PLANT PHENOTYPING NETWORK: DATA STANDARDS

National EOSC CZ Conference 2025 - Long Live Research Data, Ostrava, Czech Republic

Data Standards in Phenotyping

Data Standards in Phenotyping

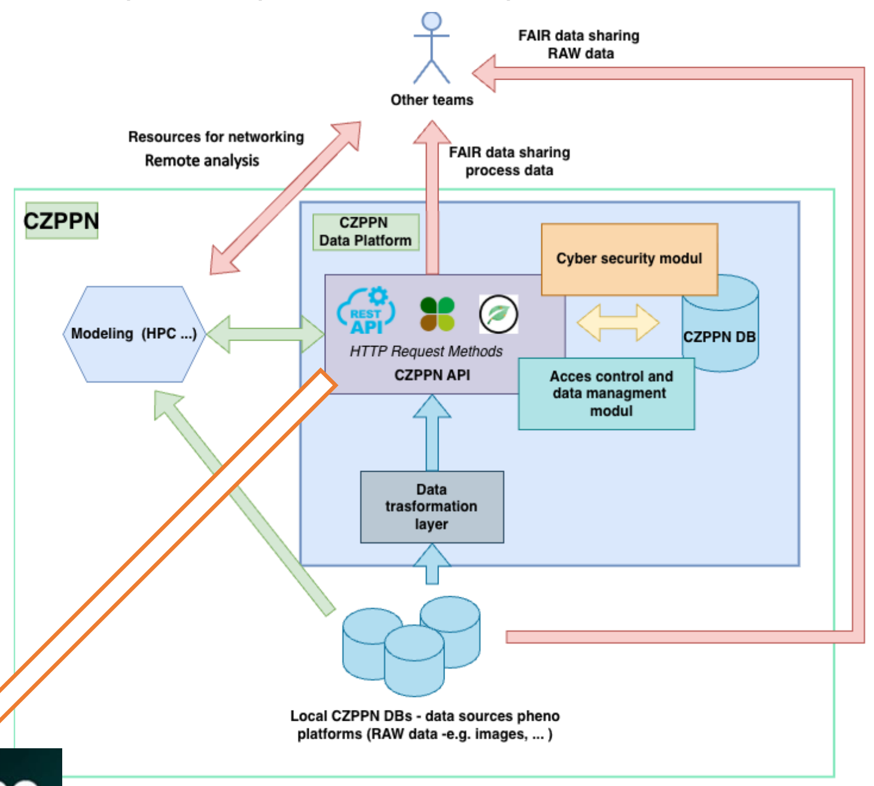
Data standards in plant phenotyping provide a common framework for describing and exchanging the diverse datasets generated by imaging, sensor systems, and environmental monitoring. Using established standards such as *MIAPPE* for metadata, controlled ontologies, and *BrAPI* interfaces ensures data consistency, interoperability, and *FAIR* compliance. These standards enable reliable integration, comparison, and reuse of phenotyping data across platforms and research teams.

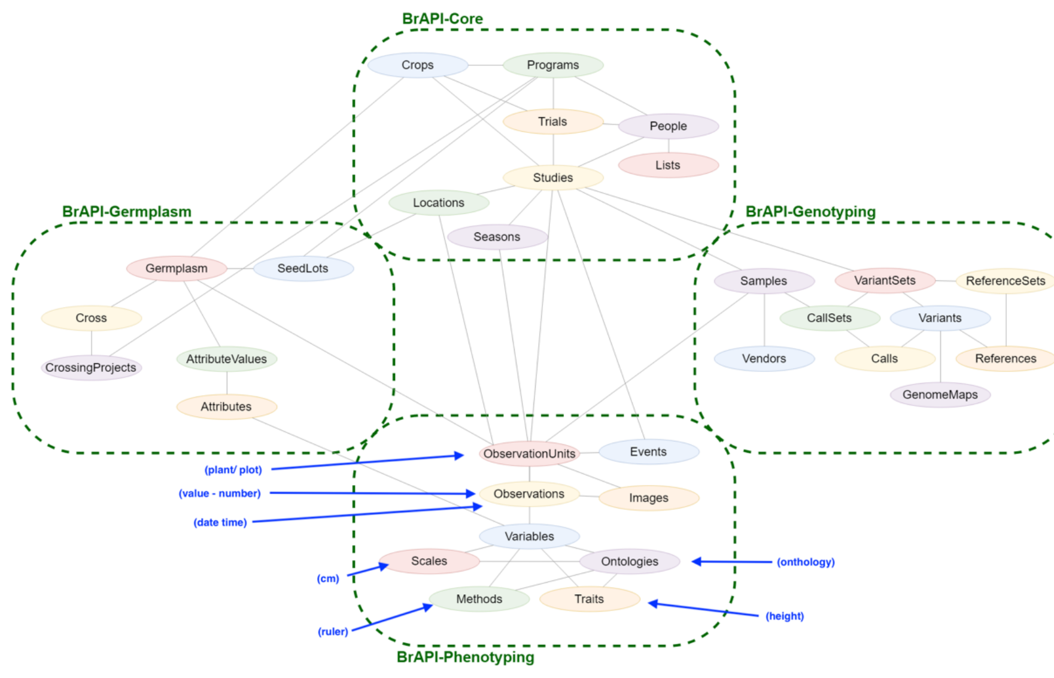
MIAPPE (Minimum Information About a Plant Phenotyping Experiment) is a comprehensive metadata standard designed to ensure consistent, structured, and interoperable documentation of plant phenotyping experiments. It defines a detailed set of descriptors covering experimental design, plant material, growth conditions, treatments, environmental monitoring, measurement protocols, and data processing workflows. *MIAPPE* provides a harmonized semantic framework built on controlled vocabularies and ontologies, enabling precise and machine-interpretable annotation of all experimental components. This significantly enhances reproducibility, cross-study integration, and long-term data usability across phenotyping infrastructures.

From a technical perspective, implementing *MIAPPE* requires establishing a standardized metadata pipeline that captures and validates all mandatory and recommended fields throughout the experiment's lifecycle. Practically, this includes creating *MIAPPE-compliant* data templates (e.g., ISA-Tab, ISA-JSON), integrating *MIAPPE* metadata fields directly into data acquisition software, and enforcing ontology-based picklists for traits, environmental parameters, and protocols. Metadata should be stored in structured formats (CSV, JSON, or relational database tables) and validated using automated schema-checking tools. Integration with *BrAPI* endpoints ensures programmatic access to MIAPPE-compliant datasets, while mapping tools such as the *EMPHASIS Data Stewardship Wizard or PhenoMeNal* workflows can assist in generating complete *MIAPPE* packages. When adopted as part of a centralised data platform, MIAPPE enables seamless ingestion, cataloguing, and *FAIR-compliant* dissemination of phenotyping datasets. See *miappe.org*.

BrAPI (Breeding API)) is an open, *RESTful web-service* specification designed to standardise the exchange of plant breeding and phenotyping data across heterogeneous information systems. It defines a comprehensive set of endpoints that cover key domains, including germplasm, trials, studies, environmental parameters, observation units, raw and processed measurements, and high-volume imaging data. *BrAPI* uses *JSON* as its primary data format, ensuring both machine readability and compatibility with modern analytical workflows and cloud-native environments.

Technically, *BrAPI* provides robust mechanisms for working with large data sets, including pagination, filtering, standardised query parameters, and format validation. It integrates security protocols such as OAuth2, enabling controlled access to sensitive or license-protected data. This makes *BrAPI* easy to use in cloud, containerised, or hybrid infrastructure environments. Within *CZPPN*, *BrAPI* serves as the primary programmatic access point between local repositories, the transformation layer, and the central data platform. The implementation uses Node.js REST services that map PostgreSQL schemas to *BrAPI* resources via middleware. This enables automated data integration with analytical models, FAIR catalogues (PhenoSERVICE, EOSC), and external services, making CZPPN a fully interoperable part of the European data infrastructure. See brapi.org.





BrAPI High Level Domain Map

Data platform is being developed within the emerging *CZPPN* (Czech Plant Phenotyping Network) to support unified data management across all facilities. It will enable structured processing, integration, and long-term organization of phenotyping datasets. The platform will also provide standardized data sharing based on *MIAPPE* metadata and *BrAPI* interfaces, supporting reuse within *CZPPN* and by external research team

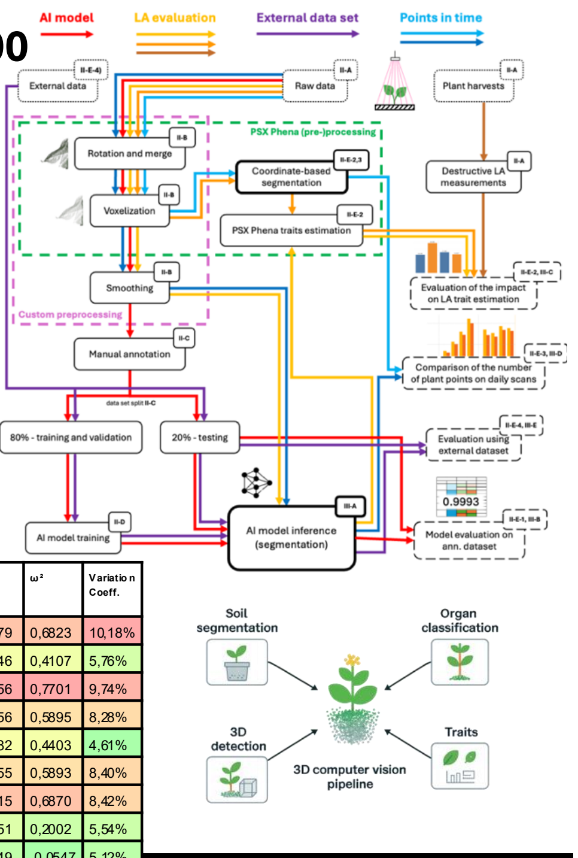
AI/ML and Computer Vision

3D computer vision using Plant Scans from Phenospex PlantEye F600

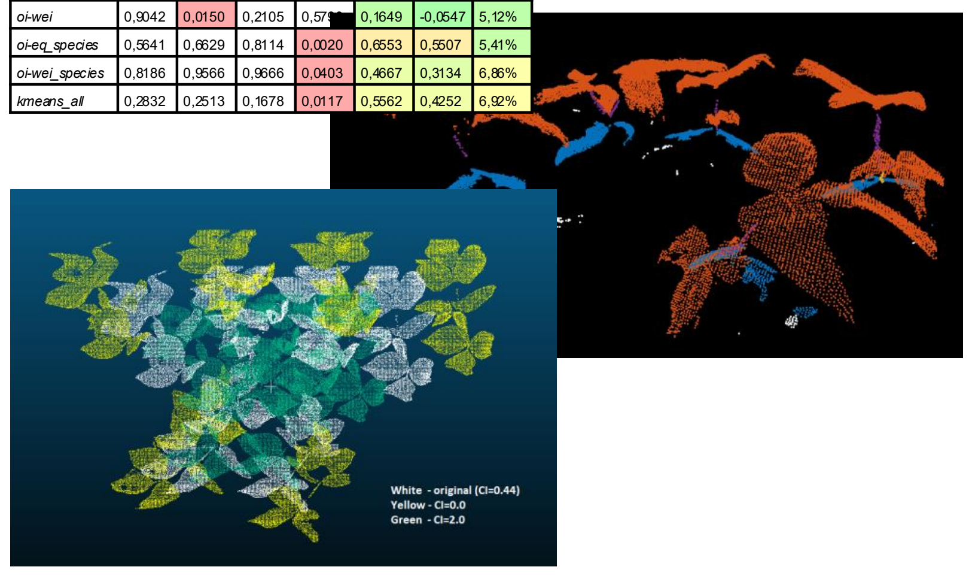
Soil and background segmentation for PlantEye F600 microplot scans. We published a novel method for segmentation of background vs. plant in the **point cloud** using **RGB+XYZ+NIR** and an MLP-based AI model. This outperforms simple coordinate cuts: it **preserves under rim points** and **refines trait estimates** derived from the scanned point cloud.

Leakage-safe data splitting for computer vision models. To avoid inflated scores from train/test contamination, we enforce **group-constrained CV**(plants/pots/sessions kept strictly separate) with light **stratification** by species/age/plant-count. We quantify scene difficulty with a **Complexity/Overlap Index (OI)** that captures geometric/spectral overlap; Reporting focuses on **AP@IoU (0.3/0.5/0.7)** plus **count AUC metrics**(ARMSE/AMAE/ARSQ) for stability across thresholds. In practice, **OI-guided folds** yield **more coherent** behavior across metrics, while individual random splits behave differently.

Development of novel augmentation methods. We tune methods by complexity: **geometric** (rot/scale/jitter), **spectral** (intensity/illumination), **occlusion/drop** (simulated leaf overlap), and **noise**. Combining **OI-guided splits** with **targeted augmentation** delivers **more stable detection** and **better out-of-domain performance** than baseline practices.



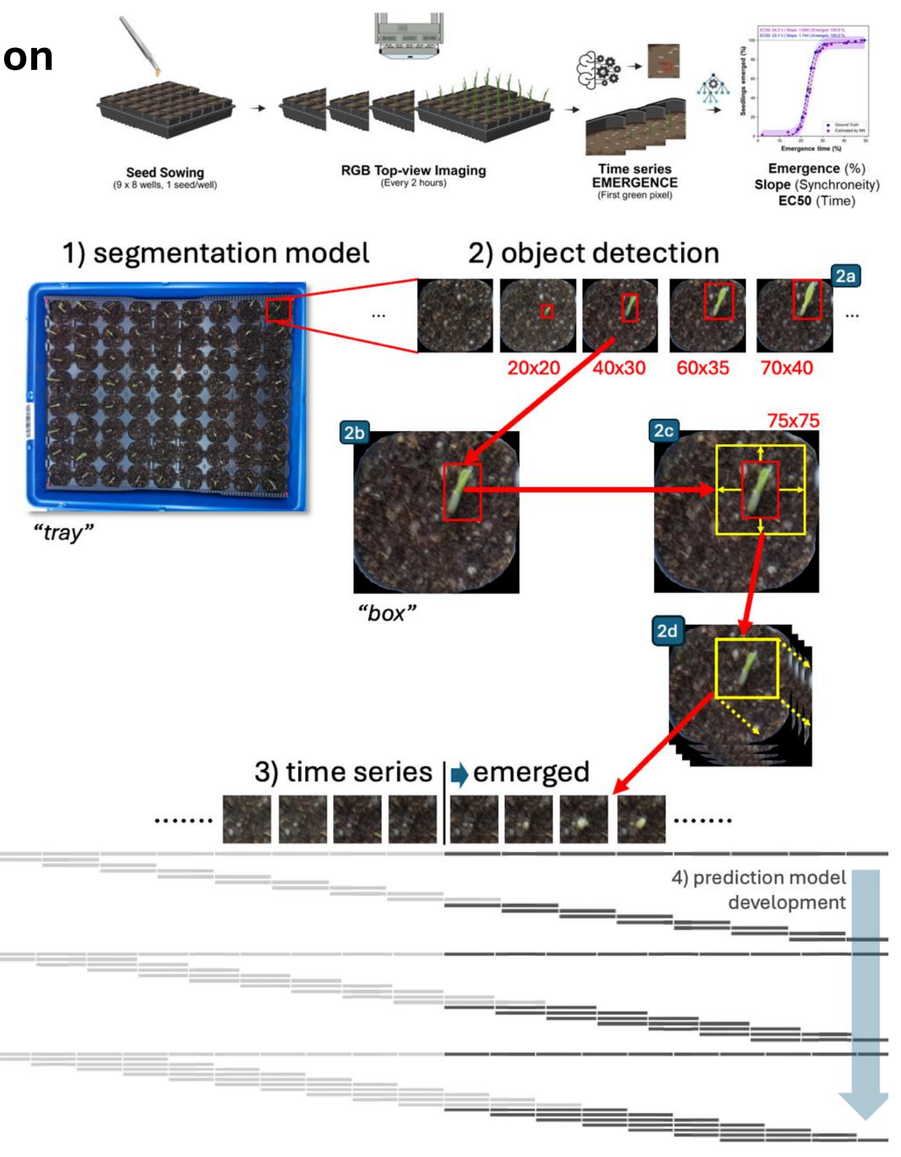
Split method	Shapiro-Wilk p-val	Levene p-val	Bartlett p-val	ANOVA p-val	η²	ω²	Variation Coeff.
rnd_0	0,2167	0,0082	0,0746	0,0002	0,7579	0,6823	10,18%
rnd_1	0,7896	0,2406	0,3339	0,0139	0,5446	0,4107	5,76%
rnd_2	0,6657	0,5871	0,8649	0,0000	0,8256	0,7701	9,74%
rnd_3	0,1520	0,4312	0,3242	0,0010	0,6856	0,5895	8,28%
pcl	0,3312	0,4003	0,8651	0,0097	0,5682	0,4403	4,61%
pcl_num	0,5541	0,0275	0,3357	0,0010	0,6855	0,5893	8,40%
ppl	0,4454	0,2254	0,1836	0,0001	0,7615	0,6870	8,42%
oi-eq	0,8669	0,4432	0,2965	0,1121	0,3751	0,2002	5,54%
oi-wei	0,9042	0,0150	0,2105	0,57[illegible]	0,1649	-0,0547	5,12%
oi-eq_species	0,5641	0,6629	0,8114	0,0020	0,6553	0,5507	5,41%
oi-wei_species	0,8186	0,9566	0,9666	0,0403	0,4667	0,3134	6,86%
kmeans_all	0,2832	0,2513	0,1678	0,0117	0,5562	0,4252	6,92%



Cost-effective image time-series emergence detection

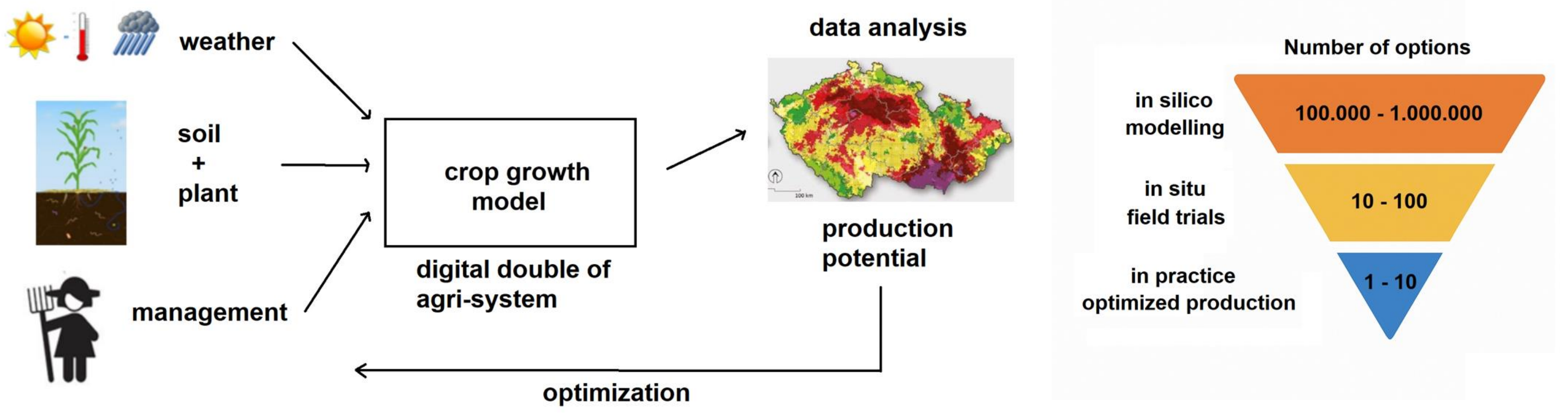
We develop an **open pipeline** for automated scoring of seedling emergence from RGB time-series trays. Full trays are segmented into wells (Mask R-CNN R-50-FPN), sequences are cropped around each plant using YOLOv8n (→ 75×75 patches), features are extracted with EfficientNetV2-B0 and classified over time via TCN/LSTM. Two independent experiments (Raspberry Pi array vs. PSI RGB system) demonstrate robust transfer. Raw per-frame accuracies reach ~0.966–0.984 and improve after post-processing and ±2 h tolerance; outputs are cumulative emergence curves and per-plant timestamps; code and data organization are provided.

We provide **open-source code** with reproducible configs and documented baselines. The pipeline is **modular**—tray/well segmentation (AI model), data reduction (object detection + algorithm), image classification model, temporal head, and post-processing are plug-replaceable. Users can **retrain on their own data** using ready-made code and training instructions. (available soon at GitHub: https://github.com/kit-pef-czu-cz/emergence-detection)



Modelling

Agricultural modelling as a tool to optimize crop production



Modelling using **crop growth models** such as **APSIM** allows to perform **large scale analysis of an agri-system**. Due to the in-silico nature of the research, it is possible to explore vast array of possibilities to try to find promising scenarios for future testing in the field. We developed an array of tools (C# programs, python scripts etc.) to help automate this process by deploying **batch processing and parallel data processing**.

When dealing with **millions of simulations**, **large amount of data** is generated (~TB). Also, during the analysis, additional statistical parameters are calculated and added to the data. In the last step, we calculate **IoG ("Index of Goodness")** in an effort to capture the performance of a given simulation by a single number. This requires **complex mathematical modelling** and further increase in the size of data.

Due to different parameters being changed in this exploratory research, it is useful to keep **different version of the datasets available** (in different stages of processing). Therefore, the need for **large capacity storage** that is capable of maintaining the data while allowing **easy access** and the **ability to re-upload the semi-processed data** back into it.

Sorghum as drought resistant alternative to maize in the Czech Republic

We are currently working with **sorghum** to analyze its potential in the Czech Republic. The modelling allows us to find promising combinations of management practices to **optimize the yield**. But more than the yield itself, the main advantage of sorghum over maize is **lesser need for fertilization** and **higher resistance to stress induced by droughts**.

The modelling also provides information on areas where breeders can improve the sorghum crop to be better suited for a Czech climate (temperate - continental, type D). The results show that one of the main areas of interest will be **acclimatization for cold**, since the parameter for cardinal temperature (minimum) has high impact on yield and biomass production.

Another promising results come in the form of **production stability** with regards to low nitrogen inputs. That would mean savings for farmers in terms of **fertilization costs**.

Authors, Affiliations, Acknowledgment

Michal Stočes[1], **Jan Jarolímek**[1], **Michael Anderle**[1], Jana Kholová[1,2], Jan Pavlík[1], Jan Masner[1], Lukáš Spichal[2], Pavel Klimes[2]

[1] Department of Information Technologies, Faculty of Economics and Management, Czech University of Life Sciences Prague

[2] Czech Advanced Technology and Research Institute (CATRIN), Palacky University Olomouc, Czech Republic

The results and knowledge presented herein have been obtained owing to the support of the projects "Precizní zemědělství a digitalizace v ČR" (Precision Agriculture and Digitalization in Czech Republic), reg. no. **QK23020058** and internal grant agency of the Faculty of Economics and Management, Czech University of Life Sciences Prague, grant no. **IGA 2023B0005**.