JSON Web Token) authentication or OAuth2—should system needs evolve. These enhancements would allow for token expiration, role-based permissions, and federated identity integration, aligning with best practices in enterprise-grade data platforms.

The overall goal is to strike a balance between cost, security, scalability, and ease of participation for land-grant institutions with varying levels of technical infrastructure and resources.

6. Future Directions

As MERLIN continues to mature, its development roadmap is centered around expanding both technical capabilities and strategic integrations to deepen its utility across the Cooperative Extension and agricultural research ecosystems. One of the most promising areas for expansion involves the integration of advanced AI-driven analytics. While MERLIN currently provides foundational support for ExtensionBot's RAG models, future iterations of the platform will include native support for predictive modeling, content clustering, and longitudinal trend analysis across aggregated datasets. These enhancements will enable institutions to anticipate user information needs, detect emerging topics in agricultural science, and optimize the delivery of educational content based on observed patterns in engagement and usage.

Another critical direction involves broadening MERLIN's interoperability with external data systems and knowledge bases. This includes establishing bidirectional integrations with national and global agricultural research databases, such as USDA's Ag Data Commons, international CGIAR repositories, and state-specific digital agriculture platforms. By participating in these larger data ecosystems, MERLIN can serve as a conduit for linking Extension content to complementary datasets such as agronomic trials, climate data, pest surveillance, and market trends—empowering users to perform more complex, multi-source analyses. These integrations will also position MERLIN as a central node in national efforts to standardize and federate agricultural data for AI applications.

Additionally, MERLIN will evolve to support a wider array of user-driven customization and tagging tools. Building on its current architecture, future versions will allow institutions to assign domain-specific metadata, develop taxonomies for localized programming, and participate in shared ontologies for Extension content. These capabilities will enable better content discovery, facilitate cross-institutional collaboration, and strengthen semantic alignment between Extension knowledge and external research domains. By encouraging shared language and classification standards, MERLIN can become an essential infrastructure for synthesizing knowledge across traditionally siloed programs and institutions.

From a platform development standpoint, MERLIN's roadmap includes transitioning key components into cloud-native microservices, enabling modular deployment and improved scalability for institutions of varying size and capacity. This architectural evolution will also support more granular access control, user role management, and audit logging, ensuring that the platform meets emerging compliance requirements while maintaining usability for technical and non-technical users alike.

Finally, MERLIN’s long-term vision includes support for real-time program evaluation and impact tracking. By combining usage analytics from ExtensionBot with behavioral insights and content performance data, the platform will empower institutions to measure outcomes, adapt curricula, and iterate on outreach strategies with greater agility. Over time, MERLIN could serve not only as a knowledge repository, but as a strategic intelligence system for Cooperative Extension—enabling data-informed decision-making at both the local and national levels.

7. Conclusion

MERLIN represents a significant advancement in the digital infrastructure supporting Cooperative Extension and agricultural research. By unifying fragmented data sources, enabling structured ingestion from land-grant institutions, and powering AI-driven tools like ExtensionBot, MERLIN fills a critical gap in how Extension knowledge is collected, standardized, and shared. Through its flexible architecture—supporting both human and machine-readable formats—MERLIN not only facilitates streamlined access to Extension publications, but also sets a new standard for data interoperability, discoverability, and reuse.

As the first unified national platform dedicated for Extension content, MERLIN has created an unprecedented opportunity for educators, researchers, and practitioners to collaborate across institutional and geographic boundaries. It simplifies access to research-backed resources, reduces duplication of effort, and fosters innovation in programming and decision-making. For educators, it accelerates curriculum development. For researchers, it enables new forms of analysis and trend detection. And for farmers and community stakeholders, it ensures equitable access to practical, science-based information.

Looking ahead, MERLIN’s continued development positions it as a cornerstone of the evolving agritech and AI-in-Extension ecosystem. With its potential to support predictive analytics, federated data integration, and real-time impact measurement, MERLIN is not merely a data platform—it is a strategic enabler of Cooperative Extension’s mission in the 21st century. By transforming how knowledge is managed and mobilized, MERLIN will play a central role in advancing the future of agronomy and agricultural outreach.

8. References

- Citations referenced above.
    -
    - 1-https://cast-science.org/wp-content/uploads/2025/03/CAST_AI-in-Agriculture.pdf
    - 2-https://medium.com/digitalgreen-techblog/under-the-hood-of-farmer-chat-journey-to-an-optimised-production-ready-rag-powered-chatbot-589bf5716e27
    -