New insights on agricultural innovation paths based on a data-driven taxonomy of technologies

Hannah Gerits
Food and Agriculture Organization of the United Nations, Rome, Italy – hannah.gerits@fao.org
Christian Mongeau
Food and Agriculture Organization of the United Nations, Rome, Italy – christian.mongeau@fao.org
Yue Zhang
Food and Agriculture Organization of the United Nations, Rome, Italy – yue.zhang@faol.org
Valeria Pesce
Food and Agriculture Organization of the United Nations, Rome, Italy – valeria.pesce@fao.org
Carola Fabi
Food and Agriculture Organization of the United Nations, Rome, Italy – carola.fabi@fao.org
Delgermaa Chuluunbaatar
Food and Agriculture Organization of the United Nations, Rome, Italy – delgermaa.chuluunbaatar@fao.org

Abstract [1]

A persistent challenge in agricultural statistics is the lack of standardized frameworks for categorizing emerging technologies, hindering cross-national comparisons and evidence-based policy development. This study presents an AI-driven approach to systematically categorize agrifood technology patents through a dual-taxonomy framework.

Using more than 3.5 million agrifood patent applications from PATSTAT (1980–2023), we develop a two-layered taxonomy that simultaneously classifies patents by (1) underlying technology characteristics and (2) intended application domain. A fine-tuned LLM extracts structured information from patent titles and abstracts, capturing both technical specifications and functional objectives. The extracted data undergoes unsupervised hierarchical clustering, generating four hierarchical levels — macro (Food Value Chain stage), meso, micro, and nano — complemented by a parallel technology axis spanning five domains.

Unlike traditional patent classification systems relying on predetermined categories, our approach discovers emergent patterns in the data, revealing previously unrecognized connections between technologies and applications. The dual-taxonomy structure enables multidimensional analysis by both technical approach (e.g., sensor technologies, biotechnology) and agricultural purpose (e.g., precision farming, sustainable production).

The resulting taxonomy serves as the foundational classification layer of FAO's Agrifood Systems Technologies and Innovations Outlook Knowledge Base (ATIO KB), supporting policymakers, investors, researchers, and innovators in making informed decisions to accelerate agrifood systems transformation. The framework addresses critical needs in agricultural statistics by providing standardized categorization for cross-national comparisons, systematic tracking of innovation patterns, identification of technology clusters relevant to sustainable development goals, and enhanced capacity for evidence-based policy formulation in digital agriculture.

This work contributes to the growing field of agricultural data science by demonstrating how advanced AI techniques can transform patent data into actionable intelligence for agricultural policy and innovation management. The approach offers a scalable, adaptable framework that can evolve with emerging

[1] The text and materials in this manuscript are free from any copyright violations.

technologies while maintaining consistency in categorization standards essential for longitudinal agricultural statistics and policy analysis.

Keywords: Agrifood innovation; Food Value Chain; Hierarchical taxonomy; Innovation indicators; Large language models

1. Introduction and related work

1.1 The Challenge of categorising agricultural technologies

Systematically categorising agricultural technologies presents challenges that existing classification frameworks have struggled to resolve. At the data level, the sector generates heterogeneous information from diverse and incompatible sources, making it difficult to construct unified, analysis-ready datasets: At the conceptual level, the rapid pace of agricultural innovation compounds the problem: as new technology classes emerge, the absence of standardised descriptive conventions creates a lag between innovation and systematic understanding [1]. In response, researchers have developed multidimensional classification frameworks that assess technologies along several criteria simultaneously: sensor types, communication protocols, application contexts, and implementation constraints [2]. More holistic approaches additionally incorporate behavioural determinants such as digital literacy, scalability, trust, and infrastructure readiness [3]. These frameworks represent a genuine advance over single-criterion schemes, but they are designed for manual application at the scale of a literature review or a targeted technology survey, not for automated, large-scale processing of the millions of patent documents that constitute the global agrifood innovation record, nor do they provide the functional, value-chain-anchored structure needed to map innovations onto their position in the food system.

Beyond the conceptual challenge, the real-world scenario that prompted this research is an actual effort to build a qualitative-information catalog of agrifood innovations: the FAO Agri-food systems Technologies and Innovations Outlook (ATIO) Knowledge Base (KB). The ambition of building a comprehensive KB that is capable of representing and analyzing the landscape of innovation over time dictates that categorization cannot be done manually and that the categories or classes themselves cannot be static and cannot be pre-defined: they have to be drawn from the reality of the existing innovations and be constantly updated.

The KB's coverage spans all types of innovations, from technological to social, institutional, financial, and policy. In this context, the patents data-driven taxonomy will cover only the technological domain, but the methodology can be applied to other domains over time, as suitable corpora at a similar scale to the PATSTAT database are identified.

1.2 Patent data as a source for innovation intelligence

Patent data have become one of the most widely used sources for measuring and tracking technological innovation. Unlike R&D expenditure, which measures inputs to the inventive process, patents are direct outputs: completed inventions documented and submitted for examination [4] [5] [6]. Each granted patent has passed external validation requiring novelty, inventive step, and industrial applicability [7] [8]. Filing costs and strategic considerations further act as an economic filter, meaning patents tend to represent innovations of positive anticipated significance [6] [9]. Patent databases are also publicly accessible, cover most industrialised economies, and provide longitudinal records extending back to the nineteenth century [4] [10].

For innovation intelligence, the informational richness of patent records is as important as their coverage. Patents contain detailed bibliographic data enabling multidimensional analysis across technological fields, geographies, actors, and time periods [11] [12]. The IPC and CPC systems provide standardised, granular technology-field codes for tracking and comparing innovation activity across jurisdictions [10] [13]. A key methodological consideration is the treatment of multi-office filings: the same invention is routinely filed in several national offices, producing multiple records for a single inventive act. Grouping

applications into DOCDB patent families removes this duplication and provides a cleaner unit of analysis [5], approach adopted here, working with deduplicated DOCDB families from PATSTAT.

Patent data are not without limitations. Most fundamentally, patents capture only a subset of innovation: some inventions are not patentable, and even patentable ones are not always filed — firms may prefer trade secrecy or consider protection costs unwarranted [10] [14]. The propensity to patent varies substantially across sectors, technology types, and countries, meaning patent counts do not reflect innovation activity uniformly across the Food Value Chain (FVC) [5]. Patent value is also highly skewed: a small number represent breakthrough innovations while the majority have limited commercial significance [5]. Finally, counts from different offices are not always directly comparable due to differences in examination standards and filing costs across jurisdictions [6]. Family-level deduplication partially mitigates the last concern, but the preceding limitations remain relevant for contextualising the findings.

1.3 Existing classification systems and their limitations

Within information science, Knowledge Organization Systems (KOS) — encompassing taxonomies, classification schemes, thesauri, authority files, and ontologies — serve a common purpose: supporting retrieval, navigation, semantic consistency, and interoperability [15].

Several KOS exist for agrifood innovations (e.g., AGROVOC, CGIAR's Scaling Readiness, AgFunder's AgriFoodTech taxonomy), but none constitute a unified, empirically grounded, and scalable framework for mapping innovation landscapes. The literature points less to the absence of classification efforts than to their fragmentation: existing systems generally capture specific dimensions of innovation rather than the multidimensional structure of the landscape as a whole. The most widely used frameworks for economic and technological activity (ISIC and its regional derivatives NACE and NAICS) organise establishments by production process [16] [17] [18], making the agrifood technology sector structurally difficult to isolate. The Central Product Classification (CPC) offers product-level granularity but remains weak for software-centric platforms and multi-function solutions spanning hardware, data services, and agronomic advisory functions (UNSD, 2024). The International Patent Classification (IPC) offers the closest approximation to a technology taxonomy and is useful for patent landscaping, but classifies inventions by technical feature rather than by functional use-case or position in the FVC, and is limited to patented inventions only [19]. Across these systems, a common structural problem remains: they are essentially single-axis systems applied to a fundamentally multi-axis innovation landscape.

This limitation provides the rationale for the present study. Consistent with FAO's ATIO KB concept note [20] [21], the agrifood innovation domain is both rapidly evolving and inherently multidimensional, requiring at minimum a distinction between technology type and use case. To our knowledge, no existing vocabulary combines these two dimensions at the scale and granularity required to map the global agrifood patent record. The aim of this study is therefore not to replace existing semantic infrastructures, but to develop a complementary, data-driven taxonomy tailored to the analytical requirements of large-scale agrifood innovation mapping.

1.4 Patent-based innovation analysis in agriculture

Patent data have been applied to agricultural innovation analysis across a wide but fragmented set of specialised domains. Existing reviews tend to focus on specific technological subfields (biotechnology, nanotechnology, new plant varieties, green and intelligent agriculture) or on particular geographic regions [22]. Within these domains, patent data have been used to study the relative roles of private firms, public research institutions, and collaborative partnerships in generating agricultural innovation, particularly in agbiotech [23]. Longitudinal analysis has revealed clear shifts in technology focus over time, e.g., the mid-1990s peak in genetic engineering tools for plant DNA manipulation followed by a transition toward high-yielding commercial varieties in the 2000s [24] [25].

Methodologically, the field has relied primarily on the IPC A01 hierarchy to retrieve and organise relevant patents [26]. Beyond retrieval, researchers have attempted to impose functional structure on patent

corpora: one established approach classifies patents by use (breeding, cultivation, and processing) to map innovation across stages of agricultural production [27]. More sophisticated approaches build transdisciplinary ontology schemas linking agricultural applications to constituent sub-technologies [28], or combine text mining with scientometric analysis to identify technology opportunities within agrifood domains [29]. At the frontier of scale, recent work has applied Main Path Analysis to corpora exceeding two million patents, identifying twelve major technology pathways tracing the agro-industry's evolution from GPS-based management to AI, IoT, and blockchain applications [30]. Despite this progress, existing studies share a common structural limitation: they are anchored either in a single technology domain or a narrow production stage, and none provides a systematic, cross-cutting taxonomy that simultaneously maps the full agrifood value chain and the multi-domain technological character of innovations at the scale of the global patent record.

1.5 Text mining and NLP approaches to patent classification

Patent documents present distinctive challenges for automated text analysis: they combine highly specialised technical terminology, complex claim structures, and strategically ambiguous language designed to maximise legal coverage [31] [32]. The principal motivation for automated approaches has been to reduce the human effort required to classify patent applications while gaining the flexibility to adapt quickly to emerging technology clusters outside established classification codes [33]. Early text mining methods followed a keyword-based workflow using TF-IDF weighting, inter-document similarity, and clustering algorithms [34]. Topic modelling, and LDA in particular, extended this by representing each patent as a mixture of latent topics inferred from term co-occurrence in claims [35]; combining LDA with supervised classifiers such as SVMs has been shown to enable automatic patent classification without expert judgement [36] [37].

The shift to dense embeddings marked a significant advance. Word embedding models such as Word2Vec and GloVe overcome core limitations of term-frequency methods, including sensitivity to synonyms, jargon, and vocabulary drift [38]; training on domain-specific patent corpora amplifies these gains, with embeddings trained on over five million patents increasing classification precision by seventeen percentage points over general-domain embeddings [39]. At the sentence level, PatentSBERTa achieved 54% accuracy and F1 scores above 66% on nearly 1.5 million patents [40] [41]. Fine-tuned BERT on datasets exceeding two million patents further outperformed CNN-based approaches, demonstrating that patent claims alone provide sufficient signal for state-of-the-art classification [42], with domain-adapted variants becoming standard baselines across IPC and CPC hierarchies [40] [43].

The most recent frontier involves LLMs applied directly to patent analysis under zero-shot, few-shot, and retrieval-augmented conditions, demonstrating competitive performance against fine-tuned encoder models [43] [31]. In parallel, unsupervised semantic clustering has emerged as an alternative to supervised classification for large corpora where manual labelling is impractical [33] [44]. The present paper combines both lines of work: LLMs are used not for supervised classification but for structured semantic extraction from patent abstracts (producing labelled representations of each family's functional use-case and technology domain) which are then embedded and clustered unsupervised to build a novel taxonomy from scratch, without reference to pre-existing IPC codes or manually defined category sets.

1.6 The ATIO knowledge base and the role of a data-driven taxonomy

The FAO Agrifood Systems Technologies and Innovations Outlook (ATIO) is a global initiative aimed at supporting agricultural stakeholders -, including policymakers, statisticians and researchers - in monitoring innovation trends, identifying technology gaps, and informing evidence-based policy and investment decisions in the agrifood sector. A central component of the initiative is the ATIO Knowledge Base (KB), a federated and AI-assisted system for organizing information on agrifood innovations across the full innovation life cycle. The KB integrates content from heterogeneous sources, including formal research, technology databases, and, progressively, grassroots innovation processes, and restructures this information through a common semantic architecture. This architecture combines an innovation profile with multiple curated classification systems — such as innovation type, use case, readiness level, adoption level, themes, actor types, and geographic categories — designed to support search, filtering, comparison,

and interoperability across the system. A central component of this architecture is the ATIO data-driven taxonomy, which is intended to serve as the primary typology for the initiative. Unlike the other classifications in the system, which are designed by experts and manually curated, the data-driven taxonomy is generated computationally from patent data and designed to reflect the actual structure of the innovation landscape rather than a predetermined thematic framework. It is intended to evolve annually as new patent applications are published.

The work presented in this paper constitutes the methodological foundation for this data-driven taxonomy. It addresses the core challenge identified in the ATIO design: that no existing classification is simultaneously scalable to millions of patent records, capable of separating technology type from application domain, and structured to support the longitudinal and cross-national comparisons required for agricultural statistics and policy analysis.

1.7 Contribution of this paper

This paper describes the methodology and results of a data-driven dual taxonomy of agrifood patents at scale, simultaneously classifying by technology type and agricultural application, and serving as the foundation for FAO's ATIO Knowledge Base. It makes three main contributions.

First, it develops a dual taxonomy classifying patent families simultaneously by technology type and agricultural application. By separating these two dimensions, the framework moves beyond conventional patent classifications that capture technical subject matter but do not distinguish between what a technology is and what it is for.

Second, it contributes a methodological pipeline for large-scale agrifood innovation mapping, combining patent-family consolidation, multilingual normalisation, LLM-based structured extraction, embedding, clustering, and hierarchical taxonomy construction — providing a reproducible, scalable, and updatable approach for deriving interpretable taxonomic structures from heterogeneous patent data.

Third, the paper positions this taxonomy within a broader knowledge organization and policy intelligence framework. The resulting classification serves as the methodological foundation of the data-driven taxonomy layer of FAO's ATIO Knowledge Base, contributing both to the study of agrifood innovation and to the design of semantic infrastructures that support discovery, comparison, and longitudinal analysis of technological change.

2. Methods

This section describes the end-to-end methodology used to construct the dual-taxonomy framework for agrifood innovation intelligence. The approach combines a fine-tuned large language model for structured information extraction with unsupervised clustering to discover emergent thematic groupings, without relying on predefined categories. The resulting framework classifies patent families simultaneously along two independent axes regarding the underlying technology and the intended agricultural application, and organises each axis into a four-level hierarchy: macro (the FVC stage, described in Section 2.3), meso, micro, and nano.

2.1 Corpus construction and semantic normalisation

The analytical corpus was drawn from the European Patent Office's PATSTAT Global database (Spring 2024 edition), which aggregates filings from national and regional patent offices worldwide. Patent families were selected by filtering against a curated set of agrifood-relevant International Patent Classification (IPC) prefixes covering primary agricultural production (A01), food processing and foodstuffs (A21–A23), veterinary instruments (A61D), fertilisers (C05), peptides and biotech-relevant chemistry (C07G, C07K), fats, oils and waxes (C11B, C11C), fermentation and microbiology (C12C–C12P), the sugar industry (C13), and water supply (E03B), among others. Applications with a filing date between 1980 and 2023 were retained.

To prevent duplicate counting, only the earliest application from each DOCDB patent family was retained, aggregating all IPC codes for full technological coverage. Thus, analysis focused on patent families rather than individual applications.

Titles and abstracts for each family were combined into a single input for processing. Non-English texts were translated to English. This ensured consistent language for model training and cross-national comparison.

2.2 Methodological approach to structured extraction: rationale and development

A key challenge was separating patent text into two dimensions: what the technology is (technical mechanism) and what it is for (agricultural use case). In patent abstracts these aspects are often intertwined, making separation difficult.

Before settling on the final approach, several alternatives were evaluated. Patent titles alone were too brief to support reliable extraction. Part-of-speech tagging could not distinguish technical from functional terms without supervision. Sentence-level segmentation failed because patents typically mix technical and functional language within the same sentences. Rule-based keyword matching lacked generalization, and zero-shot classification with sentence embeddings required prototype construction and performed poorly in the patent domain.

A feasibility study was conducted to evaluate whether LLMs could perform this dual extraction reliably at scale. Three instruction-tuned models (Mistral-7B-Instruct, LLaMA-3-8B-Instruct, DeepSeek- 6.7B-Instruct) were tested on a stratified sample of 700 patents. Each model was prompted to extract a plain-English sentence and a short tag list for each of the two dimensions. Outputs were evaluated across 180 combinations of model, input variant (full text vs. first abstract sentence), embedding model, and representation strategy, using a multi-stage decision logic covering output quality (empty output and semantic leakage rates), robustness (sensitivity to input variant and cross-dimension overlap), and clusterability (silhouette score, Davies–Bouldin index, semantic diversity). Mistral-7B-Instruct consistently produced the best trade-off across these criteria, with low cross-dimension overlap (mean cosine similarity < 0.16) and high semantic diversity in the resulting embeddings.

However, deploying Mistral-7B-Instruct over the full corpus of 3.5M+ patents was not feasible, due to substantial GPU infrastructure costs and the risk of hallucinations not easily caught by automated flags. Fine-tuning a smaller, more efficient model on domain-specific annotated examples was therefore recommended, expected to achieve higher precision and consistency than prompting a larger general-purpose model while remaining tractable at full corpus scale. Gemma 3 1B-IT (Google), a 1-billion-parameter instruction-tuned model, was selected as the base for this fine-tuning process, as described in Section 2.4.

2.3 Food Value Chain classification

Each patent family was assigned a primary FVC stage as the first substantive processing step. This FVC assignment constitutes the macro level of both taxonomy axes: all subsequent clustering steps are performed independently within each FVC stratum, ensuring that meso, micro, and nano clusters remain coherent within a single stage of the agrifood system. The FVC schema comprises six categories (Production, Aggregation, Processing, Distribution, Consumption, and Cross-Cutting) plus an Other category for filings outside the agrifood domain.

Production covers innovations in agricultural, aquaculture, and livestock systems where the commodity remains in its original form, including agricultural inputs, cultivation and precision farming, fisheries, post-harvest handling, and AI/robotic field applications. Aggregation covers technologies for collecting and consolidating commodities without altering their form, including logistics, tracking, and protective transport. Processing covers transformations of the raw commodity, including milling and extraction, thermal and chemical preservation, fermentation, and packaging. Distribution covers movement of commodities to wholesale, retail, or end consumers. Consumption covers in-home food preparation, safety monitoring, and nutritional tracking. Cross-Cutting is assigned when a technology operates across multiple FVC stages or when available text is insufficient to determine a single predominant stage with confidence.

Classification was performed using GPT-4o-mini via the OpenAI Batch API. This commercial model was preferred for this step (rather than the fine-tuned Gemma model) because the task involves assigning each

patent to one of a small number of well-defined, stable categories, for which a well-prompted general-purpose model is well-suited, and because it is a one-time assignment rather than the four-field structured extraction that justified dedicated fine-tuning. Input texts were compressed using LLMLingua-2 prior to batching to manage token-budget constraints. The model was prompted with detailed category definitions and instructed to return a structured JSON response containing the assigned label and a brief rationale; labels outside the permitted set were discarded.

2.4 Structured information extraction via fine-tuned language model

Each patent family was processed to extract four structured semantic fields forming the foundation of the dual taxonomy: a technology description (one sentence describing the technical mechanism of the invention, explicitly decoupled from its intended purpose); technology labels (two to five terms capturing the technology type and key technical components); a use-case description (one sentence beginning with "To …" describing the real-world agricultural problem or application, without reference to the technical mechanism); and use-case labels (two to five terms capturing the application domain, subdomain, and target entities). The strict separation between these two pairs of fields is the architectural principle underlying the dual-taxonomy design: the technology dimension captures how innovations work, while the use-case dimension captures what they are for.

As mentioned above, Gemma 3 1B-IT was selected as the base model. Fine-tuning was conducted in two sequential phases using Low-Rank Adaptation (LoRA), a parameter-efficient technique that updates a small subset of model weights, applied to attention and MLP projection layers with rank $r = 32$ and scaling factor $\alpha = 64$.

In the first phase, Supervised Fine-Tuning (SFT), the model was trained on approximately 25,000 annotated patent examples. Each example pairs a patent text with a validated four-field XML output produced by iterative prompt-based extraction and manual review. The training objective was applied exclusively to output tokens, with the prompt masked, so the model learned to generate structured fields without being penalised for the input format. In the second phase, Group Relative Policy Optimisation (GRPO), the SFT checkpoint was further refined using reinforcement learning. GRPO generates multiple candidate outputs per input, scores each against a set of reward functions, and updates the model to favour higher-scoring outputs. Nine reward functions were applied, covering structural compliance (correct XML tag structure and balance), content constraints (use-case field beginning with "To", label count within the two-to-five range), and semantic quality (cross-field distinctiveness, intra-field non-repetitiveness). This two-phase strategy (progressing from format compliance to semantic quality) produced a model capable of generating structured, coherent extractions at full corpus scale.

Inference over the complete patent corpus was conducted using vLLM for GPU-accelerated batch processing. Each generated output was scored against the same reward functions used during training. Records failing any of the structural or quality thresholds were excluded from downstream clustering, ensuring that the taxonomy construction stage operates on a clean, coherent extraction set.

2.5 Semantic representation and embedding

The extracted structured fields were transformed into dense vector representations using Paecter (mpi-inno-comp/paecter), a transformer-based sentence encoder specifically developed for patent text and selected on the basis of the feasibility study results reported in Section 2.2. Two separate embedding sets were produced for each patent family: a technology embedding, encoding the technology description enriched with technology labels, and a use-case embedding, encoding the use-case description enriched with use-case labels. Labels were appended to the corresponding description before encoding to strengthen the semantic signal. All embedding vectors were L2-normalised so that cosine similarity is equivalent to the dot product throughout the clustering pipeline.

Prior to clustering, additional quality filters were applied within each FVC stratum. Records with fewer than eight words or fewer than 40 non-whitespace characters were excluded as uninformative — a threshold informed by a preliminary feasibility study of text-length effects on extraction quality, which identified the bottom five percentile of patent texts as disproportionately short and associated with a sharp

decline in output reliability. Semantic consistency between the original patent text and each of the four extracted fields was further assessed by cosine similarity; records where any individual similarity fell below 0.30, or where the minimum across all four fields fell below 0.20, were excluded as semantically incoherent.

2.6 Hierarchical taxonomy construction

The taxonomy was induced through unsupervised hierarchical clustering, applied independently to the technology and use-case embedding spaces and independently within each FVC stratum. Unlike traditional patent classification systems relying on predetermined categories, this approach discovers emergent thematic groupings directly from the semantic content of the corpus.

Clustering was performed using GPU-accelerated K-means with scalable K-means++ initialisation. The number of clusters K at each hierarchical level (meso, micro, and nano) was selected automatically by running K-means across a candidate range with three independent random seeds, evaluated on five criteria: cluster stability (mean pairwise Adjusted Rand Index across seed pairs); semantic coherence (mean Normalised Pointwise Mutual Information of top cluster-characteristic terms); term distinctiveness (Monroe et al. log-odds scoring against the rest of the corpus); a cosine-based silhouette score; and cluster size entropy, penalising highly imbalanced partitions. The optimal K was selected by rank aggregation across these criteria with deterministic tie-breaking, ensuring that selected granularity reflects a combination of stability, interpretability, and balance without manual metric weighting.

The hierarchy was constructed by nesting within each FVC stratum. Meso clusters were computed on the full within-stratum corpus; micro clusters by re-running the procedure on each meso cluster's subset; and nano clusters within each micro cluster, providing a fourth level of granularity for large or internally diverse groups. The use-case and technology taxonomies were built in parallel over their respective embedding spaces. Within each meso use-case cluster, a separate clustering procedure was additionally applied to the technology embeddings of member patents, yielding technology subclusters characterising the distinct technical approaches deployed within each application domain. The relationship between use-case meso clusters and technology subclusters is summarised as a cross-tabulation forming the multidimensional crosswalk at the heart of the dual-taxonomy framework.

Each patent was also assigned a macro technology domain (Biological, Mechanical, Material, Chemical, or Electronic/Energy) derived from two independent signals: a lookup table mapping the primary LLM-extracted technology label to a macro domain, and an IPC-derived assignment based on hierarchical mapping of formal classification codes. The LLM-derived signal serves as the primary label, reflecting the semantic content of the extracted fields. The IPC-derived signal provides independent cross-validation: agreement between the two signals supports the semantic consistency of the structured extraction; disagreement flags records for potential review.

2.7 Cluster labelling and evidence anchoring

Cluster-characteristic terms were identified using c-TF-IDF, a variant of TF-IDF adapted to the cluster setting, in which within-cluster term frequencies are normalised by cluster size and weighted by a log-ratio IDF component that rewards terms frequent in the target cluster but rare in others. Complementary salient term analysis identified terms with the highest frequency ratio between the target cluster and its nearest neighbours, supporting disambiguation of closely related clusters.

Representative patent families were selected for each cluster using a three-band sampling strategy: families closest to the centroid (core band), families from the median similarity range (middle band), and families from the lower portion of the distribution excluding the bottom five percent (breadth band), with near-duplicates removed using bigram Jaccard similarity. These representative records serve as empirical anchors, allowing domain experts to inspect concrete examples of the innovations captured by each category and verify the coherence of the automated grouping.

Cluster labels were generated at the meso level using GPT-4o-mini. As with the FVC classification step, a general-purpose commercial model is appropriate here because the task (assigning a descriptive name and rationale to a cluster defined by its top terms and representative documents) does not require the

specialised structured extraction for which Gemma was fine-tuned. The model was prompted with each cluster's top c-TF-IDF terms, salient discriminating terms relative to neighbouring clusters, and a sample of representative abstracts from all three sampling bands. It was instructed to produce purpose-oriented labels at three levels of granularity: a short title (three to six words), a medium phrase (seven to twelve words), and a paragraph-length description with rationale. An out-of-scope flag was also produced to identify clusters whose content suggests they may not genuinely belong to the agrifood domain despite passing the IPC filter; flagged clusters are subject to manual review and potential exclusion before the taxonomy is finalised. Clusters with overlapping or ambiguous short labels were detected automatically and submitted to a dedicated relabelling pass.

3. Results and discussion

3.1 Distribution of innovation across the food value chain

The taxonomy has been published as a navigation dashboard at https://datalab.review.fao.org/datalab/dashboard/oin-taxonomy/. A full analysis of the innovation patterns it reveals will be the object of a separate paper, as the present contribution focuses on the methodological framework. The following subsections provide selected initial insights, a discussion on strengths and weaknesses, and an outline of future work.

The distribution of patent families across the six FVC categories reveals strong concentration in upstream stages (see Table 1 and Figure 1). Production accounts for 45.6% of families and Processing for 37.5%; together they cover over 83% of the corpus. Cross-Cutting technologies represent 10.3%, Consumption 5.8%, while Aggregation (0.7%) and Distribution (0.1%) are markedly underrepresented.

Table 1 – Patent family counts and key indicators by Food Value Chain (FVC) stage.

FVC Stage	Families	Share (%)	Meso clusters	Patents / family
Production	930,262	45.6	20	1.29
Processing	764,276	37.5	2	1.23
Cross-Cutting	211,009	10.3	4	1.41
Consumption	119,251	5.8	2	1.21
Aggregation	13,277	0.7	4	1.22
Distribution	1,296	0.1	10	1.49
Total	2,039,371	100.0	42	1.28

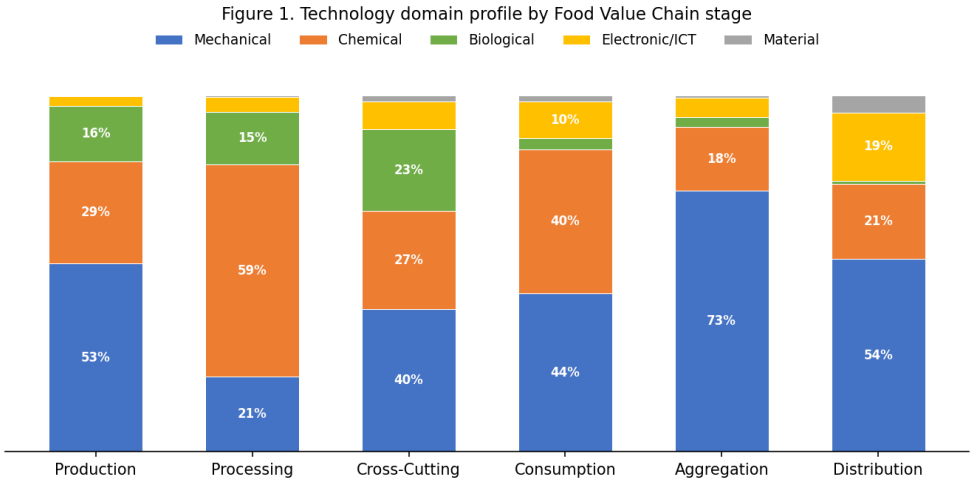

Figure 1 – Technology domain profile by FVC stage. Each bar shows the percentage share of technology assignments within that stage; because some families carry multiple technology labels, shares reflect domain intensity rather than exclusive classification.

This upstream concentration is not surprising from an economic standpoint: input technologies (seeds, fertilisers, agrochemicals, food formulations, processing machinery) are highly amenable to patent protection because they can be clearly delimited, manufactured at scale, and generate defensible commercial returns. Downstream stages rely more heavily on trade secrets, logistics know-how, and

software innovations that either fall outside the selected IPC scope or are protected through non-patent means.

The patents-per-family ratio provides a complementary signal on commercial scope. A higher ratio means each unique invention is filed in more national offices on average, implying broader perceived commercial relevance and international deployment. Cross-Cutting (1.41) and Distribution (1.49) show the highest ratios, consistent with innovations applicable across multiple value chain stages or in global supply chains being disproportionately filed across multiple jurisdictions.

The technology domain profile (Figure 1) further reveals cross-cutting patterns that IPC codes cannot express directly. Processing is dominated by Chemical technologies (59%), reflecting the centrality of formulation and preservation chemistry, while Aggregation is overwhelmingly Mechanical (73%), consistent with its focus on handling and logistics equipment. Cross-Cutting patents display a notably higher share of Biological technologies (23%), pointing to innovations whose applicability spans multiple value chain stages. These patterns (simultaneously visible across the technology and application axes) illustrate the analytical value of the dual-taxonomy design.

3.2 Strengths and limitations of the framework

The dual-taxonomy design reveals relationships that IPC codes cannot express directly. IPC is effective for retrieval and for identifying technical subject matter, but does not systematically distinguish between what a technology is and what it is used for. By decomposing patents simultaneously along a technology axis and an agricultural application axis, the framework makes visible cross-cutting patterns such as the deployment of similar technical modalities across different value-chain stages, or the convergence of multiple technical approaches around the same agricultural challenge. This makes it possible to identify not only dominant technology domains but also emerging application niches and under-developed use cases over time.

The taxonomy also moves beyond single-axis classification logic, offering a data-driven, multi-axis representation of agrifood innovation. Because clustering is performed on semantically normalised text rather than fixed labels, the resulting taxonomy can capture recurring patterns that reflect the actual structure of the patent corpus at a given moment — identifying emergent clusters not predefined in expert-built vocabularies. A further advantage is scalability: once the pipeline is established, the taxonomy can be updated regularly, maintaining comparability over time and supporting longitudinal analysis of technological change.

These strengths notwithstanding, a number of limitations hinder the application of this taxonomy to information systems, particularly within the foreseen ATIO KB context. The first concerns the patent data themselves: patent records do not provide a complete representation of agrifood innovation, due to geographic concentration in high-income patent offices, under-representation of incremental or informal innovations, coverage limited to patentable technological innovations, and linguistic limitations. A key limitation lies in validation and model dependence: no single external gold standard exists against which clusters can be fully validated, different runs of the routine produce different results, establishing hierarchical relationships is challenging given multiple possible grouping dimensions, and granularity is inconsistent due to the varying number of dimensions used to cluster. These limitations suggest that the taxonomy should be treated as a feeding mechanism to a more consistent semantic backbone rather than as a definitive representation of agrifood innovation.

3.3 Future directions

Several directions for further development are envisaged. First, annual updates are planned through automated ingestion of new PATSTAT editions, ensuring the framework evolves with the innovation landscape. Second, the methodology is designed to be extended beyond patents — to scientific publications, project databases, and grassroots innovation repositories — broadening coverage to innovations not captured by the patent system.

Work has also begun exploring whether faceted classification theory could help address issues of hierarchy, granularity, and consistency by clarifying the feature dimensions used for grouping (e.g., Output/Noun: general-purpose type; Method; Approach; Use: use case; Benefit/impact; Context; Object — inspired and adapted from Ranganathan's PMEST dimensions [45] [46]). These dimensions are proposed as a starting hypothesis for discussion, not as a fixed or finalised framework, and are being examined through the lens of existing product family design methodological frameworks. This approach would complement the category-based taxonomies and the use of post-coordinated searches (e.g. on combinations of types and use cases) with a new precoordinated feature-based typology: innovation families. The FAO Data Lab and the FAO Office of Innovation will continue collaborating on building a comprehensive typology of agrifood technologies to enable insights and trend analyses.

References

[1] Raouhi, E.M., Lachgar, M., Hrimech, H., & Kartit, A. (2023). Unmanned Aerial Vehicle-based Applications in Smart Farming: A Systematic Review. International Journal of Advanced Computer Science and Applications.

[2] Elbasi, E., Mostafa, N., AlArnaout, Z., Zreikat, A.I., Cina, E., Varghese, G.H., Shdefat, A.Y., Topcu, A.E., Abdelbaki, W., Mathew, S., & Zaki, C. (2023). Artificial Intelligence Technology in the Agricultural Sector: A Systematic Literature Review. IEEE Access, 11, 171-202.

[3] Ofosu-Ampong, K., Abera, W., Mesfin, T., & Abate, T. (2025). Digital agro-advisory tools in the global south: a behavioural analysis of impacts, and future directions. Discover Agriculture, 3.

[4] Carlino, G.A., & Kerr, W.R. (2014). Agglomeration and Innovation. Handbook of Regional and Urban Economics, 5, 349-404.

[5] Dechezleprêtre, A., Ménière, Y., & Mohnen, M. (2017). International patent families: from application strategies to statistical indicators. Scientometrics, 111, 793 - 828.

[6] Zhang, N., Sun, C., Xu, M., Wang, X., & Deng, J. (2023). Catching up of Latecomer Economies in ICT for Sustainable Development: An Analysis Based on Technology Life Cycle Using Patent Data. Sustainability.

[7] Zhang, S., Yuan, C., & Wang, Y. (2019). The Impact of Industry–University–Research Alliance Portfolio Diversity on Firm Innovation: Evidence from Chinese Manufacturing Firms. Sustainability.

[8] Khan, N.A., Qu, H., Qu, J., Wei, C., & Wang, S. (2021). Does Venture Capital Investment Spur Innovation? A Cross-Countries Analysis. SAGE Open, 11.

[9] Shao, H., Jin, Q., Guo, Y., Zhou, F., Wider, W., & Lu, L. (2025). The relationship between the structure of firms' human capital and corporate innovation performance. PLOS One, 20.

[10] Garsous, G., & Worack, S. (2021). Trade as a channel for environmental technologies diffusion. OECD Trade and Environment Working Papers.

[11] Petruzzelli, A.M., Rotolo, D., & Albino, V. (2014). Determinants of Patent Citations in Biotechnology: An Analysis of Patent Influence Across the Industrial and Organizational Boundaries. ArXiv, abs/1403.2096.

[12] Xue, X., Tan, X., Huang, Q., Zhu, H., & Chen, J. (2022). Exploring the Innovation Path of the Digital Construction Industry Using Mixed Methods. Buildings.

[13] Guevara-Ramírez, W., Martínez-de-Alegría, I., Río-Belver, R.M., & Alvarez-Meaza, I. (2022). Strategic management of patents on electrochemical conversion fuel cells and batteries in Latin America as a mechanism for moving towards energy sustainability. Journal of Applied Electrochemistry, 53, 625 - 644.

[14] Markatou, M. (2012). Measuring 'Sustainable' Innovation in Greece: A Patent Based Analysis.

[15] Hodge, G. (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. Washington, DC: Council on Library and Information Resources.

[16] US Census Bureau, 2022

[17] Eurostat, 2024

[18] UNSD, 2024

[19] WIPO (2024). Guide to the International Patent Classification.

[20] FAO. 2025. Agrifood Systems Technologies and Innovations Outlook (ATIO): The co-design of the ATIO Knowledge Base report of the consultation process and outcomes.

[21] FAO. 2025b. Agrifood Systems Technologies and Innovations Outlook Knowledge Base (ATIO KB): Concept Note.

[22] Li, H., Zhao, Y., Li, Y., & Wang, Y. (2024). Turning inward in difficulties: R&D human resource slack, technological diversification, and independent innovation. PLOS ONE, 19.

[23] Bertoni, D., Cavicchioli, D., Donzelli, F., Ferrazzi, G., Frisio, D.G., Pretolani, R., Ricci, E.C., & Ventura, V. (2018). Recent Contributions of Agricultural Economics Research in the Field of Sustainable Development. Agriculture.

[24] Khomenko, E.V., & Rakhvalova, D. (2021). Intellectual property commercialization as a factor of sustainable development of regional agricultural enterprises. E3S Web of Conferences.

[25] Caprarulo, V., Ventura, V., Amatucci, A., Ferronato, G., & Gilioli, G. (2022). Innovations for Reducing Methane Emissions in Livestock toward a Sustainable System: Analysis of Feed Additive Patents in Ruminants. Animals : an Open Access Journal from MDPI, 12.

[26] Cheng, L., Zhang, S., Lou, X., Huang, J., Rao, F., & Bai, R. (2021). How Does Tie Strength Dispersion within Inter-Organizational Networks Affect Agricultural Technological Innovation? Evidence from China. Land.

[27] Wang, H., Wang, Q., Xiao, Y., Chen, H., Su, Z., & Xiang, C. (2024). Collaborative Network, Technological Progress and Potato Production in China. Potato Research, 68, 1331 - 1353.

[28] Trappey, A.J., Lin, J.G., Chen, K., & Chen, M.M. (2023). Global Patent Technology Portfolio Study of Agricultural UAV Innovations. TE.

[29] Thavorn, J., Muangsin, V., Gowanit, C., & Muangsin, N. (2021). A Scientometric Assessment of Agri-Food Technology for Research Activity and Productivity. Publ., 9, 57.

[30] Rustenova, E., Ibyzhanova, A., Aidaraliyeva, A., Barykin, S., Rudenko, L., Mottaeva, A.B., & Voronova, O. (2026). Digital and Energy Transitions in the Agro-Industry: An Economic Analysis of Technology Diffusion and Sustainable Innovation Pathways. International Journal of Energy Economics and Policy.

[31] Yoo, Y., Zhang, X., & Cao, L. (2025). Self-Filtered Distillation with LLMs-generated Trust Indicators for Reliable Patent Classification. ArXiv, abs/2510.05431.

[32] Jiang, L., & Goetz, S. (2024). Natural language processing in the patent domain: a survey. Artificial Intelligence Review, 58.

[33] Bergeaud, A., Potiron, Y., & Raimbault, J. (2016). Classifying patents based on their semantic content. PLoS ONE, 12.

[34] Hu, J., Li, S., Yao, Y., Yu, L., Yang, G., & Hu, J. (2018). Patent Keyword Extraction Algorithm Based on Distributed Representation for Patent Classification. Entropy, 20.

[35] Garechana, G., Río-Belver, R.M., Bildosola, I., & Cilleruelo-Carrasco, E. (2019). A method for the detection and characterization of technology fronts: Analysis of the dynamics of technological change in 3D printing technology. PLoS ONE, 14.

[36] Yun, J., & Geum, Y. (2020). Automated classification of patents: A topic modeling approach. Comput. Ind. Eng., 147, 106636.

[37] Grzeszczyk, T.A., & Grzeszczyk, M.K. (2021). Improving the Discovery of Technological Opportunities Using Patent Classification Based on Explainable Neural Networks. European Research Studies Journal.

[38] Hain, D.S., Jurowetzki, R., Buchmann, T., & Wolf, P. (2020). A text-embedding-based approach to measuring patent-to-patent technological similarity. Technological Forecasting and Social Change.

[39] Risch, J., & Krestel, R. (2019). Domain-specific word embeddings for patent classification. Data Technol. Appl., 53, 108-122.

[40] Bekamiri, H., Hain, D.S., & Jurowetzki, R. (2021). PatentSBERTa: A deep NLP based hybrid model for patent distance and classification using augmented SBERT. Technological Forecasting and Social Change.

[41] Bekamiri, H., Hain, D.S., & Jurowetzki, R. (2022). A Survey on Sentence Embedding Models Performance for Patent Analysis. ArXiv, abs/2206.02690.

[42] Lee, J., & Hsiang, J. (2020). Patent classification by fine-tuning BERT language model. World Patent Information.

[43] Emer, L., Lippi, M., Mina, A., & Vandin, A. (2026). Large Language Models for Patent Classification: Strengths, Trade-offs, and the Long Tail Effect. ArXiv, abs/2601.23200.

[44] Comb, M., & Martin, A. (2024). Mining digital identity insights: patent analysis using NLP. EURASIP Journal on Information Security, 2024.

[45] Ferreira, A. C., Maculan, B. C. M. S., & Naves, M. M. L. (2017). Ranganathan and the faceted classification theory. Transinformação, 29(3), 279–295.

[46] Broughton, V. (2023). Facet Analysis: The Evolution of an Idea. Cataloging & Classification Quarterly, 61(5–6), 411–438.