Agris on-line Papers in Economics and Informatics

Volume XVIII Number 1, 2026

# Standardized Data Infrastructures for Plant Phenomics: A Review of MIAPPE and BrAPI Integration within High-Performance

Vladimír Voral, Michael Anderle, Michal Stočes, Pavel Ambruz, Pavel Šimek, Jan Jarolímek, Jiří Vaněk

Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

## Abstract

The increasing complexity and volume of plant phenotypic data have driven the emergence of new computational and standardization frameworks to enable data integration, reproducibility, and reuse. This systematic literature review examines the current state of software tools, data models, and interoperability standards in plant phenomics, focusing on the implementation of the FAIR (Findable, Accessible, Interoperable, Reusable) principles. Using a structured PRISMA-based methodology, we analyze two major community driven initiatives MIAPPE and BrAPI as representative solutions for standardized data description and exchange. Furthermore, the study evaluates the role of High-Performance Computing (HPC) and deep learning in addressing computational challenges associated with large-scale datasets, including multi-sensor and 3D capture technologies. Special consideration is given to data governance, encompassing secure access, ethical use, and GDPR compliance within expanding phenomics ecosystems. The synthesis identifies persistent gaps in data harmonization and semantic alignment, proposing future research directions toward more integrated, secure, and scalable infrastructures. This review emphasizes that the success of plant phenomics depends on bridging the gap between standard definitions and their practical implementation within high-performance workflows.

## Keywords

MIAPPE, BrAPI, plant phenomics, high-performance computing (HPC), FAIR principles, data management, ontology, data governance, interoperability, open science.

Voral, V., Anderle, M., Stočes, M., Ambruz, P., Šimek, P., Jarolímek, J. and Vaněk, J. (2026) "Standardized Data Infrastructures for Plant Phenomics: A Review of MIAPPE and BrAPI Integration within High-Performance Computing Frameworks", *AGRIS on-line Papers in Economics and Informatics*, Vol. 18, No. 1, pp. 111-131. ISSN 1804-1930. DOI 10.7160/aol.2026.180109.

## Introduction

The proliferation of digital technologies and high-throughput data acquisition platforms has transformed plant science into a data intensive discipline. Modern phenomics, defined as the systematic study of plant traits through the integration of imaging, sensor technologies, and computational analytics, plays a crucial role in understanding genotype environment interactions. As the scale and complexity of data grows, ensuring that phenomics datasets are interoperable, reproducible, and reusable has become a central challenge for the research community.

A significant barrier to progress in this domain lies in the lack of standardized data models and exchange protocols. Phenotypic data are often stored in heterogeneous formats, annotated inconsistently, and distributed across numerous databases. This fragmentation hinders comparative analyses and limits the potential of integrative approaches that link phenotypic, genomic, and environmental data. To address these issues, international initiatives such as MIAPPE and BrAPI have been established to promote standardization and interoperability in phenomics research.

The MIAPPE standard defines a structured metadata schema to describe plant phenotyping experiments, covering elements such as experimental design, plant material, environmental conditions, and observed traits (Papoutsoglou et al., 2020). The accompanying data model enables harmonization and improves data discoverability across platforms. Complementarily, the BrAPI framework provides a web service specification that facilitates programmatic communication between breeding and phenotyping databases (Shelby

et al., 2019). Together, MIAPPE and BrAPI serve as key enablers of FAIR data management, enhancing transparency, re-producibility, and collaborative data sharing across institutions.

Alongside these standardization efforts, advances in computational infrastructure notably High-Performance Computing (HPC) and machine learning have significantly expanded analytical capacities in phenomics. HPC enables large-scale processing of multispectral images, time-series data, and multi-omics datasets, thereby supporting complex modeling and predictive breeding applications (Wallace et al., 2018; Abdul Hamid & Singh, 2024). These computational approaches not only accelerate scientific discovery but also contribute to the development of decision-support tools in precision agriculture and crop improvement.

As phenomics data ecosystems evolve, data governance and security emerge as equally critical dimensions. Responsible data management requires embedding ethical, legal, and technical safeguards such as role-based access control, data anonymization, and GDPR compliance into data infrastructures (Umbach, 2024). Ensuring security-by-design and adherence to open science principles strengthens public trust, encourages collaboration, and enhances the societal impact of research (Wilkinson et al., 2016, Papoutsoglou et al., 2020).

This review provides a systematic synthesis of current literature on software tools, data standards, and computational infrastructures in plant phenomics. It identifies key technological developments, evaluates the degree of interoperability achieved, and highlights emerging challenges related to scalability, ethical stewardship, and cross-domain data integration.

## Breeding APIs and the evolution of BrAPI standards

In recent times, there has been a notable higher volume in the scale of datasets, leading to their dispersal across multiple platforms. As the demand for analyses requiring the interoperability of data from various origins rises, bridging the gap between disparate systems poses a considerable challenge. In addressing this challenge, was introduced the public plant Breeding Application Programming Interface (BrAPI), was introduced to enhance interoperability among breeding applications. BrAPI stands as a standardized web service API specification, conceived through collaborative efforts within a global community. The emergence of BrAPI enabled applications, known as BrAPPs, demonstrates the practical utility of this standard in aggregating datasets from diverse sources. This is further complemented by regional initiatives that integrate high-throughput measurements into high-performance computing infrastructures for agricultural research (Stočes et al., 2025). Acknowledged as a pivotal technology for various significant breeding system initiatives, the development of such a standard has gained recognition. The inaugural version of the API primarily focuses on furnishing services for system integration and the retrieval of fundamental breeding data encompassing germplasm, study, observation, and marker data. Several applications leveraging BrAPI compatibility, termed BrAPPs, have already been developed, capitalizing on the increasing adoption of BrAPI across numerous databases (Selby et al., 2019).

The introduction of BrAPI v2 marks a significant evolution of the standard, transitioning to a modular architecture that includes Core, Phenotyping, Genotyping, and Germplasm modules. This update enables more robust real-world applications for data integration and facilitates collaboration across diverse breeding and genetics communities, effectively addressing the challenges of managing large-scale, fragmented datasets in modern plant science (Selby et al., 2025). Specifically, for phenomics, BrAPI v2 provides enhanced support for high-throughput imaging data and complex phenotypic observations, allowing for more efficient programmatic exchange between large-scale phenotyping platforms and downstream analytical tools.

## High-Performance Computing frameworks in phenomics

With the increasing complexity of data structures, the growing volume of multispectral imagery, time-series data, and integrated omics layers (e.g., the combination of phenomics and genomic data), the computational demands of phenomics research have risen substantially. In this context, High-Performance Computing (HPC) plays a crucial role. HPC provides the infrastructure necessary to process massive datasets in parallel, train large-scale machine learning models, and perform simulations that would be unfeasible or extremely time-consuming on conventional computing systems (Sterling et al., 2024, Choudhary at al., 2020).

HPC enables researchers to perform tasks more efficiently, such as automated seg-mentation of plant images, extraction of phenotypic traits, or yield prediction based on field-derived data. To ensure the FAIRness of HPC workflows, the use of standardized software containers and distributions like Bioconda is essential for reproducibility across different high-performance infrastructures (Georgiou et al., 2020, Grüning et al., 2018). For instance, the use of deep neural networks to process hyperspectral imagery facilitates accurate wheat yield estimations, an undertaking that would be nearly impossible without access to high-performance computational resources. Reflecting on the first decade of this development, deep learning has emerged as the dominant technique in plant phenotyping (Kartal et al., 2025), shifting the focus from traditional image analysis to high-capacity pipelines that extract high-level agronomic phenotypes (Ubbens et al., 2025). This evolution underscores the critical role of HPC in handling the "explosion" of multi-modal data from diverse sensors and 3D capture technologies (Ubbens et al., 2025).

In combination with standardized interfaces such as BrAPI, HPC not only accelerates computational workflows but also enhances the scalability and reproducibility of scientific experiments. Modern Breeding APIs are increasingly being linked to high-capacity pipelines, such as TASSEL-GBS, that leverage HPC clusters to perform genome-wide associations directly on integrated phenomics datasets (Glaubitz et al., 2014). This is critical for achieving robust, reproducible scientific outcomes and for enabling the integration of heterogeneous datasets from multiple sources.

HPC is particularly indispensable in the domain of predictive breeding, where historical and current phenotypic data are integrated with genotypic profiles to develop robust models with practical applications in agriculture and food security. The integration of genomic, phenomics, and environmental data through advanced computational techniques allows for more accurate predictions and more efficient selection of resilient and high-yielding cultivars (Yunbi et al., 2022).

Thus, the application of high-performance computing in phenomics not only accelerates scientific progress but also paves the way toward more efficient and sustainable approaches to plant production.

### Aim of the review and anticipated research gaps

1. Gap in understanding how MIAPPE guidelines, BrAPI standards, and ontologies can be effectively combined to achieve better data harmonization. There is a lack of studies investigating the effectiveness of various methods in achieving this goal. This includes the application of retrospective harmonization and the impact of data transformation strategies on the quality of harmonized data.
2. Limited research on the selection of appropriate software tools for data harmonization, despite the increasing trend towards open science and FAIR principles. This includes tools that support the use of MIAPPE, BrAPI, and ontologies.
3. Gap in integrating standardized data workflows with High-Performance Computing (HPC) infrastructure. While HPC is essential for large-scale phenomics, there is insufficient research on how to maintain data standards (MIAPPE/BrAPI) throughout the automated processing and analysis in HPC environments.
4. Gap in understanding the practical challenges and opportunities in implementing these principles in real-world scenarios, although there has been progress in promoting data sharing and reusability. This includes issues related to data management, interoperability, and the use of tools like PHIS.

## Materials and methods

The methodology of this study follows the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines (Tricco et al., 2018) to ensure transparency and reproducibility. A scoping review approach was selected as particularly suited for mapping evidence in a rapidly evolving field, identifying main concepts, and uncovering knowledge gaps within a heterogeneous body of literature (Tricco et al., 2018). The primary objective was to map the current landscape of standardized data infrastructures, focusing on the integration of MIAPPE, BrAPI, and High-Performance Computing (HPC) frameworks in plant phenomics.

### Research Questions (RQ)

To guide the literature search and synthesis, we formulated three key research questions:

- RQ1: What is the current state of data sharing (interoperability) and data harmonization in plant biology?
- RQ2: Which types of data and data models are utilized for phenotypic data harmonization?
- RQ3: Do tools and standards for data management and harmonization currently exist to support large-scale phenomics?

### Search strategy and data source

The literature search was conducted across two primary databases: Web of Science (WoS) and Scopus, covering the period from 1990 to 2025. The selection was limited to original research papers and reviews with a focus on plant biology, agriculture, and in-formation technology. To ensure the quality of the synthesized evidence, only peer-reviewed articles and indexed conference proceedings were considered for inclusion. The search was performed using a combination of keywords and Boolean operators, following a strategy of sorting terms from generic to specific to capture the most relevant literature at the intersection of phenomics and data science.

### Search Queries (SQ) and analysis

The systematic search was executed using a multi-layered approach, employing eleven distinct search queries (SQ1–SQ11) designed to capture the intersection of data standards, semantic technologies, and computational infrastructure. The strategy followed a progression from broad thematic areas to highly specific technical combinations involving MIAPPE, BrAPI, and High-Performance Computing (HPC). The full search strings used for both Web of Science (WoS) and Scopus are detailed in the list below:

- SQ1: TS=("Ontology" OR "Data management" OR "Fair Principles" OR "Open science").
- SQ2: TS=("Ontology" OR "Data management" OR "FAIR principles" OR "Open science" OR "Data sharing" OR "BrAPI").
- SQ3: TS=("Ontology" OR "Data management" OR "Fair Principles" OR "Open science" OR "Data sharing" OR "BrAPI" OR "Miappe" OR "Phenomics Databases" OR "Phenomics Platforms" OR "PHIS").
- SQ4: TS=("Ontology" OR "Data management" OR "Fair Principles" OR "Open science" OR "Data sharing" OR "BrAPI" OR "Miappe" AND "Phenomics Databases" AND "Phenomics Platforms" AND "PHIS").
- SQ5: TS=("Ontology" OR "Data management" OR "Fair Principles" OR "Open science" OR "Data sharing" AND "BrAPI" AND "Miappe" AND "Phenomics Databases" AND "Phenomics Platforms" AND "PHIS").
- SQ6: TS=("Ontology" OR "Data management" OR "Fair Principles" AND "Open science" AND "Data sharing" AND "BrAPI" AND "Miappe" AND "Phenomics Databases" AND "Phenomics Platforms" AND "PHIS").
- SQ7: TS=("Ontology" OR "Data management" AND "Fair Principles" AND "Open science" AND "Data sharing" AND "BrAPI" AND "Miappe" AND "Phenomics Databases" AND "Phenomics Platforms" AND "PHIS").
- SQ8: TS=("HPC").
- SQ9: TS=("HPC" OR "supercomputing" OR "high-performance" OR "parallel computing" OR "cluster computing").
- SQ10: TS=("HPC" OR "supercomputing" OR "high-performance" OR "parallel computing" OR "cluster computing" OR "Ontology" OR "Data management" OR "Fair Principles" OR "Open science" OR "Data sharing" OR "BrAPI" OR "Miappe" OR "Phenomics Databases" OR "Phenomics Platforms" OR "PHIS").
- SQ11: TS=("HPC" OR "supercomputing" OR "high-performance" OR "parallel computing" OR "cluster computing" AND "Ontology" AND "Data management" AND "Fair Principles").

To quantify the research landscape, we performed a bibliometric analysis of the results obtained from Web of Science (WoS) and Scopus. This analysis focused on the volume of reviews versus original research articles, publication trends over time, and the impact of the top cited publications within each query. The specific parameters and the resulting document count for each query are summarized in Table 1.

| Search ID | Reviews (WoS) | Original Article (WoS) | Peak Publication Year | WoS (Fielded Search) | WoS (Query Builder) | Scopus (All Fields) |
|---|---|---|---|---|---|---|
| SQ1 | 5,629 | 109,885 | 2022: 12,263 | 3,465,450 | 163,794 | 2,762,160 |
| SQ2 | 6,995 | 121,924 | 2024: 14,596 | 3,794,765 | 183,749 | 255,093 |
| SQ3 | 7,082 | 123,043 | 2024: 14,701 | 3,952,857 | 185,134 | 25,705 |
| SQ4 | 6,991 | 121,828 | 2024: 14,586 | 3,794,765 | 183,634 | 7 |
| SQ5 | 5,629 | 109,885 | 2022: 12,263 | 3,465,450 | 163,794 | 2 |
| SQ6 | 4,032 | 104,956 | 2022: 11,225 | 1,942,388 | 155,498 | 2 |
| SQ7 | 2,332 | 86,560 | 2022: 9,015 | 133,883 | 121,931 | 2 |
| SQ8 | 459 | 14,003 | 2024: 1,781 | 88,830 | 24,383 | 108,478 |
| SQ9 | 27,187 | 532,548 | 2024: 53,682 | 1,399,556 | 671,897 | 1,520,993 |
| SQ10 | 34,253 | 654,883 | 2024: 68,223 | 5,225,105 | 855,486 | 7,545 |
| SQ11 | 27,179 | 531,948 | 2024: 53,657 | 754,312 | 670,578 | 690 |

Source: Own processing

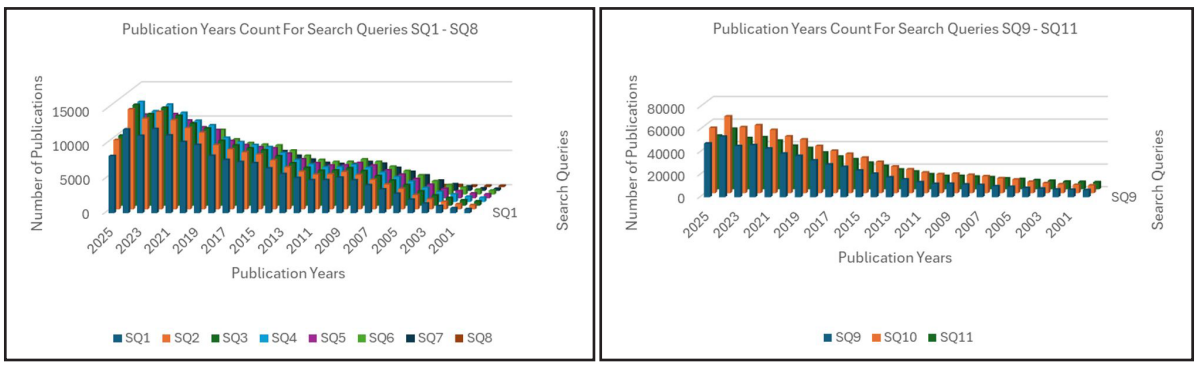
Table 1: Summary of systematic search queries and document retrieval counts across Web of Science (WoS) and Scopus databases.

The quantitative evolution of the research landscape was visualized by analyzing publication frequency over time for each search query. Due to the significant variance in retrieval volume where broad thematic searches yielded hundreds of thousands of records compared to hundreds in specialized queries the longitudinal analysis was divided into two comparative visualizations focusing on the peak development period from 2000 to 2025.

Figure 1(a) illustrates the annual publication trends for queries SQ1 through SQ8, representing the foundational growth of data stewardship, FAIR principles, and basic HPC concepts. Figure 1(b) focuses on the high-volume datasets retrieved in SQ9 through SQ11, reflecting the massive scale of research in supercomputing and parallel architecture. Both figures demonstrate a sharp exponential increase in scientific output, particularly following the global push for data interoperability after 2010 and the formalization of the FAIR principles in 2016.

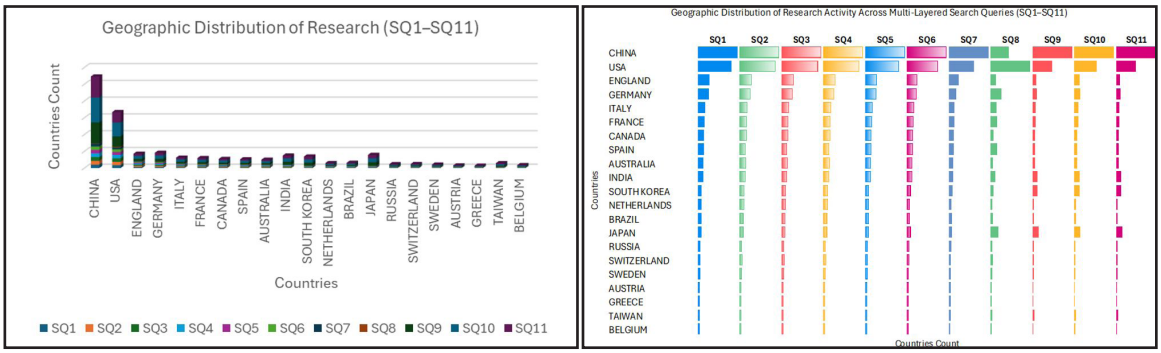
In addition to temporal trends, the geographic distribution of research activity was assessed to identify global hubs of expertise. This spatial analysis provides insights into the regional prioritization of high-performance computing (HPC) infrastructures and their application in data-driven life sciences.

Figure 2 presents a synthesized overview of the global contribution across all search queries. Figure 2(a) utilizes a comparative matrix format (heat map) to visualize the publication density for the top 21 countries, highlighting how research leadership scales with complexity. Complementary to this, Figure 2(b) employs a stacked bar chart to illustrate the cumulative research output of these nations across the entire query spectrum (SQ1–SQ11). This visualization allows for a direct




Source: Own processing

Figure 1: Comparative analysis of annual publication trends (2000–2025) derived from Web of Science: (a-left) Longitudinal growth of foundational topics represented by search queries SQ1–SQ8; (b-right) Evolution of high-volume research areas associated with supercomputing and large-scale data integration in SQ9–SQ11.




Source: Own processing

Figure 2: Geographic distribution of research activity across search queries (SQ1–SQ11): (a-left) Comparative matrix (heat map) of publication density for the top 21 countries; (b-right) Stacked bar chart illustrating the cumulative research output by country, categorized by search queries SQ1 through SQ11, highlighting the total volume and internal distribution of thematic focus.

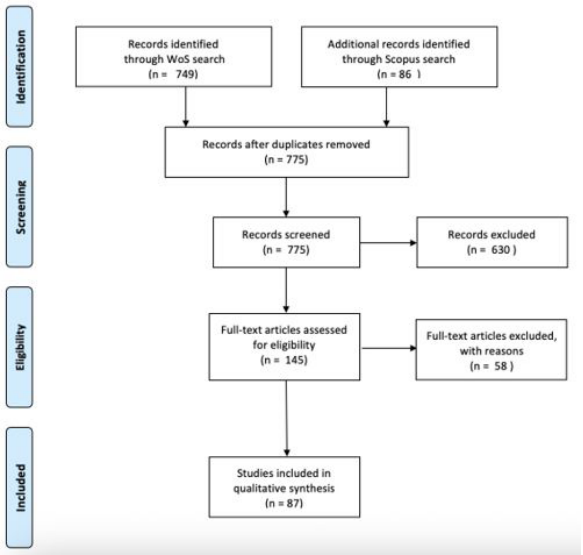
comparison of total volume while maintaining the visibility of individual query contributions per country.

To ensure full transparency while maintaining conciseness in the main text, the complete set of bibliometric analyses for all eleven search queries including individual annual growth charts and detailed geographic distribution maps for the top 10 countries is pro-vided in Appendix A (Figures A1–A22).

### Study selection (PRISMA flowchart)

The systematic screening process followed the multi-stage approach defined by the PRISMA-ScR guidelines. The initial identification phase across WoS and Scopus yielded 835 records. Following the removal of duplicates, 775 unique records were screened based on titles and abstracts, resulting in the exclusion of 630 articles that did not meet thematic focus on plant phenomics or standardized metadata.

The remaining 145 articles underwent full-text assessment for eligibility. Of these, 58 were excluded due to a lack of technical detail regarding infrastructure implementation or insufficient alignment with MIAPPE/BrAPI standards. Ultimately, 87 studies were selected for qualitative synthesis and inclusion in this review. The complete selection workflow is visualized in the PRISMA flow diagram (Figure 3).




Source: Own processing

Figure 3: PRISMA flow diagram of the study selection process, illustrating the stages of identification, screening, eligibility, and final inclusion of the 87 selected studies focusing on plant phenomics data standards and HPC integration.

## Results and discussion

This section synthesizes the findings from the literature review, categorizing the current state of the art into the data paradigm, security frameworks, phenotyping infrastructures, and semantic standards.

The development of phenomics marks the transition from individual trait observation to large-scale, technology assisted quantification of phenotypes. Phenomics can be defined as the systematic acquisition and analysis of phenotypic traits using automated sensor networks, imaging systems, robotics, and computational tools (Tardieu et al., 2017).

The field emerged as a distinct area of inquiry around 2010, stimulated by technological innovations introduced by companies such as Lemna-Tec and by collaborative initiatives including the International Plant Phenotyping Network (IPPN), which has since sought to establish phenomics as a fully recognized scientific discipline (Poorter et al., 2023).

Phenomics extends the classical concept of plant phenotyping by integrating multi-dimensional data streams describing plant structure, function, and temporal dynamics under diverse environmental conditions. These datasets enable the systematic linking of genotype × environment interactions to plant performance, thereby bridging molecular biology, physiology, and ecology within a unified experimental and computational framework (Tardieu et al., 2017).

### The data paradigm: from acquisition to wisdom

Much of the recent progress in science and society is fundamentally driven by advances in data centered disciplines and the development of technologies that facilitate large-scale data acquisition, processing, and interpretation. The exponential growth of data collection technologies, such as high throughput sensors, Internet of Things (IoT) devices, imaging systems, and automated data loggers, has drastically transformed the landscape of research and innovation (Nicore et al., 2020, Wilkinson et al., 2016). Simultaneously, advances in data analytics and computational infrastructure, particularly High-Performance Computing (HPC) and cloud-based systems, have enabled the efficient processing of increasingly complex datasets. These technological shifts underpin the broader movement toward data intensive science, in which data are not merely by products of experimentation but central assets that drive discovery.

### Evolution of the data paradigm

Over the past two decades, the philosophy of data management has evolved through several distinct paradigms. The Open Data movement, emerging in the early 2010s, advocated for unrestricted access to scientific data, encouraging researchers to make their data publicly available to promote transparency and reuse. However, as data volumes and heterogeneity grew, it became clear that openness alone was insufficient to ensure meaningful reuse and interoperability.

This realization led to the formulation of the FAIR Guiding Principles (Findable, Accessible, Interoperable, Reusable) in 2016, which provided a structured framework for responsible data stewardship (Wilkinson et al., 2016). The FAIR paradigm emphasized machine readability, persistent identifiers, and standardized metadata essential elements for ensuring that data can be discovered, understood, and reused by both humans and computational systems. In parallel, the introduction of the General Data Protection Regulation (GDPR) in 2018 established a legal framework for the ethical handling of personal and sensitive data, further reinforcing accountability and privacy in data-driven research.

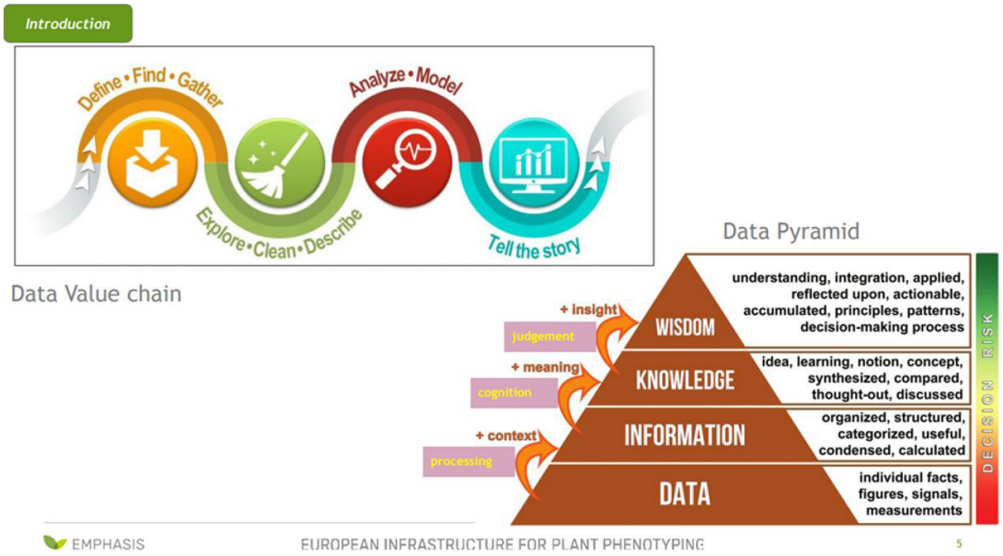
More recently, the concept of Digital Responsibility Goals (DRGs) has been proposed as a next stage in this evolution (Meier et al., 2021). The DRG framework extends beyond FAIR-ness and compliance by emphasizing digital literacy, algorithmic transparency, and trustworthy data use. It recognizes that for data to deliver societal value, users must not only have access to quality data but also possess the competencies and ethical awareness required to use it responsibly. In this view, the emphasis shifts from mere data availability to data quality, interpretability, and societal impact.

### From data to wisdom: the knowledge hierarchy

This progression aligns with the widely accepted DIKW (Data Information Knowledge Wisdom) hierarchy, which conceptualizes the transformation of raw data into actionable insight (Ackoff, 1989). The transformation process from raw phenotypic observations to actionable wisdom is illustrated in (Figure 4). Within this framework:

- Data represents raw observations or measurements.
- Information emerges when data is structured and contextualized.
- Knowledge is derived through interpretation, linking information to understanding.
- Wisdom reflects the capacity to apply knowledge judiciously for decision-making and societal benefit.

In the context of phenomics, this paradigm is particularly relevant. Phenomics re-search produces vast amounts of heterogeneous data ranging from sensor readings and imaging data to environmental and genetic metadata. Transforming these diverse datasets into knowledge requires robust data management pipelines, interoperability standards (e.g., MIAPPE and BrAPI), and computational models




Source: Based on the Data–Information–Knowledge–Wisdom (DIKW) model (Ackoff, 1989; Rowley, 2007)

Figure 4: The DIKW hierarchy represents the transformation from raw phenotypic data to actionable wisdom.

capable of extracting biologically meaningful patterns. Ultimately, achieving "data-to-wisdom" in phenomics de-pends not only on data collection but also on the responsible, transparent, and collaborative use of data across the research ecosystem.

### Data and end-users in phenomics

The primary users of phenomics data include plant biologists, geneticists, breeders, agronomists, and data scientists. Their needs vary from fundamental understanding of plant environment interactions to applied goals such as yield prediction, stress tolerance improvement, and resource efficient cultivation. For these stakeholders, standardized, high-quality data is essential to ensure comparability and reproducibility across experiments, institutions, and technological platforms.

Thus, the data paradigm in phenomics extends beyond scientific methodology: it encompasses ethical, technical, and educational dimensions that collectively determine the field's ability to generate not only data and information but actionable knowledge and wisdom for sustainable agriculture and food security.

Phenotyping data serves a dual purpose within the life sciences. From a fundamental research perspective, they provide an empirical basis for understanding how genetic variation translates into phenotypic diversity and environmental adaptability. High-resolution phenotypic datasets support the study of plant responses to abiotic and biotic stresses, photosynthetic efficiency, and biomass accumulation, contributing to predictive modeling of plant performance (Fiorani et al., 2013).

From an applied perspective, phenomics data underpin agricultural innovation and crop improvement. They facilitate data driven breeding strategies by enabling the identification and selection of genotypes with desirable agronomic traits such as enhanced drought tolerance, improved nutrient use efficiency, and increased yield stability. The integration of phenomics datasets into breeding pipelines accelerates the evaluation of genotype × environment × management interactions and informs decision making in precision agriculture and resource optimization (Ghanem et al., 2015). Beyond breeding, phenomics analyses are also instrumental in agronomy, agrochemical testing, crop nutrition research, and post-harvest quality characterization.

### Data security and digital ethics in phenomics

Within the current data paradigm, it is essential to consider data security and digital ethics. As the volume and complexity of phenomics data increase, so do the risks of misuse, unauthorized access, data manipulation, or loss of integrity. Modern data management therefore encompasses not only standards such as FAIR, but also security-by-design principles, which assume the integration of security mechanisms directly into the design of data systems (Arend et al., 2022, Aksenova et al., 2024). Key tools include

role-based access control (RBAC) based on user roles and permissions, authentication and authorization protocols such as OAuth2 or OpenID Connect, data transmission encryption (TLS/SSL), and anonymization or pseudonymization of sensitive records (UK Data Service, 2025). In the context of phenomics, it is also crucial to implement audit logs and system monitoring, which enable retrospective tracking of data access and manipulation (Aksenova et al., 2024). These principles are particularly critical when data are related to proprietary breeding projects or hosted on cloud platforms with open access policies (Minssen et al., 2020).

From an ethical and legal perspective, it is now essential to consider European and international data protection frameworks, particularly the General Data Protection Regulation (GDPR) (European Commission, 2025, GDPR-info.eu, 2024, European Union, 2018). Although phenomics data is not inherently personal, their combination with metadata (e.g., geolocation, timestamps, or sensor data linked to specific operators) can lead to the identification of individuals or locations, thereby triggering regulatory obligations (OECD, 2022, Wong et al., 2021). Research infrastructures must therefore ensure not only technical data security, but also clear rules for access control, data retention, classification of data categories, and the management of consent for data use (OECD, 2022, Wong et al., 2021). Responsible data stewardship in this context not only protects the rights of users and research participants but also fosters trust among stakeholders within open and federated data ecosystems, where the trustworthiness of data is as crucial as its technical quality and interoperability (Prifti et al., 2023).

## Phenotyping infrastructure, repositories, and data types

Phenotyping represents a fundamental component of contemporary biological and agricultural research, providing quantitative descriptions of an organism's observable characteristics that arise from the interaction between its genotype and the environment. The concept of phenotyping originated in the 1960s, initially referring to the morphological and physiological characterization of organisms through direct observation and measurement. With the advent of advanced sensing technologies, imaging, and automation, phenotyping has evolved into a high-throughput, data-driven scientific activity that enables systematic and reproducible quantification of complex biological traits (Fiorani et al., 2013).

## High-Throughput platforms and HPC integration

The lack of common standards for describing phenotypic data has long hindered efficient data exchange and reuse, leading to the formulation of the MIAPPE (Minimum In-formation About a Plant Phenotyping Experiment) recommendations. This initiative de-fines the necessary scope of metadata and utilizes the ISA-Tab format for the practical organization of experimental data, thereby directly supporting the replicability and comparability of results (Ćwiek-Kupczyńska et al., 2016).

Ground-based robotic phenotyping platforms have emerged as a high-throughput solution for capturing multi-dimensional plant traits, yet they generate massive volumes of sensor data that challenge traditional processing capabilities. The integration of these robotic systems with High-Performance Computing (HPC) is essential for real-time data analytics, image reconstruction, and the management of heterogeneous datasets in field conditions (Rui and Changying, 2022; Yuan et al., 2023).;

## Taxonomy of phenotyping data and 3D analysis

Phenotyping data encompasses a wide spectrum of measurements that collectively describe plant form, function, and response to the environment. Common categories of data include morphological, physiological, biochemical, environmental, and temporal data, each contributing distinct insights into plant behavior (Bosilj et al., 2018).

- Morphological data describe physical attributes such as plant size, shape, leaf area, color, and architecture, typically derived from 2D or 3D imaging systems, laser scanning, or structured light sensors.
- Physiological data capture biological processes such as transpiration, photosynthesis, chlorophyll fluorescence, or stomatal conductance, often obtained through hyperspectral or thermal imaging.
- Biochemical and molecular data include metabolite, protein, and enzyme profiles obtained from analytical platforms such as mass spectrometry or NMR, linking molecular variation to observed traits.

- Environmental data document growth conditions including light, temperature, humidity, and soil composition, which are crucial for interpreting genotype environment interactions.
- Temporal data represent longitudinal measurements that track develop-mental or stress related changes in plants across time.

Recent developments in 3D imaging and point cloud analysis have expanded the ability to capture complex plant geometries and canopy architectures. Such techniques have reached advanced levels in disciplines such as geospatial science, robotics, and industrial computer vision, where 3D object reconstruction, LiDAR-based mapping, and AI-driven image segmentation are standard. These approaches are increasingly being adapted for biological research, offering transferable solutions for data storage formats, spatial annotation, and metadata integration that could accelerate progress in plant phenotyping.

## Data models and repository infrastructures

The increasing complexity of phenotypic datasets has stimulated efforts to formalize data models capable of ensuring structural consistency, semantic clarity, and interoperability. Traditional relational databases have proven insufficient for handling multimodal and temporal data typical of phenotyping workflows. Consequently, hybrid data architecture combining object relational, NoSQL, and graph-based models have emerged as more suitable options, supporting flexible schema definitions and efficient storage of large-scale sensors and imaging data.

Several open data repositories and infrastructure projects support phenotyping data storage and sharing, including PHIS (Plant Phenomics Information System), e!DAL-PGP (Plant Genomics and Phenomics Research Data Repository), and institutional databases integrated into ELIXIR or EMPHASIS networks. These repositories facilitate data deposition, versioning, and long-term accessibility in accordance with FAIR principles (Findable, Accessible, Interoperable, Reusable). The PGP (Plant Genomics and Phenomics Research Data Repository) serves as a key infrastructure for the publication of complex phenomics and genomic data, utilizing the e!DAL software framework. This system enables the assignment of persistent identifiers (DOIs) and automates the data submission process, thereby fulfilling FAIR principles even for voluminous datasets, such as phenotyping im-age collections or mass spectrometry data (Arend et al., 2016).

## Case study: The TERRA-REF project and HPC

The TERRA-REF project demonstrates the necessity of high-performance computing (HPC) environments for processing open access reference datasets generated by high-resolution field scanners. This infrastructure supports the integration of hyperspectral imagery, 3D structures, and environmental measurements, providing an open-source computational pipeline that enables the calibration of algorithms for extracting plot-level phenotypes from massive datasets (Xu et al., 2022).

Phenomics inherently produces high-dimensional and heterogeneous data, and the processing of such data requires advanced information and communication technologies (ICT) and HPC to ensure reproducibility (Tardieu et al., 2017).

Despite rapid technological advancement, phenotypic traits remain inherently multifactorial, which complicates the development of robust models and remains a key obstacle to data comparability (Tardieu et al., 2017, Poorter et al., 2023).

## Semantic standards: harmonization through ontologies

The field of plant phenotyping has rapidly evolved over the past two decades, driven by the increasing integration of sensor technologies, automation, and computational analysis. This development has led to the generation of diverse and complex datasets that capture structural, functional, and environmental aspects of plant performance. However, this heterogeneity also creates major challenges in data modeling, storage, harmonization, and interoperability across platforms and research infrastructures.

Given the diversity of experimental setups and data formats, data harmonization has become a critical focus area. The processing and interpretation of high-dimensional and heterogeneous data typically comprising multispectral imagery, 3D structural reconstructions, and sensor based temporal measurements require standardized data management frameworks capable of ensuring data interoperability and reproducibility. The establishment of community standards reflects ongoing efforts to promote harmonization and facilitate data exchange across phenotyping platforms and research infrastructures.

### Metadata frameworks and MIAPPE

The most widely adopted community driven standard is MIAPPE (Minimum Information About a Plant Phenotyping Experiment), which defines a structured metadata schema describing experimental context, environmental conditions, biological material, and measured traits. MIAPPE is maintained as an open-source specification with its documentation and schema hosted on GitHub (MIAPPE Contributors, 2024). This standard is open community server accessible on GitHub, where there are available latest changes and documentation (MIAPPE Contributors, 2024). For practical implementation, this initiative utilizes the ISA-Tab format for the organization of experimental data, thereby directly supporting the replicability and comparability of results.

### Programmatic interoperability: The breeding API (BrAPI)

To complement MIAPPE, the Breeding API (BrAPI) provides a web service interface for accessing, querying, and exchanging phenotypic and genotypic data across systems. Together, these standards support interoperability between databases, enabling efficient data integration, visualization, and computational analysis across research infrastructures. The practical integration of these standards is demonstrated in the proposed connection between institutional platforms and the BrAPI module, where client applications manage ontologies and data flows between phenotyping platforms (e.g., PSI, Phenospex) and centralized repositories. Such systems must ensure that metadata is generic and ex-tensible to allow for future mapping to diverse data consumers (Stočes et al., 2023). BrAPI serves as a standardized web service API specification, conceived through collaborative efforts within a global community, which is essential for transforming raw phenotypic observations into actionable biological and agronomic knowledge.

### Semantic integration and reference ontologies

Bridging the gap between phenotypic and genotypic data requires validated trait names and measurement scales, a task addressed by the Crop Ontology (CO). The CO supports the harmonization of data annotation through synchronized trait dictionaries, allowing for automatic annotation of field observations and direct cross-referencing with other standards like the Plant Ontology (Matteis et al., 2012).

- The Planteome project: Provides a centralized suite of reference ontologies, such as the Plant Ontology and Plant Trait Ontology, which serve as common standards for the semantic integration of genomics and phenomics da-ta. By providing annotations for 95 plant taxa, this resource enables the linking of gene function and expression to specific phenotypes, thereby facilitating data discovery and reuse across diverse research communities (Cooper et al., 2018).
- Semantic tagging: The complexity of omics datasets necessitates the use of ontologies for semantic tagging, which significantly increases data interoperability and understanding of experimental conditions (Dumschott et al., 2023).
- Practical application: These principles are demonstrated in the GnpIS repository, which utilizes a generic, ontology-driven data model integrated with MIAPPE and BrAPI standards to ensure long-term access and cutability of phenotypic datasets across diverse species (Pommier et al., 2019).

### From data interoperability to data wisdom

Beyond technical interoperability, the responsible and meaningful use of phenotyping data by end users is central to achieving what has been termed data wisdom the transformation of raw data into actionable biological understanding and informed decision-making. This requires not only robust data infrastructures but also the adoption of open science practices, transparent metadata documentation, and reproducible analytical workflows. In this context, the convergence of phenomics with computational biology, artificial intelligence, and systems modeling is expected to enhance data reusability and accelerate scientific discovery in plant research and agriculture.

In summary, phenomics constitutes a rapidly developing interdisciplinary domain that integrates plant biology, engineering, and data science to quantify the complexity of plant behavior under real world conditions. Phenotyping data not only enables the mechanistic understanding of genotype environment interactions but also support translational applications in breeding and sustainable agriculture. Given the scale and heterogeneity of phenomics datasets, future progress depends on continued advancements in computational infrastructure and the adoption

of standardized, interoperable data models such as MIAPPE and BrAPI, which are essential for transforming raw phenotypic observations into actionable biological and agronomic knowledge.

## Discussion

This systematic literature review provides a comprehensive overview of the current landscape of data management and interoperability in plant phenomics, highlighting both achievements and remaining challenges. The adoption of standards such as MIAPPE and BrAPI, together with adherence to FAIR principles, has markedly improved the transparency, discoverability, and reusability of phenotyping data. Yet, the translation of these principles into operational practices remains fragmented and often limited to well-funded infrastructures.

## Summary of main findings

This systematic review identified and synthesized the key developments in data standardization, interoperability, and computational infrastructure for plant phenomics. The analysis of available literature focused primarily on MIAPPE, BrAPI, and related FAIR data principles revealed significant progress in establishing community driven frameworks for standardized data management. These frameworks represent major milestones toward making plant phenotyping data Findable, Accessible, Interoperable, and Reusable (FAIR), facilitating cross platform integration and reuse.

Despite these advances, the adoption of standards across the research community remains uneven. While MIAPPE provides a robust metadata schema for describing experimental context, its implementation varies between institutions and platforms, often due to limited awareness, technical capacity, or resource constraints. Similarly, although BrAPI has gained recognition as a key interoperability layer between phenotyping and breeding databases, its deployment is still concentrated within a few leading infrastructures.

Furthermore, the review confirmed persistent gaps in data harmonization and se-mantic alignment. The integration of MIAPPE compliant metadata with ontologies and BrAPI endpoints remains an ongoing challenge, particularly when linking heterogeneous datasets across species or experimental designs. The lack of harmonized trait ontologies and incomplete mapping between existing vocabularies (e.g., Crop Ontology, Plant Ontology, and Environment Ontology) constrain automated interoperability and limit the reusability of data in downstream computational analyses.

From a technological standpoint, High-Performance Computing (HPC) and cloud infrastructures have become indispensable for managing the increasing complexity of phenomics data. However, the current state of the field indicates that while the theoretical framework for FAIR and interoperable phenomics data is well established, the practical realization of this vision remains fragmented. Modern phenotype monitoring systems often lack unified development standards across sensors, communication, and data analysis modules, which limit their scalability. Furthermore, due to current computational and algorithmic constraints, much of the data processing still occurs offline (Yuan et al., 2023). Despite the theoretical benefits, significant struggles remain in the practical 'FAIRification' of existing datasets (Papoutsoglou at al., 2023). Implementing FAIR principles requires a systematic approach that ad-dresses data organization, metadata standards, and secure access protocols (Krisnawijaya et al., 2024).

## Future vision

Based on the findings of this review, several key research directions and open questions emerge that should guide future work in the plant phenomics data community:

1. Bridging the gap between standard definition and implementation: Although MIAPPE and BrAPI provide clear frameworks for describing and exchanging data, many research infrastructures have yet to implement them fully. Future work should focus on developing lightweight tools, APIs, and templates that simplify adoption without requiring extensive technical expertise. The integration of MIAPPE into popular data management platforms or laboratory information systems could serve as a practical accelerator.
2. Semantic harmonization through ontology alignment: Ontologies are the backbone of interoperability, yet inconsistencies in trait naming and semantic structure still impede data integration. Efforts should concentrate on cross-linking domain ontologies (e.g., Plant Ontology, Crop Ontology, and Environment Ontology) and align them with MIAPPE metadata schemas to facilitate automatic data discovery and integration across repositories. However,

as noted in recent research, most agricultural ontologies currently lack explicit evaluation procedures. To address this, a structured evaluation framework focusing on criteria such as clarity, coherence, and completeness is essential for matching specific evaluation methods to the intended purpose, whether for decision support or system interoperability (Goldstein et al., 2021).

3. Integration of HPC and AI-based data analytics: The increasing scale of phenomics datasets demands tighter coupling between standard compliant data infrastructures and HPC/AI pipelines. Research should focus on establishing modular, reusable workflows for image analysis, time series modeling, and genotype phenotype association studies that are strictly compatible with FAIR data principles.
4. Advancing data ethics and digital responsibility: As phenomics expands into open and federated data ecosystems, the community must address issues of data security, intellectual property, and equitable access. Future infra-structures should incorporate digital responsibility goals (Meier et al., 2021), ensuring that users are digitally literate, that algorithms are transparent, and that data is trustworthy and ethically managed.
5. Reframing the research questions: A critical reflection raised by Vincent Vadez ("Got all the answers! But what was the question?") (Varshney et al., 2018) underscores that progress in phenomics is often limited not by data availability, but by the lack of well formulated research questions. Future research should move beyond mere data accumulation to focus on hypothesis driven, integrative studies that translate data into actionable biological knowledge and agronomic innovation.
6. Towards an ecosystem of interoperable wisdom: The goal of the data paradigm transforming data into knowledge nd wisdom requires a shift from technological readiness to community maturity. This includes the establishment of governance models, training programs, and incentive structures that reward FAIR and open data practices while ensuring data quality, re-producibility, and trust.

## Conclusion

This systematic literature review has provided a comprehensive synthesis of the current state of data management and interoperability in plant phenomics. The findings demonstrate that while the adoption of standards such as MIAPPE and BrAPI, alongside the FAIR principles, has significantly enhanced the transparency and discoverability of phenotypic datasets, a gap persists between theoretical frameworks and their operational implementation.

The synthesis identified three critical areas requiring urgent attention: the inconsistent integration of existing standards across diverse research platforms, the insufficient alignment of semantic frameworks and ontologies, and the need for more robust data governance mechanisms to address emerging ethical and security concerns. Overcoming these obstacles will necessitate coordinated community efforts, institutional support for long term data stewardship, and closer collaboration between data scientists, biologists, and infrastructure developers.

Ultimately, the integration of High-Performance Computing (HPC), artificial intelligence, and standardized APIs are essential for achieving scalable and reproducible phenomics research. However, technological readiness must be matched by community maturity. By fostering a culture of openness, semantic harmonization, and digital responsibility, the field can transition toward a mature data ecosystem. Such an ecosystem will be capable of transforming complex, high dimensional datasets into the actionable wisdom required to address global challenges in sustainable agriculture and food security.

## Acknowledgements

This work was supported by the EC's Digital Europe Programme in the project AGRITECH EU grant agreement No. 101123258.

The results and knowledge included herein have been obtained owing to support from the following institutional grant. Internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. IGA 2025A1012.

*Corresponding author:*
*Ing. Vladimír Voral*
*Department of Information Technologies, Faculty of Economics and Management*
*Czech University of Life Sciences Prague*
*Kamýcká 129, 165 00 Prague-Suchdol, Czech Republic*
*Email: voral@pef.czu.cz*

## References

[1] Abdul Hamid, N. A. W. and Singh, B. (2024) "High-Performance Computing Based Operating Systems, Software Dependencies and IoT Integration", In: K. A. Ahmad, K. A., Abdul Hamid, N. A. W., Jawaid, M., Khan, T. and Singh, B. (eds.) "*High Performance Computing in Biomimetics*", pp. 175-204, Series in BioEngineering. Springer, Singapore. E-ISBN 978-981-97-1017-1, ISSN 2196-8861. DOI 10.1007/978-981-97-1017-1_8.

[2] Ackoff, R. L. (1989) "From data to wisdom", *Journal of Applied Systems Analysis*, Vol. 16, pp. 3-9. ISSN 0308-9541.

[3] Aksenova, A., Johny, A., Adams, T., Gribbon, P., Jacobs, M. and Hofmann-Apitius, M. (2024) "Current state of data stewardship tools in life science", *Frontiers in Big Data*, Vol. 7, p. 1428568. E-ISSN 2624-909X. DOI 10.3389/fdata.2024.1428568.

[4] Arend, D., Junker, A., Scholz, U., Schüler, D. and Selbig, J. (2022) "From data to knowledge - big data needs stewardship, a plant phenomics perspective", *The Plant Journal*, Vol. 110, No. 1, pp. 12-28. E-ISSN 1365-313X, ISSN 0960-7412. DOI 10.1111/tpj.15804.

[5] Arend, D., Junker, A., Scholz, U., Schüler, D., Wylie, J. and Lange, M. (2016) "PGP repository: a plant phenomics and genomics data publication infrastructure", *Database (Oxford)*, Vol. 2016, p. baw033. ISSN 1758-0463. DOI 10.1093/database/baw033.

[6] Bosilj, P., Duckett, T. and Cielniak, G. (2018) "Connected attribute morphology for unified vegetation segmentation and classification in precision agriculture", *Computers in Industry*, Vol. 98, pp. 226-240. ISSN 0166-3615. DOI 10.1016/j.compind.2018.02.003.

[7] Bosilj, P., Duckett, T., Cielniak, G. and Pearson, S. (2018) "Quantitative phenotyping of plants using three-dimensional imaging and machine vision", *Computers and Electronics in Agriculture*, Vol. 153, pp. 69-79. ISSN 1872-7107.

[8] Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., Mungall, C., Sinn, B. T., Cavaliere, D., Carbon, S., Dunn, N. A., Smith, B., Qu, B., Preece, J., Zhang, E., Todorovic, S., Gkoutos, G., Doonan, J. H., Stevenson, D. W., Arnaud, E. and Jaiswal, P. (2018) "The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics", *Nucleic Acids Research*, Vol. 46, No. D1, pp. D1168 - D1180. E-ISSN 1362-4962. DOI 10.1093/nar/gkx1152.

[9] Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., van Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S.A., Scholz, U., van Schriek, M., Seren, Ü., Usadel, B., Weise, S., Kersey, P. and Krajewski, P. (2016) "Measures for interoperability of phenotypic data: minimum information requirements and formatting", *Plant Methods*, Vol. 12, No. 44. E-ISSN 1746-4811. DOI 10.1186/s13007-016-0144-4.

[10] Dumschott, K., Dörpholz, H., Laporte, M. A., Brilhaus, D., Schrader, A., Usadel, B., Neumann, S., Arnaud, E. and Kranz, A. (2023) "Ontologies for increasing the FAIRness of plant research data", *Frontiers in Plant Science*, Vol. 14, p. 1279694. E-ISSN 1664-462X. DOI 10.3389/fpls.2023.1279694.

[11] European Commission. (2025) "*Legal framework of EU data protection*". [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en [Accessed: Dec. 15, 2025].

[12] European Union. (2018) "General Data Protection Regulation (GDPR)", *Official Journal of the European Union*, L119, pp. 1-88. E-ISSN 1725-2423.

[13] Fiorani, F. and Schurr, U. (2013) "Future scenarios for plant phenotyping", *Annual Review of Plant Biology*, Vol. 64, pp. 267-291. ISSN 1545-2123. DOI 10.1146/annurev-arplant-050312-120137.

[14] Frontiers in Plant Science. (2023) "*Phenomics as an emerging research discipline*". [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2023.1233794/full [Accessed: Sept 21, 2025].

[15] GDPR-info.eu. (2024) "*General Data Protection Regulation (GDPR) - Legal Text*". [Online]. Available: https://gdpr-info.eu [Accessed: Sept 21, 2025].

[16] Georgiou, Y., Zhou, N., Zhong, L., Hoppe, D., Pospieszny, M., Papadopoulou, N., Nikas, K., Nikolos, O. L., Kranas, P., Karagiorgou, S., Pascolo, E., Mercier, M. and Velho, P. (2020) "Converging HPC, Big Data and Cloud Technologies for Precision Agriculture Data Analytics on Supercomputers", In: Jagode, H., Anzt, H., Juckeland, G., Ltaief, H. (eds) *High Performance Computing. ISC High Performance 2020*. Lecture Notes in Computer Science, Vol. 12321, Springer, Cham. E-ISBN 978-3-030-59851-8. DOI 10.1007/978-3-030-59851-8_25.

[17] Ghanem, M. E., Marrou, H. and Sinclair, T. R. (2015) "Physiological phenotyping and its application to the breeding of drought-tolerant crops", *Frontiers in Physiology*, Vol. 6, 362. E-ISSN 1664-042X. DOI 10.3389/fphys.2015.00362.

[18] Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q. and Buckler, E. S. (2014) "TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline", *PLoS One*, Vol. 9, No. 2, p. e90346. ISSN 1932-6203. DOI 10.1371/journal.pone.0090346.

[19] Goldstein, A., Fink, L. and Ravid, G. (2021) "A Framework for Evaluating Agricultural Ontologies", *Sustainability*, Vol. 13, No. 11, p. 6387. ISSN 2071-1050. DOI 10.3390/su13116387.

[20] Grüning, B. Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch C. H., Valieris, R., Köster, J. and Biocomda Team (2018) "Bioconda: sustainable and comprehensive software distribution for the life sciences", *Nature Methods*, Vol. 15, No. 7, pp. 475-476. ISSN 1548-7105. DOI 10.1038/s41592-018-0046-7.

[21] Kartal, S., Choudhary, S., Stočes, M., Šimek, P., Vokoun, T. and Novák, V. (2020) "Segmentation of Bean-Plants Using Clustering Algorithms", *AGRIS on-line Papers in Economics and Informatics*, Vol. 12, No. 3, pp. 36-43. ISSN 1804-1930. DOI 10.7160/aol.2020.120304.

[22] Kartal, S., Masner, J., Kholová, J., Galba, A., Murugesan, T., Baddam, R., Mikes, V. and Kánská, E. (2025) "AI-Driven Background Segmentation for High-Throughput 3D Plant Scans", *IEEE Access*, Vol. 13, pp. 136027-136037. DOI 10.1109/ACCESS.2025.3594406. ISSN 2169-3536.

[23] Krisnawijaya, N. N. K., Tekinerdogan, B., Catal, C., van der Tol, R. and Herdiyeni, Y. (2025) "Implementing FAIR principles in data management systems: A multi-case study in precision farming", *Computers and Electronics in Agriculture*, Vol. 230, p. 109855. ISSN 0168-1699. DOI 10.1016/j.compag.2024.109855.

[24] LeBauer, D., Maxwell, B., Demieville, J., Fahlgren, N., French, A., Garnett, R., Hu, Z., Huynh, K., Kooper, R., Li, Z., Maimaitijiang, M., Mao, J., Mockler, T., Morris, G., Newcomb, M., Ottman, M., Ozersky, P., Paheding, S., Pauli, D., Pless, R., Quin, W., Riemer, K., Rohde, G., Rooney, W., Sagan, V., Shakoor, N., Stylianou, A., Thorp, K., Ward, R., White, J., Willis, C. and Zender, C. (2020) "*Data From: TERRA-REF, An open reference data set from high resolution genomics, phenomics, and imaging sensors*", [Dataset], Dryad. DOI 10.5061/dryad.4b8gtht99.

[25] Matteis, L., Skofic, M., Portugal, A., Mclaren, G., Hyman, G. and Arnaud, E. (2012) "Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice", *Frontiers in Physiology*, Vol. 3, p. 326. E-ISSN 1664-042X. DOI 10.3389/fphys.2012.00326.

[26] Meier, P., Deksnyte, G. and Winter, R. (2021) "Digital Responsibility Goals: A Framework for the Responsible Use of Data and Algorithms", *Business & Information Systems Engineering*, Vol. 63, No. 6, pp. 665-678. ISSN 1867-0202. DOI 10.3233/SHTI220377.

[27] MIAPPE Contributors. (2024) "*Minimum Information About a Plant Phenotyping Experiment (MIAPPE) Specification, Version 1.2*". [Online]. Available: https://github.com/MIAPPE [Accessed: Oct 25, 2025].

[28] Minssen, T., Rutz, B. and van Zimmeren, E. (2020) "Clinical trial data transparency and GDPR compliance", *Science and Public Policy,* Vol. 47, No. 2, pp. 228-238. ISSN 1471-5430. DOI 10.1093/scipol/scaa014.

[29] Nicora, G., Vitali, F., Dagliati, A., Geifman, N. and Bellazzi, R. (2020) "Integrated Multi-Omics Analyses in Oncology: A Review of Machine Learning Methods and Tools", *Frontiers in Oncology*. Vol. 10. E-ISSN. 2234-943X. DOI 10.3389/fonc.2020.01030.

[30] OECD. (2022) "*Responding to societal challenges with data: Access, sharing, stewardship and control*", OECD Publishing. [Online]. Available: https://www.oecd.org/en/publications/responding-to-societal-challenges-with-data_2182ce9f-en.html [Accessed: Oct 25, 2025].

[31] Papoutsoglou, E. A., Farian D., Arend, D., Arnaud, E., Athanasiadis, I. N., Chaves, I., Coppens, F., Cornut, G., ...Pommier, C. (2020) "Enabling reusability of plant phenomic datasets with MIAPPE 1.1", *New Phytologist*, Vol. 227, No. 1, pp. 260-273. E-ISSN 1469-8137. DOI 10.1111/nph.16544.

[32] Papoutsoglou, E. A., Athanasiadis, I., Visser, R. and Finkers, R. (2023) "The benefits and struggles of FAIR data: the case of reusing plant phenotyping data", *Scientific Data*, Vol. 10. E-ISSN 2052-4463. DOI 10.1038/s41597-023-02364-z.

[33] Pommier, C., Michotey, C., Cornut, G., Roumet, P., Duchêne, E., Flores, R., Lebreton, A., Alaux, M., Durand, S., Kimmel, E., Letellier, T., Merceron, G., Laine, M., Guerche, C., Loaec, M., Steinbach, D., Laporte, M. A., Arnaud, E., Quesneville, H. and Adam-Blondon, A. F. (2019) "Applying FAIR Principles to Plant Phenotypic Data Management in GnpIS", *Plant Phenomics*, Vol. 2019, p. 1671403. ISSN 2643-6515. DOI 10.34133/2019/1671403.

[34] Prifti, K., Krijger, J., Thuis, T. and Stamhuis, E. (2023) "From Bilateral to Ecosystemic Transparency: Aligning GDPR's Transparency Obligations with the European Digital Ecosystem of Trust", In: Kuhlmann, S., De Gregorio, F., Fertmann, M., Ofterdinger, H. and Sefkow, A. (eds.) *Transparency or Opacity : A Legal Analysis of the Organization of Information in the Digital World*, 1st ed., pp. 115, Nomos. ISBN 978-3-7560-0027-2. DOI 10.5771/9783748936060-115.

[35] Rowley, J. (2007) "The wisdom hierarchy: representations of the DIKW hierarchy", *Journal of Information Science*, Vol. 33, No. 2, pp. 163-180. E-ISSN 1741-6485. DOI 10.1177/0165551506070706.

[36] Selby, P., Abbeloos, R., Backlund, J. E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O. E., Birkett, C., Calaminos, V. C., Carceller, P., Cornut, G. ... The BrAPI consortium (2019) "BrAPI - an application programming interface for plant breeding applications", *Bioinformatics*, Vol. 35, pp. 4147-4155. ISSN 1367-4811. DOI 10.1093/bioinformatics/btz190.

[37] Selby, P., Abbeloos, R., Adam-Blondon, A.F., Agosto-Pérez, F. J., Alaux, M., Alic, I., Al-Shamaa, K., Aparicio, ... BrAPI Consortium. (2025) "BrAPI v2: real-world applications for data integration and collaboration in the breeding and genetics community", *Database*, Vol. 2025, p. baaf048. ISSN 1758-0463. DOI 10.1093/database/baaf048.

[38] Sterling, T., Anderson, M. and Brodowicz, M. (2024) "*High performance computing: Modern systems and practices*", 2nd ed., Elsevier. ISBN 9780128230350. DOI 10.1016/C2013-0-09704-6.

[39] Stočes, M., Jarolímek, J., Anderle, M., Kholová, J., Pavlík, J., Masner, J., Spichal, L. and Klimes, P. (2025) "*Plant Phenotyping Network: Data Standards*". Poster, National EOSC CZ Conference 2025, Ostrava, Czech Republic. [Online]. Available: https://www.eosc.cz/media/4054734/stoces_a1_height-poster-stoces.pdf [Accessed: Dec. 12, 2025].

[40] Stočes, M., Vaněk, J., Jarolímek, J., Novák, V., Masner, J., Šimek, P., Kánská, E., Havránek, M., Kubata, K. and Voral, V. (2023) "Agriculture Data Platform - Institutional Data Repository - Selected Aspects", *AGRIS on-line Papers in Economics and Informatics*, Vol. 15, No. 4, pp. 127-133. DOI 10.7160/aol.2023.150409.

[41] Tardieu, F., Cabrera-Bosquet, L., Pridmore, T. and Bennett, M. (2017) "Plant phenomics, from sensors to knowledge", *Current Biology*, Vol. 27, No. 15, pp. R770 - R783. E-ISSN 0960-9822. DOI 10.1016/j.cub.2017.05.055.

[42] Tricco, A. C., Lillie, E., Zarin, W., O´Brien, K. K., ColQuhoun, H., Levac., D., Moher, D., Peters, M. D. J...Straus, S. E. (2018) "PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation", *Annals of Internal Medicine*, Vol. 169, No. 7, pp. 467-473. E-ISSN 1539-3704. DOI 10.7326/M18-0850.

[43] Ubbens, J., Stavness, I., Pound, M.P. and Wei Guo. (2025) "Deep learning in plant phenotyping: the first ten years", *Plant Phenomics*, Vol. 7, No. 4, p. 100062. ISSN 2643-6515. DOI 10.1016/j.plaphe.2025.100062.

[44] UK Data Service. (2025) "FAIR data principles". [Online]. Available: https://ukdataservice.ac.uk/learning-hub/research-data-management/plan-to-share/fair-data-principles/ [Accessed: Sept. 22, 2025].

[45] Umbach, G. (2024) "Open Science and the impact of Open Access, Open Data, and FAIR publishing principles on data-driven academic research: Towards ever more transparent, accessible, and reproducible academic output?", *Statistical Journal of the IAOS*, Vol. 40, No. 1, pp. 59-70. ISSN 1875-9254. DOI 10.3233/SJI-240021.

[46] Varshney, R. K., Thudi, M., Pandey, M. K., Tardieu, F., Ojiewo, C., Vadez, V., Whitbread, A. M., Siddiwue, K. H. M., Nguyen, H. T., Carberry, P. S. and Bergvinson, D. (2018) "Accelerating genetic gains in legumes for the development of prosperous smallholder agriculture: integrating genomics, phenotyping, systems modelling and agronomy", *Journal of Experimental Botany*, Vol. 69, No. 13, pp. 3293-3312. ISSN 0022-0957. DOI 10.1093/jxb/ery088.

[47] Wilkinson, M. D., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomebrg, N, Boiten, J.-W...Mons, B. (2016) "The FAIR Guiding Principles for scientific data management and stewardship", *Scientific Data*, Vol. 3, p. 160018. E-ISSN 2052-4463. DOI 10.1038/sdata.2016.18.

[48] Wong, J., Henderson, T. and Ball, K. (2021) "Data protection for the common good: Developing a framework for a data protection-focused data commons", *Data & Policy*, Vol. 3, p. e41. E-ISSN 2632-3249. DOI 10.1017/dap.2021.40.

[49] Xu, R. and Li, C. (2022) "A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots", *Plant Phenomics*, Vol. 2022, pp. 1-20. ISSN 2643-6515. DOI 10.34133/2022/9760269.

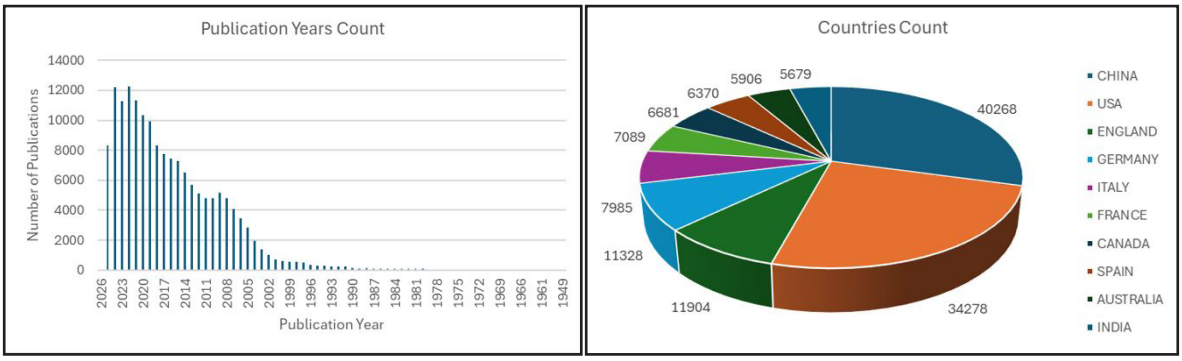
[50] Xu, Y., Zhang, X., Li, H., Zheng, H., Zhang, J., Olsen, M. S., Varshney, R. K., Prasanna, B. M. and Qian, Q. (2022) "Smart breeding driven by big data, artificial intelligence, and integrated genomic-enviromic prediction", *Molecular Plant*, Vol. 15, No. 11, pp. 1664-1695. ISSN 1674-2052. DOI 10.1016/j.molp.2022.09.001.

[51] Yuan, H., Song, M., Liu, Y., Xie, Q., Cao, W., Zhu, Y. and Ni, J. (2023) "Field Phenotyping Monitoring Systems for High-Throughput: A Survey of Enabling Technologies, Equipment, and Research Challenges", *Agronomy*, Vol. 13, No. 11, 2832. ISSN 2073-4395. DOI 10.3390/agronomy13112832.

## Appendix A: Detailed bibliometric analysis per Search Query

This appendix presents a detailed breakdown of the bibliometric data for each of the eleven search queries (SQ1–SQ11). For each query, the annual publication growth and the geographic distribution of the top 10 contributing countries are provided.

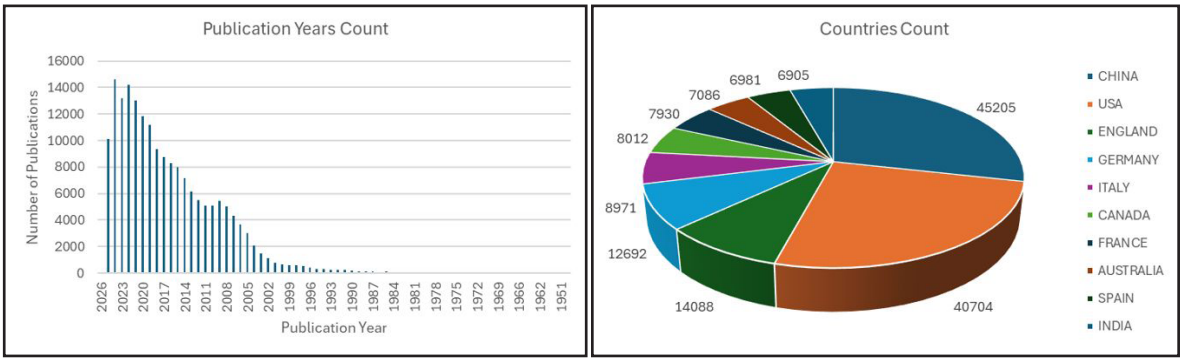
Search Query 1 (SQ1)




Source: Own processing

Figure A1: Bibliometric profile for SQ1: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

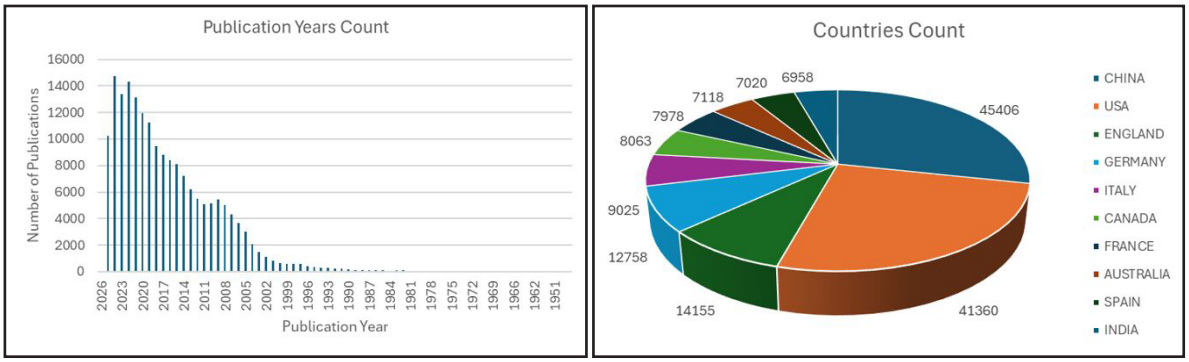
Search Query 2 (SQ2)




Source: Own processing

Figure A2: Bibliometric profile for SQ2: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

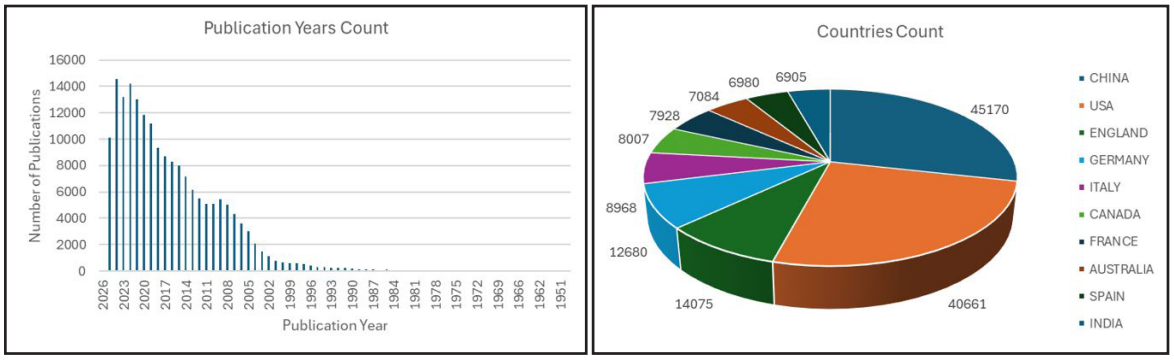
Search Query 3 (SQ3)




Source: Own processing

Figure A3: Bibliometric profile for SQ3: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

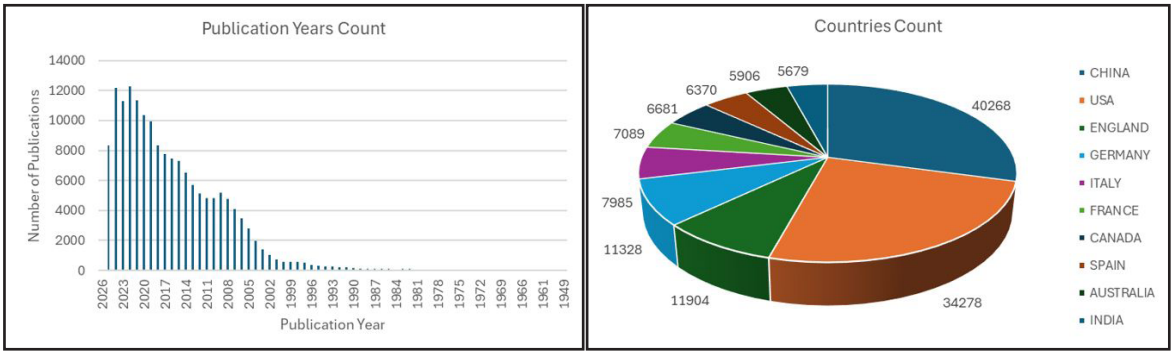
## Search Query 4 (SQ4)




Source: Own processing

Figure A4: Bibliometric profile for SQ4: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

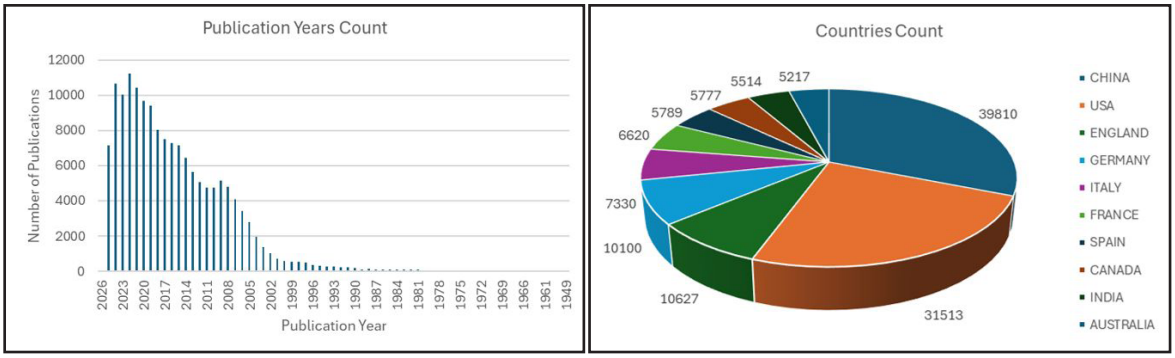
## Search Query 5 (SQ5)




Source: Own processing

Figure A5: Bibliometric profile for SQ5: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

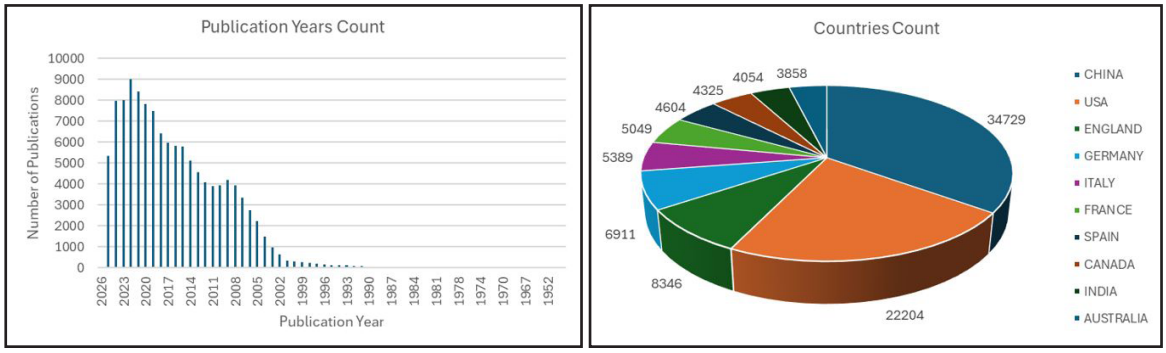
## Search Query 6 (SQ6)




Source: Own processing

Figure A6: Bibliometric profile for SQ6: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

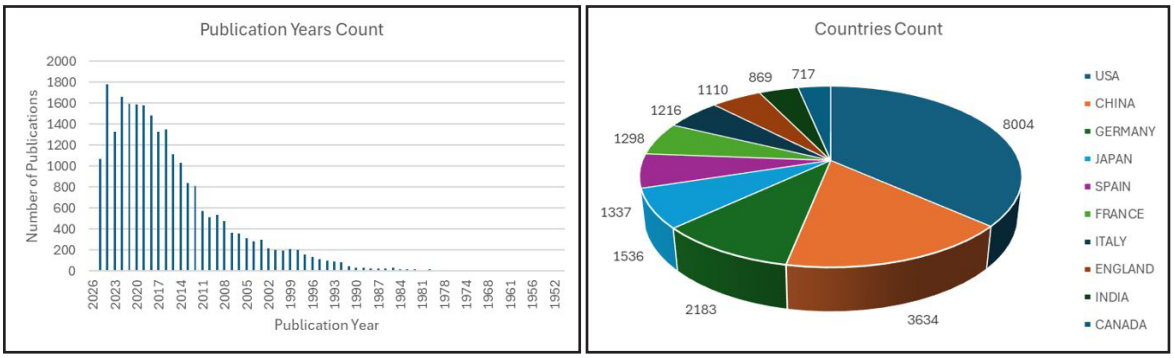
## Search Query 7 (SQ7)




Source: Own processing

Figure A7: Bibliometric profile for SQ7: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

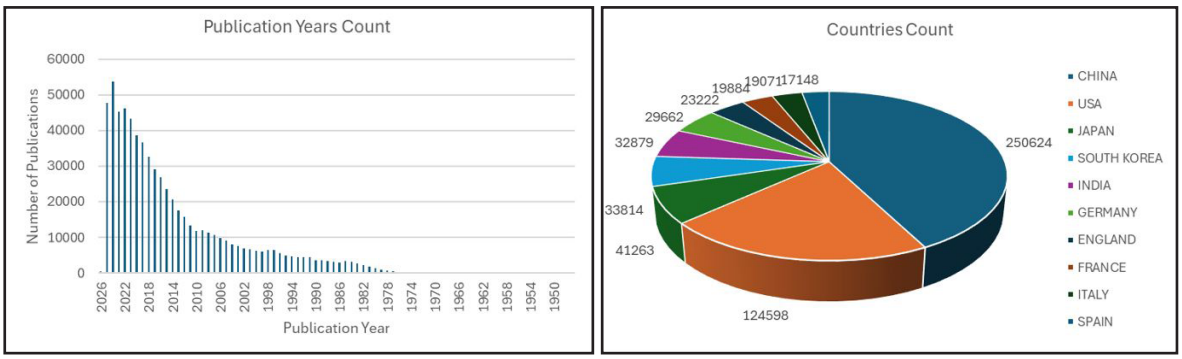
## Search Query 8 (SQ8)




Source: Own processing

Figure A8: Bibliometric profile for SQ8: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

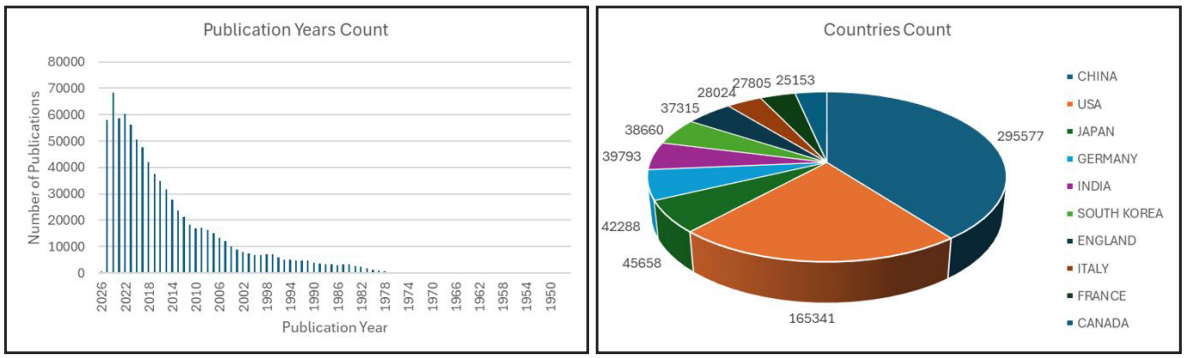
## Search Query 9 (SQ9)




Source: Own processing

Figure A9: Bibliometric profile for SQ9: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

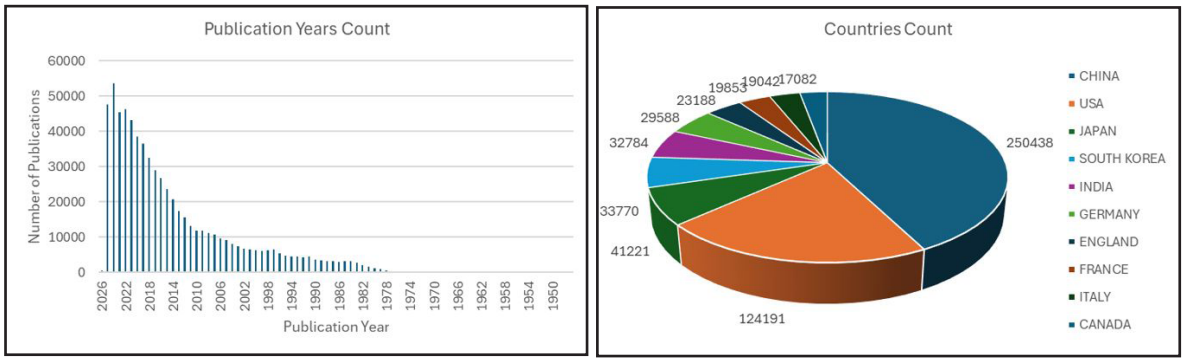
## Search Query 10 (SQ10)




Source: Own processing

Figure A10: Bibliometric profile for SQ10: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.

## Search Query 11 (SQ11)




Source: Own processing

Figure A11: Bibliometric profile for SQ11: (a-left) Annual publication trends (1949–2025); (b-right) Geographic distribution of the top 10 most active countries.