

Knowledge Graph-Driven Resource Identification and Utilization of Salvia-Derived Biomass

Isaac Johansson

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
isaacjohansson94@binghamton.edu

Walid Ray

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
walidr@ucf.edu

Abstract

The sustainable identification and utilization of Salvia-derived biomass require coordinated interpretation of highly heterogeneous data across taxonomy, phytochemistry, cell wall composition, agronomic practice, extraction processes, and supply chain logistics. Knowledge graphs provide a scalable semantic framework for representing such distributed data as interconnected entities and relations, enabling explainable resource identification and decision support in biomass valorization. This paper presents a system-level analysis of knowledge graph-driven approaches for Salvia-derived biomass resource identification and utilization. It examines architectural patterns, ontology engineering, data harmonization, and reasoning services from an infrastructure perspective rather than from a single algorithmic viewpoint. The discussion emphasizes structural trade-offs among semantic expressiveness, scalability, data quality, provenance, and governance. It further considers fairness, data sovereignty, institutional arrangements, and policy implications for sustainable bioeconomy systems. The paper develops a conceptual foundation grounded in contemporary knowledge graph research and biomass conversion literature. By integrating taxonomic marker data with process and logistics knowledge, knowledge graph systems can transform fragmented biomass information into actionable knowledge for selecting, processing, and valorizing Salvia-derived resources while maintaining robustness, equity, and long-term sustainability.

Keywords

knowledge graph; Salvia-derived biomass; resource identification; ontology engineering; sustainable bioprocessing; data governance.

1. Introduction

The transition toward circular bioeconomy systems has generated renewed interest in plant-derived biomass as a source of chemicals, materials, and energy. Among medicinal and aromatic plants, Salvia species occupy a distinctive position because they produce bioactive compounds of pharmaceutical interest and also generate lignocellulosic residues during cultivation, extraction, and processing. However, the effective utilization of Salvia-derived biomass is constrained by the fragmentation of knowledge across disciplines. Taxonomic identification, cell wall composition, phytochemical profiles, processing conditions, and supply chain constraints are typically documented in separate databases, publications, and institutional repositories. Knowledge graphs offer a general computational strategy for integrating these heterogeneous sources into a unified semantic fabric [1]. They model

domain entities and their relations explicitly, allowing reasoning across data that were not originally designed to interoperate [2].

Large-scale knowledge graphs such as Wikidata and DBpedia have demonstrated the practical value of representing encyclopedic and factual knowledge in graph form [3], [4]. These resources illustrate how entity-centric modeling can support query answering, disambiguation, and cross-domain linkage. At the same time, domain-specific knowledge graphs require more carefully designed ontological commitments because the underlying data are often heterogeneous, uncertain, and sensitive to context. Knowledge graph refinement has therefore become an important research area concerned with quality, completeness, and consistency [5]. For botanical and biochemical applications, ontology development methodologies provide guidance for defining classes, properties, and constraints [6]. Tools such as the Protégé environment support collaborative ontology engineering [7], while coordinated ontology efforts such as the Open Biological and Biomedical Ontologies Foundry promote interoperability across life science domains [8]. The FAIR principles further specify that data should be findable, accessible, interoperable, and reusable [9].

In *Salvia*-related research, cell wall component characteristics have been proposed as identification markers for *Salvia miltiorrhiza* and its close relatives [10]. Such marker data are valuable because taxonomic ambiguity can affect both phytochemical quality and biomass conversion performance. Bioactive compound profiles, especially tanshinones and phenolic acids, are central to the medicinal value of *Salvia miltiorrhiza* [11]. At the same time, residual biomass can be directed toward bioenergy and biomaterial pathways that depend on lignocellulosic composition [12]. Thermochemical conversion processes such as pyrolysis are sensitive to feedstock variability [13]. Integrating these diverse perspectives within a knowledge graph can support more coherent resource identification and utilization decisions. This paper therefore examines the structural, semantic, and governance considerations that shape such systems rather than proposing a narrow computational procedure.

2. Background and Related Work

The conceptual origins of knowledge graphs lie in knowledge representation research concerned with formalizing categories, relations, and logical constraints [14]. Modern knowledge graphs extend these ideas by combining ontology-like schemas with large volumes of instance data. A knowledge graph typically contains nodes representing real-world entities, edges representing typed relations, and ontologies that constrain the meaning of those relations [1]. The graph structure enables traversal-based queries and path reasoning that are difficult to express in traditional relational databases without extensive joins. In recent years, knowledge graphs have been used in domains ranging from biomedical research to industrial data integration, where they support heterogeneous data exploration and decision support [2].

The *Salvia* genus includes several species of medicinal and economic importance, but taxonomic discrimination can be challenging when morphological features are insufficient. Chemical markers, including diterpenoids and polyphenolic acids, are frequently used to distinguish species and evaluate quality [11]. More recently, cell wall component characteristics have been examined as complementary identification markers for *Salvia miltiorrhiza* and close relatives [10]. This type of marker is particularly relevant for biomass utilization because cell wall polymers such as cellulose, hemicellulose, and lignin directly influence pretreatment requirements, enzymatic digestibility, and thermochemical conversion

yields. Linkage of marker data to biomass processing knowledge therefore creates a semantic bridge between botanical identification and downstream engineering.

Biomass systems are often discussed in the context of energy production, biomaterials, and circular resource management. Feedstock composition, moisture content, ash content, and polymer distribution affect conversion efficiency and product quality [12]. Pyrolysis, for example, is a thermochemical pathway whose bio-oil yield and composition depend strongly on lignocellulosic structure [13]. In medicinal plant processing, the residual biomass after extraction may contain useful polysaccharides, fibers, or lignin, but its valorization requires an understanding of both upstream botanical identity and downstream process constraints. Knowledge graphs can represent these interdependencies explicitly, allowing stakeholders to query across botanical, chemical, and engineering datasets in a unified manner. Ontology-driven data integration in the life sciences has already demonstrated the value of such semantic harmonization [8].

3. Knowledge Graph Architecture for Salvia Biomass Resources

A knowledge graph architecture for Salvia-derived biomass can be organized into several interacting layers: data acquisition, ontology and schema management, graph storage, reasoning, and application services. At the schema level, core classes might include Taxon, BiomassLot, ChemicalProfile, CellWallMarker, GeographicOrigin, ExtractionProcess, ConversionPathway, and Product. Important relations include hasTaxon, hasMarker, producedIn, extractedBy, convertedBy, and yields. Additional attributes can capture measurement values, units, timestamps, and provenance. This design supports the representation of both biological identity and engineering context within a single graph model.

The ontology layer benefits from reusing established vocabularies and ontologies wherever possible. Reuse reduces duplication and improves semantic interoperability with external datasets. Methodological guidance for ontology development emphasizes clarifying scope, reusing existing ontologies, defining classes and properties, and validating against competency questions [6]. Collaborative ontology editing environments such as Protégé facilitate iterative refinement and versioning [7]. The Open Biological and Biomedical Ontologies Foundry further provides principles for orthogonal ontology design and coordinated evolution [8]. For Salvia biomass applications, ontology modules could cover plant taxonomy, chemical composition, process engineering, and provenance. The cell wall marker dataset described in [10] can be mapped to a CellWallMarker class with relations to taxonomic entities and measurement methods.

Graph storage and query services must accommodate both structured and semi-structured data. Industry-scale knowledge graphs show that practical systems often combine curated schemas with large volumes of automatically extracted triples [17]. In the Salvia domain, curated data may come from laboratory measurements and taxonomic references, while automatically extracted data may come from literature mining, agronomic databases, and sensor records. Reasoning services can support consistency checking, class inference, and relationship discovery. For example, a knowledge graph may infer that a biomass lot belongs to a particular processing suitability category if its cell wall marker profile matches patterns associated with high lignin content. Such inference adds value but also requires careful attention to uncertainty and provenance.

4. Data Integration and Semantic Harmonization

Data integration is the central challenge in constructing a *Salvia* biomass knowledge graph. Source datasets differ in granularity, vocabulary, measurement units, and temporal resolution. A cell wall marker table may report monosaccharide composition in mass percentages, while a pyrolysis dataset may report feedstock composition in terms of proximate analysis. A knowledge graph must harmonize these representations without losing measurement context. The FAIR principles provide a broad framework for ensuring that source data and derived knowledge remain discoverable and reusable [9]. Linked Data practices support interlinking across distributed datasets using stable identifiers and standardized vocabularies [15].

Quality assessment is critical because erroneous links or inconsistent values can propagate through reasoning and decision services. Knowledge graph refinement techniques address issues such as missing types, incorrect relations, and duplicate entities [5]. Quality frameworks for Linked Data emphasize dimensions including accuracy, completeness, consistency, and provenance [16]. In *Salvia* biomass integration, entity resolution may need to reconcile species names, accession numbers, biomass batch identifiers, and geographic coordinates. Without robust identity management, a graph may merge distinct biomass lots or fail to connect a marker profile with its corresponding extraction residue. Provenance records are therefore as important as the asserted triples themselves.

A concrete integration scenario illustrates these challenges. Suppose a research group records cell wall component characteristics for several *Salvia* accessions [10]. A second group records extraction yields of tanshinones from the same or related accessions [11]. A third group models the pyrolysis behavior of residual biomass with different lignocellulosic profiles [13]. In a knowledge graph, the accession entity serves as the anchor linking marker data, phytochemical data, and conversion data. The graph can then answer queries such as which residual biomass lots with high lignin content are more suitable for biochar production or which close relatives of *Salvia miltiorrhiza* share similar cell wall marker patterns. This integrated perspective is difficult to achieve with isolated spreadsheets or discipline-specific databases.

5. Resource Identification and Decision Support

Resource identification in *Salvia*-derived biomass involves matching available biomass materials to potential utilization pathways. A knowledge graph supports this process by making taxonomic, compositional, and process-related constraints explicit and queryable. For example, cell wall component markers can help distinguish *Salvia miltiorrhiza* from morphologically similar relatives [10]. Such identification is important because different species may contain different concentrations of bioactive compounds and different lignocellulosic structures. When a biomass lot is linked to a verified taxonomic identity, downstream reasoning can retrieve relevant processing constraints and expected performance indicators.

Decision support can be organized around competency questions such as which biomass lots meet the compositional requirements for a given conversion process, which extraction residues are available near a processing facility, and which species substitution risks are associated with a particular biomass batch. The knowledge graph reasoning layer can evaluate typed paths connecting biomass lots, marker profiles, geographic origins, and conversion pathways. Unlike black-box predictive models, graph-based recommendations can expose the evidence chain linking a recommendation to specific measurements and provenance records. This explainability is valuable in regulatory and commercial contexts where decisions must be justified.

The integration of phytochemical and thermochemical knowledge enables multi-objective resource allocation. A biomass lot with high concentrations of tanshinones may be more valuable for pharmaceutical extraction, while its residues may be better suited for pyrolysis if lignin content is high [11], [13]. The graph can represent these alternatives as distinct pathways from the same biomass entity. Decision support queries can then evaluate trade-offs among economic value, energy yield, and environmental impact. Because the graph preserves provenance and measurement metadata, stakeholders can audit why a particular pathway was recommended and under what assumptions. This characteristic aligns with emerging demands for transparent and responsible data-driven systems.

6. System-Level Trade-offs and Architectural Considerations

The design of a Salvia biomass knowledge graph involves recurring trade-offs between semantic expressiveness and computational scalability. Highly expressive ontologies can capture subtle distinctions among cell wall markers, taxonomic relationships, and process conditions, but they increase modeling complexity and may reduce query performance. Lightweight schemas improve agility and ease of population but may fail to support sophisticated reasoning. Knowledge graph surveys indicate that practical systems often adopt a middle path in which a curated ontology is used for core entities while large amounts of instance data are stored in a more flexible graph structure [2]. Industry experience similarly suggests that schema evolution, data quality, and user access patterns are more pressing concerns than raw expressiveness in many deployments [17].

Centralization and federation present another structural tension. A centralized knowledge graph simplifies query processing and consistency enforcement but may create a data governance bottleneck. A federated approach distributes data ownership and maintenance responsibilities but requires interoperable identifiers, contracts, and query routing. For Salvia biomass, data may originate from botanical gardens, agricultural stations, extraction facilities, and research laboratories. Federation supports local autonomy, but it introduces latency, partial availability, and provenance complexity. The choice between these architectural patterns depends on institutional readiness, funding models, and the sensitivity of source data.

Data quality and uncertainty also shape system design. Biomass measurements are subject to sampling variation, instrument calibration, and seasonal effects. A knowledge graph must represent measurement context and confidence rather than only asserted values. Refinement processes can improve graph quality over time by detecting inconsistencies and suggesting corrections [5]. Quality metrics for linked data provide operational guidance for monitoring accuracy and completeness [16]. In the Salvia context, a cell wall marker value may be valid only under specific growth conditions. The knowledge graph should therefore link measurements to cultivation and analytical protocols, allowing users to assess comparability before making decisions.

7. Governance, Fairness, and Policy Implications

Governance is critical because Salvia-derived biomass knowledge graphs link biological data, traditional knowledge, agricultural records, and commercial process information. Access control, licensing, and benefit sharing must be addressed at the system level. Open data initiatives emphasize the potential for transparency and innovation [19], but openness can conflict with the interests of smallholder farmers, indigenous communities, and commercial partners. The commons governance literature demonstrates that sustainable resource institutions require clear boundaries, participation, monitoring, and conflict resolution

mechanisms [20]. A knowledge graph is not only a technical artifact but also an institutional arrangement for data stewardship.

Fairness concerns arise when data integration privileges certain actors or regions. If the graph primarily integrates data from large research institutions and industrial partners, smallholder producers may be excluded from decision-making and benefit distribution. Data sovereignty requires that communities retain control over how their knowledge is represented and shared. In Salvia biomass systems, taxonomic and agronomic knowledge may originate from local producers who have long cultivated and selected specific species. A governance framework should recognize these contributions through provenance attribution and equitable access rules. Knowledge graphs can support such governance by tracking data origin and usage policies as first-class entities.

Policy implications extend beyond data access. Regulatory frameworks for medicinal plants, biomass conversion, and biodiversity conservation vary across jurisdictions. A knowledge graph can encode regulatory constraints, certification requirements, and sustainability criteria, enabling compliance-aware resource utilization. For example, a query could identify biomass sources that satisfy organic certification and avoid protected species. However, the graph must be maintained with current legal information, and algorithmic recommendations must not replace human judgment in matters of legal and ethical significance. Transparent provenance and audit trails are therefore essential to responsible deployment.

8. Sustainability and Lifecycle Perspectives

Sustainability considerations are central to the utilization of Salvia-derived biomass. A knowledge graph can support lifecycle-oriented reasoning by linking feedstock characteristics, conversion pathways, energy inputs, and coproduct outputs. Biomass valorization is most effective when it considers the full material flow from cultivation through extraction to final conversion [12]. For Salvia species, the medicinal value of extracted compounds may justify cultivation, but residual biomass should not be discarded as waste. Instead, residues can be evaluated for bioenergy, biochar, fibers, or other coproducts depending on their cell wall composition [13].

Knowledge graph representations can capture environmental indicators such as land use, water consumption, greenhouse gas emissions, and biodiversity impact. These indicators may be associated with geographic origins, cultivation practices, and conversion technologies. Decision support can then compare pathways not only by economic yield but also by environmental performance. For example, a pathway that produces high-value tanshinones while converting lignin-rich residues into stable biochar may offer long-term carbon sequestration benefits. The graph can represent these synergistic pathways and support scenario comparison.

From a lifecycle perspective, the temporal dimension is important. Biomass composition changes with harvest time, storage conditions, and post-harvest handling. A knowledge graph that records temporal metadata can support time-aware queries and seasonal planning. It can also connect historical agronomic data with future climate scenarios to assess the resilience of Salvia cultivation. This forward-looking perspective aligns with the broader goal of building adaptive and sustainable bioeconomy infrastructures. Knowledge graphs provide a flexible substrate for this complexity, but their sustainability value depends on continuous data curation, community engagement, and institutional commitment.

9. Deployment and Infrastructure Requirements

Deploying a Salvia biomass knowledge graph requires practical infrastructure for data ingestion, storage, query processing, and user access. Data ingestion pipelines must handle structured laboratory records, semi-structured agronomic spreadsheets, and unstructured literature. Provenance capture should occur at ingestion time so that every assertion can be traced to its source. Graph storage systems must support scalable traversal and reasoning over potentially millions of entities and relations. Industry-scale deployments show that indexing, caching, and query optimization are crucial for interactive performance [17]. In the Salvia domain, query patterns may vary from exploratory scientific queries to operational supply chain lookups.

Interoperability depends on stable identifiers, controlled vocabularies, and application programming interfaces. FAIR-aligned data publication supports the discovery and reuse of distributed datasets [9]. Linked Data principles enable cross-referencing with external taxonomies, chemical databases, and geographic resources [15]. These interoperability layers reduce the cost of integrating new datasets and allow the knowledge graph to evolve as scientific knowledge advances. However, infrastructure alone is insufficient. Capacity building, documentation, and user training are necessary to ensure that diverse stakeholders can contribute to and benefit from the system.

Long-term sustainability of the knowledge graph itself requires institutional governance, persistent funding, and versioned release processes. Ontologies and data models must evolve without breaking downstream applications. Schema evolution strategies, backward compatibility policies, and deprecation workflows are therefore part of infrastructure planning. A modular architecture can isolate changes and reduce maintenance risk. In addition, edge computing and sensor data streams may eventually support real-time monitoring of biomass quality at cultivation or processing sites. Such extensions would allow the knowledge graph to move from static data integration toward dynamic resource management, provided that data quality and model governance remain rigorous.

10. Conclusion

Knowledge graph-driven approaches offer a coherent strategy for integrating the heterogeneous data required to identify and utilize Salvia-derived biomass. By representing taxonomic markers, cell wall characteristics, bioactive profiles, conversion pathways, and provenance in a unified semantic graph, these systems can support transparent and explainable resource decisions. The present analysis has emphasized system-level concerns rather than narrow algorithmic formulations. Architectural design must balance expressiveness, scalability, federation, and maintainability. Data integration requires careful ontology reuse, entity resolution, and quality assessment. Governance arrangements must address fairness, data sovereignty, and compliance. Sustainability and lifecycle perspectives connect biomass resource identification with broader circular economy objectives. Future deployments will depend on robust infrastructure, institutional commitment, and the ability to adapt to evolving scientific knowledge and policy environments. A knowledge graph is not a static database but a dynamic socio-technical infrastructure whose value emerges from continuous curation, inclusive governance, and responsible use.

References

1. Nickel, M., Murphy, K., Tresp, V., & Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1), 11–33.

2. Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys*, 54(4), 1–37.
3. Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78–85.
4. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167–195.
5. Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3), 489–508.
6. Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology*. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05.
7. Musen, M. A. (2015). The Protégé project: A look back and a look forward. *AI Matters*, 1(4), 4–12.
8. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11), 1251–1255.
9. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
10. Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
11. Wang, X., Morris-Natschke, S. L., & Lee, K. H. (2007). New developments in the chemistry and biology of the bioactive constituents of Tanshen. *Medicinal Research Reviews*, 27(1), 133–148.
12. Ragauskas, A. J., Williams, C. K., Davison, B. H., Britovsek, G., Cairney, J., Eckert, C. A., Frederick, W. J., Hallett, J. P., Leak, D. J., Liotta, C. L., Mielenz, J. R., Murphy, R., Templer, R., & Tschaplinski, T. (2006). The path forward for biofuels and biomaterials. *Science*, 311(5760), 484–489.
13. Mohan, D., Pittman, C. U., & Steele, P. H. (2006). Pyrolysis of wood/biomass for bio-oil: A critical review. *Energy & Fuels*, 20(3), 848–889.

14. Sowa, J. F. (2000). Knowledge representation: Logical, philosophical, and computational foundations. Course Technology.
15. Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The story so far. *International Journal on Semantic Web and Information Systems*, 5(3), 1–22.
16. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for Linked Data: A survey. *Semantic Web*, 7(1), 63–93.
17. Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: Lessons and challenges. *Communications of the ACM*, 62(8), 36–43.
18. McKendry, P. (2002). Energy production from biomass (part 1): Overview of biomass. *Bioresource Technology*, 83(1), 37–46.
19. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258–268.
20. Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.