ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 10/04/2024

DOI: https://doi.org/10.3897/arphapreprints.e125091

Prototype Biodiversity Digital Twin: Crop Wild Relatives Genetic Resources for Food Security

Desalegn Chala, Erik Kusch, Claus Weiland, Carrie Andrew, Jonas Grieb, Tuomas Rossi, Tomas Martinovic, Dag Endresen

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 10/04/2024*. DOI: https://doi.org/10.3897/arphapreprints.e125091

Prototype Biodiversity Digital Twin: Crop Wild Relatives Genetic Resources for Food Security

Desalegn Chala[‡], Erik Kusch[‡], Claus Weiland[§], Carrie Andrew[‡], Jonas Grieb[§], Tuomas Rossi[|], Tomas Martinovic[¶], Dag Endresen[‡]

‡ University of Oslo, Oslo, Norway
§ Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Frankfurt am Main, Germany
| CSC – IT Center for Science Ltd., Espoo, Finland
¶ IT4Innovations, VSB - Technical University of Ostrava, Ostrava-Poruba, Czech Republic

Corresponding author: Desalegn Chala (desdchala@gmail.com) Reviewable v 1

Abstract

Amidst population growth and climate-driven crop stresses, ensuring food security demands innovative strategies. Crop wild relatives (CWR), wild plants in the same genus as the crop, offer novel genetic resources crucial for enhancing crop resilience. Here, we introduce a prototype digital twin (pDT) to aid in searching and utilising CWR genetic resources. Leveraging the MoDGP (Modeling the Germplasm of Interest) tool, the pDT enables mapping geographic areas where stress-tolerant CWR populations can be found. With its graphical user interface, it offers flexibility in selecting genetic resources from CWR tailored to enhance resilience of various crops against diverse stress factors.

Keywords

Crop Wild Relatives, Biodiversity Digital Twin, MoDGP, Destination Earth, Sustainable Development Goals

Introduction

Population growth and climate change are two of the major factors that are challenging food security. The human population has increased from one to eight billion over the past 200 years and is expected to reach 11 billion by the end of this century (Roser et al. 2023, United Nations 2022). However, potential agricultural production is challenged by climate-change-driven biotic and abiotic stresses (Kumar et al. 2022). To meet the Sustainable Development Goal 2: Zero Hunger (SDG2), we need to boost crop yield by about 70%*[1]. For this, we need crops with adaptive capacities to changing environments. Domesticated crops have been under human selection pressure for ages and their gene pool is limited by the domestication bottleneck (Tanksley and McCouch 1997). To broaden their genetic diversity, valuable genetic resources can be found within crop wild relatives (CWR).

CWR are wild plant species closely related to cultivated crops. Broadly, they encompass all wild plants within the same genus as the crop (Maxted et al. 2006). CWR constitutes about 21% of the world's flora (Maxted and Kell 2009). CWR have survived in nature enduring various selection pressures, both biotic and abiotic. Consequently, they harbour novel genetic resources that can play pivotal roles in crop improvement efforts.

Currently, two prominent challenges hinder the utilisation of CWR. Firstly, plant breeders often depend on their established breeding lines, and the potential contributions of CWR is not investigated well. Secondly, there exists a notable absence of user-friendly tools for effective utilisation.

Plant breeders typically depend on the vast collections of plant genetic resources gathered (Bioversity International 2019, Loskutov 1999) and conserved ex-situ in several gene banks (FAO 2010). Numerous methodologies have been developed to systematically identify accessions possessing various traits from these collections. One of the earliest methods was the "core collection concept" (Frankel 1984), which aimed to characterize the entire accessions to create minimally redundant subsets to capture maximum genetic diversity with fewer samples. Initially, around 10% of accessions underwent field trials against various stresses (Frankel 1984, Brown 1989). However, for crops with extensive collections, this approach became impractical, leading to the development of the "mini-core collection" where only 10% of the core collection was evaluated (Upadhyaya and Ortiz 2001, Upadhyaya et al. 2013), leaving most collections untested.

To address this challenge, the FIGS ("Focused Identification of Germplasm Strategy") tool was introduced, building upon earlier work by Michael Mackay (Mackay 1985, Caradus et al. 1990). FIGS employs two main approaches: "FIGS filtering," which filters accessions based on expert knowledge and environmental data (Bouhssini et al. 2009), and "FIGS modelling," which predicts trait presence in uncharacterized accessions using field trial data (Sunitha et al. 2023). All these methods primarily serve to filter collections stored in gene banks.

For CWR, both collections and field evaluation data are scarce. To address this challenge, we are introducing the MoDGP (“Modelling the Germplasm of Interest”) tool in the CWR pDT. MoDGP leverages species distribution modelling, relying on occurrence data of CWR to produce habitat suitability maps, establish mathematical correlations between adaptive traits and environmental factors, and facilitates mapping geographic areas where populations possessing germplasms resilient against various biotic and abiotic stresses are potentially growing (Fig. 1).

Objectives

The main objective of the CWR pDT is to streamline the identification and utilization of novel genetic resources from CWR through automating data flow, automated modelling runs, uncertainty analysis, and timely alerts on potential genetic resources of interest for plant breeders, policymakers, and conservation scientists. Our objective includes the

creation of habitat suitability maps for all CWR with sufficient occurrence data, accessible via an intuitive graphical user interface implemented with the R shiny framework. Our model is designed to be adaptable across different crop species and traits, empowering users to address key research questions in pre-breeding, such as identifying geographic areas where populations of CWR harbouring beneficial genetic traits for enhancing crop resilience to environmental stresses are potentially growing. Additionally, the pDT facilitates the assessment of gaps in existing collection efforts, aiding in the strategic planning of future genetic resource collections.

Workflow

The workflow of the CWR pDT includes automated access of occurrence and environmental data, automated model runs to generate habitat suitability maps for CWR via an ensemble modelling technique to predict and map stress-tolerant populations of CWR for use in breeding programs (Fig. 1). Additionally, the pDT incorporates a graphical user interface to facilitate end-users' interaction with the outputs of the pDT.

Data

MoDGP relies on two types of data as input. Firstly, occurrence data from GBIF (CIAT 2024), with plans to expand sources to include ICARDA, Genesys PGR, EURISCO, RAINBIO, and more (Table 1). Genesys, a global gene bank ex-situ conserved data hub, not only provides occurrence data but also serves as a valuable source of crop trait information. RAINBIO contains georeferenced occurrences, particularly from sub-Saharan tropical Africa, which can be filtered for CWR data. Other data sources are listed in Table 1. Secondly, environmental variables such as climate (bioclimate data), soil, and topographic data are utilized as predictor variables in raster format. We use climate data from ERA5, soil data from SoilGrids, and elevation data from SRTM DEM (Table 1). At each occurrence point for each CWR species, values of environmental variables are extracted and prepared as input for MoDGP.

Model

MoDGP uses different high performing species distribution modelling algorithms such as generalized additive modelling (GAM; Wood 2011), generalized boosted regression modelling (gbm; Greg et al. 2024), and maximum entropy modelling (MaxEnt; Phillips et al. 2006) to produce habitat suitability maps of model targets. The algorithms in MoDGP function by relating occurrence points to environmental variables to produce habitat suitability maps.

We aim to run models for all CWR with unique occurrence data exceeding 40. To represent the absence data, we identify 10,000 points where other species of the same genus are

present, but the model target is absent. These points are chosen within a buffer area of 15 km from known presence points.

To mitigate multicollinearity, we stack all predictor variables and extract their values at both the presence and absence points. Then, we compute Pearson's pairwise correlations and from variables exhibiting a correlation coefficient exceeding |0.8|, only one variable with the lowest variable inflation factor is selected for model runs. Each model is replicated twice using two methods: bootstrapping and substitution. In each replication, 75% of the data are allocated for training, with the remaining used for evaluation. Consequently, we generate 12 habitat suitability maps for each species as three algorithms replicated twice employing two replication methods.

Results from all algorithms are evaluated against test data using area under the ROC curve (AUC) and True Skill Statistics (TSS). Maps from less performing models i.e. with AUC < 0.7 and/or TSS < 0.4 are dropped and only maps from high performing algorithms and models settings are kept.

The selected maps are combined through an ensemble approach and binary maps are produced using the maximum sum sensitivity and specificity threshold to distinguish between suitable and non-suitable pixels. Values of abiotic stresses are extracted from suitable pixels, and range of tolerance to stress factors are generated as response curves. CWR of a given crop are ranked based on their range of tolerances to stress factors. For model targets with high tolerance to these factors, geographic areas where plants presenting the desired genotypes are potentially growing will be mapped and provided.

FAIRness

We will comprehensively document the entire workflow, spanning from the initial input data through each processing step and modelling, culminating in the generated output. We will ensure that the occurrence data utilized for modelling is referenced using persistent identifiers whenever feasible. Additionally, references to climate, soil, and topographic data will be provided. All data employed in the models will be made publicly accessible and free for sharing and usage, with appropriate acknowledgment. The outputs from pDT and the modelling tools utilized to generate these outputs will also be openly available to the public as FAIR Digital Objects (FDOs, Wittenburg et al. 2023).

FDOs integrate persistent identifiers and structured metadata to enable cross-domain interoperability, crucial for platforms like the European Open Science Cloud (EOSC*[2]), aligning with FAIR principles emphasizing machine actionability (Jacobsen et al. 2020, European Commission 2018). Leveraging an ecosystem of services and registries, including RO-Crate for packaging (Sefton et al. 2023) and schema.org/Bioschemas for structured metadata (Gray et al. 2017), we implement web-compliant FDOs. This approach aids integration with European initiatives like the European Green Deal*[3], utilizing two FDO types to describe computational workflows and capture FAIR data from simulations (Fig. 2).

All developed model codes and scripts will be published as open source in the BioDT repository on GitHub (https://github.com/BioDT).

Performance

CWR pDT aims to run tens of thousands of CWR species using different algorithms and model replications. This is an ideal case for utilizing parallel processing as the different model runs are independent. In preparation for parallelizing the execution, the R environment has been containerized with Docker and the container image can be pulled and executed on the CPU partition of the LUMI supercomputer through Apptainer / Singularity and on a cloud through Docker. Initial tests have been run on LUMI-C with this setup, but the parallelization scheme is not fully implemented yet. The large parallel computing capacity of LUMI-C is expected to be highly suitable for achieving the aimed large scale model processing. In case of smaller workloads, the containerised solution is directly executable also on cloud environments.

Interface and outputs

To provide the best experience of interaction with pDT for multiple end-user groups such as pre-breeders, researchers, conservation scientists, and academicians, we are developing a web interface based on the R Shiny (https://rstudio.github.io/shiny/authors.html) application. The interface will feature dropdown menus for crops and their corresponding

1. wild relatives,
2. habitat suitability maps, and
3. abiotic stress ranges among others.

This will allow users to effectively map the optimal overlap between environmental stress factors and habitat suitability to identify geographic areas where populations resilient to stresses can potentially thrive.

End users can collect samples from mapped areas of interest and test the performances of the genotypes. The user interface also enables users to constrain or relax the tolerance thresholds and decide geographical areas from which the germplasm of interest can be obtained. It should also enable them to prioritize the germplasms to be tested based on quality and/or access. Distribution models capture potentially suitable habitats and thus may help the discovery of new populations and identify gaps in collection efforts or ex-situ conservation. With improvements in online occurrence data, the validity of models can also improve over time improving the robustness of the models. The modelling tools will also be made available to users.

Integration and sustainability

To ensure the long-term availability and accessibility of the pDT CWR, a pilot for the integration into the Big Data processing services of the Destination Earth Data Lake (DEDL, Duatis Juarez et al. 2023) is under development together with the platform operator EUMETSAT*[4].

A major objective of the pilot study is the implementation of data pipelines between DEDL as a data aggregator, processing platform and provider of earth observation data and the pDT CWR which will serve as a blueprint to facilitate the integration of more Digital Twins into DestinE's core infrastructures. Comprehensive mappings between BioDT's core semantic artefacts such as schema.org/Bioschemas (fundamental for RO-Crate) and specifications used in DEDL such as SpatioTemporal Asset Catalogs (STAC*[5]) will be provided as FAIR Semantic Mappings to foster the reusability of all resulting data products (Broeder et al. 2021), and subsequently mobilized through BioDT's mapping tool mapping.bio (Wolodkin et al. 2023).

Application and impact

While plant breeders often rely on their breeding lines and landraces, CWR offer not only vast diversity but have also undergone several (and ongoing) selection pressures and thus encompass novel genetic resources. Representing approximately 21% of the plant kingdom (Maxted et al. 2006), assuming that a third of them have adequate occurrence data available, we here aim to provide outputs for roughly 7% of the plant kingdom, equivalent to around 26,600 plant species. Different populations of these species exhibit adaptations to various crop stresses. The CWR pDT makes this abundant resource accessible through a graphical user interface, allowing plant breeders to choose among several populations of the 26,600 species.

The suitability maps produced by pDT serve diverse purposes, including in-situ conservation, restoration, ex-situ conservation, and seed collection gap analysis. As the pDT is envisioned to operate automatically on a regular basis, its results are continuously updated, offering real-time outputs. These outputs are available globally and can be tailored to match different geographic scales, from country to continental levels.

In general, applications and impacts of the pDT can fall into two categories:

1. **Climate change adaptation:** Plant breeders can utilize the pDT to map populations of CWR possessing novel genetic resources, aiding in the development of crops with high resilience to stresses induced by climate change.
2. **Conservation:** By identifying geographic regions hosting populations of CWR with adaptive traits, the tool facilitates targeted conservation efforts, thereby aiding in the conservation of genetic diversity. The CWR pDT aslo plans to integrate ecogeographic land characterization (ELC) maps via the CAPFITOGEN tool (Parra

Quijano et al. 2021). These maps illustrate adaptive scenario classes that can be overlaid onto protected areas to assess conservation of diverse adaptive trait populations. Moreover, the maps facilitate gap analysis in ex-situ gene banks, thereby improving both ex-situ and in-situ conservation efforts.

Acknowledgements

This study has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101057437 (BioDT project, https://doi.org/10.3030/101057437). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

We acknowledge the EuroHPC Joint Undertaking and CSC – IT Center for Science, Finland for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC – IT Center for Science and the LUMI consortium, through Development Access calls.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Bioversity International (2019) Bioversity Collecting Mission Database. Version 1.10. Global Biodiversity Information Facility https://doi.org/10.15468/ulk1iz
- Bouhssini ME, Street K, Joubi A, Ibrahim Z, Rihawi F (2009) Sources of wheat resistance to Sunn pest, Eurygaster integriceps Puton, in Syria. Genet Resour Crop Evol 56: 1065-1069. https://doi.org/10.1007/s10722-009-9427-1
- Broeder D, Budroni P, Degl'Innocenti E, Le Franc Y, Hugo W, Jeffery K, Weiland C, Wittenburg P, Zwolf CM (2021) SEMAF: A Proposal for a Flexible Semantic Mapping Framework. https://doi.org/10.5281/zenodo.4651420
- Brown AH (1989) Core collections: a practical approach to genetic resources management. Genome 31 (2): 818-824. https://doi.org/10.1139/g89-144
- Caradus J, Forde M, Wewala S, Mackay AC (1990) Description and classification of a white clover (Trifolium repens L.) germplasm collection from southwest Europe. New Zealand Journal of Agricultural Research 33 (3): 367-375. https://doi.org/10.1080/00288233.1990.10428433
- CIAT (2024) A global database for the distributions of crop wild relatives. Version 1.13. Crop Wild Relatives Occurrence data consortia, Centro Internacional de Agricultura Tropical - CIAT. The Global Biodiversity Information Facility. https://doi.org/10.15468/dl.nt55b5

- Duatis Juarez J, Schick M, Puechmaille D, Stoicescu M, Saulyak B (2023) Destination Earth Data Lake. EGU General Assembly 2023, Vienna, Austria, 24–28 Apr 2023. https://doi.org/10.5194/egusphere-egu23-7177
- European Commission (2018) Directorate-General for Research and Innovation, Turning FAIR into reality - Final report and action plan from the European Commission expert group on FAIR data. European Commission, Publications Office. https://doi.org/10.2777/1524
- FAO (2010) The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture, Rome. Food and Agriculture Organization of the United Nations. URL: https://www.fao.org/agriculture/crops/thematic-sitemap/theme/seeds-pgr/sow/sow2/en/
- Frankel O (1984) Genetic Perspectives of Germplasm Conservation. In: Arber WK, Llimensee K, Peacock WJ, Stralinger P (Eds) Genetic Manipulation: Impact on Man and Society. Cambridge University Press, Cambridge.
- Gray AJ, Goble C, Jimenez RC (2017) Bioschemas: From Potato Salad to Protein Annotation. 16th International Semantic Web Conference (ISWC 2017), RWTH Aachen University, October 23rd to 25th, 2017. CEUR, Vienna. Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), 4 pp. URL: https://ceur-ws.org/Vol-1963/paper579.pdf
- Greg R, Edwards D, Kriegler B, Schroedl S, Southworth H, Greenwell B, Boehmke B, Cunningham J, GBM Developers (2024) Gbm: Generalized Boosted Regression Models . 2.1.9. R CRAN. URL: https://cran.r-project.org/web/packages/gbm/
- Jacobsen A, Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo CT, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery KG, Kaliyaperumal R, Kersloot MG, Kirkpatrick CR, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, Reisen M, Rocca-Serra P, Pergl R, Sansone S-, Silva Santos LO, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson MD, Willighagen EL, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2 (1-2): 10-29. https://doi.org/10.1162/dint_r_00024
- Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR (2019) Sharing interoperable workflow provenance: A review of best practices and their practical application in CWLProv. GigaScience 8 (11). https://doi.org/10.1093/gigascience/giz095
- Kumar L, Chhogyel N, Gopalakrishnan T, Hasan MK, Jayasinghe SL, Kariyawasam CS, Kogo BK, Ratnayake S (2022) Climate change and future of agri-food production. In: Bhat R (Ed.) Future Foods.
- Loskutov IG (1999) Vavilov and his Institute. A history of the wold collection of plant genetic resources in Russia. International Plant Genetic Resources institute https://doi.org/10.13140/2.1.2632.0644
- Mackay M (1985) Maintaining Genetic Diversity in Germplasm Collections. BioScience 35 (10): 582-588.
- Maxted N, Ford-Lloyd BV, Jury S, Kell S, Scholten M (2006) Towards a definition of a crop wild relative. Biodiversity & Conservation 15 (8): 2673-268. https://doi.org/10.1007/s10531-005-5409-6

- Maxted N, Kell SP (2009) Establishment of a global network for the in situ conservation of crop wild relatives: status and needs. FAO Commission on Genetic Resources for Food and Agriculture, 112 pp. URL: https://www.fao.org/3/i1500e/i1500e18a.pdf
- Parra Quijano M, Iriondo J, Torres M, López F, Phillips J, Kell S (2021) Capfitogen 3: a toolbox for the conservation and promotion of the use of agricultural biodiversity. 3. Bogotá: Universidad Nacional de Colombia. Facultad de Ciencias Agrarias.. URL: https://www.capfitogen.net/en/access/capfitogen3-local-mode/
- Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4): 231-259. https://doi.org/10.1016/j.ecolmodel.2005.03.026
- Roser MRH, Ortiz-Ospina E, Rod&eacute s (2023) World Population Growth - Our World in Data. https://ourworldindata.org/world-population-growth. Accessed on: 2024-3-15.
- Sefton P, Ó Carragáin E, Soiland-Reyes S, Corcho O, Garijo D, Palma R, Coppens F, Goble C, Fernández J, Chard K, Gomez-Perez JM, Crusoe M, Eguinoa I, Juty N, Holmes K, Clark J, Capella-Gutierrez S, Gray AG, Owen S, Williams A, Tartari G, Bacall F, Thelen T, Ménager H, Rodríguez-Navas L, Walk P, Whitehead B, Wilkinson M, Groth P, Bremer E, Castro LJ, Sebby K, Kanitz A, Trisovic A, Kennedy G, Graves M, Koehorst J, Leo S, Portier M, Brack P, Ojsteršek M, Droesbeke B, Niu C, Tanabe K, Miksa T, La Rosa M, Decruw C, Czerniak A, Jay J, Serra S, Siebes R, de Witt S, El Damaty S, Lowe D, Li X, Gundersen S, Radifar M (2023) RO-Crate Metadata Specification 1.1.3. https://doi.org/RO-CrateMetadataSpecification1.1.3
- Sunitha NC, Prathibha MD, Thribhuvan R, Lokeshkumar BM, Basavaraj PS, Lohithaswa HC, Anilkumar C (2023) Focused identification of germplasm strategy (FIGS): a strategic approach for trait-enhanced pre-breeding. Genetic Resources and Crop Evolution 71 (1): 1-16. https://doi.org/10.1007/s10722-023-01669-7
- Tanksley SD, McCouch (1997) Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. Science 277 (5329): 1063-1066. https://doi.org/10.1126/science.277.5329.1063
- United Nations (2022) World Population Prospects 2022: Summary of Results. United Nations Department of Economic and Social Affairs, Population Division URL: https://www.un.org/development/desa/pd/content/World-Population-Prospects-2022
- Upadhyaya HD, Ortiz R (2001) A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theoretical and Applied Genetics 102 (8): 1292-1298. https://doi.org/10.1007/s00122-001-0556-y
- Upadhyaya HD, Wang Y-, Gowda CL, Sharma S (2013) Association mapping of maturity and plant height using SNP markers with the sorghum mini core collection. Theoretical and Applied Genetics 126 (8): 2003-2015. https://doi.org/10.1007/s00122-013-2113-x
- Wittenburg P, Schwardmann U, Blanchi C, Weiland C (2023) FDOs to Enable Cross-Silo Work. In: Sure-Vetter Y, Goble C (Eds) *1st Conference on Research Data Infrastructure (CoRDI) - Connecting Communities*. 12 – 14 September 2023. Karlsruhe https://doi.org/10.52825/cordi.v1i.263
- Wolodkin A, Weiland C, Grieb J (2023) Mapping.bio: Piloting FAIR semantic mappings for biodiversity digital twins. Biodiversity Information Science and Standards 7 (111979). https://doi.org/10.3897/biss.7.111979

- Wood SN (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society: Series B: Statistical Methodology 73 (1): 3-36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Endnotes

*1 https://www.un.org/sustainabledevelopment/hunger/

*2 https://eosc-portal.eu/

*3 https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

*4 https://www.eumetsat.int/international-cooperation/destine

*5 https://stacspec.org

Author-formatted, not peer-reviewed document posted on 10/04/2024. DOI:
https://doi.org/10.3897/arphapreprints.e125091

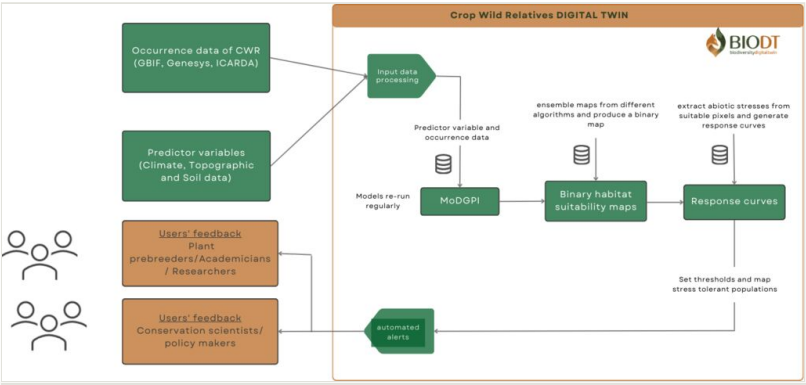

Figure 1.

Simplified workflow of the crop wild relatives prototype digital twin.

Author-formatted, not peer-reviewed document posted on 10/04/2024. DOI:
https://doi.org/10.3897/arphapreprints.e125091

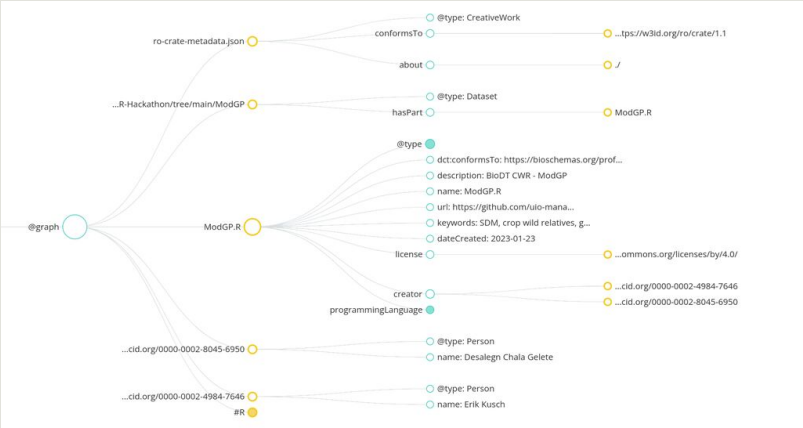

Figure 2.

Actual outline of data model employing the RO-Crate approach for workflow preservation and aggregation (Khan et al. 2019) represented as information nodes in a directed graph using machine interpretable semantic artefacts such as schema.org (e.g. http://schema.org/Dataset) as well as PIDs such as ORCID (https://orcid.org/).

Author-formatted, not peer-reviewed document posted on 10/04/2024. DOI: https://doi.org/10.3897/arphapreprints.e125091

Table 1.

Data and data sources for the crop wild relatives prototype digital twin.

Data type	Source	Webpage	Remarks
Species occurrence/ trait data	Global Biodiversity Information Facility(GBIF)	https://www.gbif.org	A global species occurrences data portal (> 2.6 billion; March 2024)
	Genesys PGR	Genesys PGR (genesys-pgr.org)	Genesys is an online platform where you can find information about Plant Genetic Resources for Food and Agriculture (PGRFA) conserved in genebanks worldwide.
	International Center for Agricultural Research in the Dry Areas (ICARDA)	https://www.icarda.org/	Usually share data with Genesys on annual basis
	RAINBIO database	https://gdauby.github.io/rainbio/index.html	Contains georeferenced occurrences of vascular plants from sub-Saharan tropical Africa.
	EURISCO crop specimens	https://eurisco.ipk-gatersleben.de/	PGRFA data portal for European genebanks
	Global Crop Wild Relative atlas	https://www.cwrdiversity.org/	Global catalog of crop wild relatives
	Plant trait database (TRY)	TRY Plant Trait Database (try-db.org)	TRY focuses on plant traits. CWR with short generation time such as herbs are particularly suitable for breeding, and the database holds remarkable importance for CWR pDT.
	NordGen Nordic catalog	https://nordic-baltic-genebanks.org/gringlobal/search.aspx	Nordic genebank PGRFA data portal
	NordGen Nordic CWR checklist	https://doi.org/10.15468/itkype	Nordic checklist of crop wild relative species

Climate	ERA5	https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels	ERA5 is a global climate reanalysis dataset produced by the European Centre for Medium-Range Weather Forecast. It simulates climate data relying on hourly weather data, offering dynamic data unlike many other climate data repositories, rendering it well-suited for biodiversity with digital twins.
Edaphic	Soil grids	https://soilgrids.org/	SoilGrids is a dataset that provides global predictions for soil properties at different depths (0-5cm, 5-15cm, 15-30cm, and 30-60cm) with a spatial resolution of 250 meters. These properties include organic carbon, pH, sand, silt, and clay fractions, among others. The dataset is built using machine learning techniques and is based on a compilation of soil samples from various sources.
Topographic	SRTM DEM	USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global | U.S. Geological Survey	The SRTM DEM is available at 90 m resolution globally