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[‡]

‡ Natural History Museum, University of Oslo, Oslo, Norway
§ Senckenberg – Leibniz Institution for Biodiversity and Earth System Research, Science IT, 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)

Academic editor: Volker Grimm

Abstract

Amidst population growth and climate-driven crop stresses such as drought, extreme weather, fungal and insect pests, as well as various crop diseases, ensuring food security demands innovative strategies. Crop wild relatives (CWR), wild plants in the same genus as the crop as well as wild populations belonging to the same species as the crop, offer novel genetic resources crucial for enhancing crop resilience against these stress factors. Here, we introduce a prototype digital twin (pDT) to aid in searching and utilising CWR genetic resources. Using the MoDGP (Modelling 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 2023). However, potential agricultural production is challenged by climate change driven biotic stresses such as drought, extreme weather, soil acidity and mineral deficiencies, as well as biotic stresses, including fungal and insect pests and various crop diseases (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). The category also includes wild populations of the same species as the cultivated crops. 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 in breeding programmes. 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 (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 characterise 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. 2012). 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 the presence of genetic resource of interest in uncharacterised 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, such as tolerance to drought and pathogens and environmental factors and facilitates mapping geographic areas where populations possessing genetic resources for resilience against various biotic and abiotic stresses are potentially growing.

Objectives

The main objective of the CWR pDT is to streamline the identification and utilisation 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, policy-makers 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 resources for enhancing crop resilience to environmental stresses are potentially growing. Additionally, in the pDT, we are developing ecogeographic land characterisation (ELC) maps to identify ELC classes that are under-represented in ex-situ seed collections. This will help to assess gaps in current collection or ex-situ conservation 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 programmes (Fig. 1, see also the model section). Additionally, the pDT incorporates a graphical user interface to facilitate end-users' interaction with the outputs of the pDT. The pDT is automated to re-run once in a year depending on availablity updates in occurrence data.

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 utilised 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 generalised additive modelling (GAM; Wood (2010)), generalised 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 (crops and crop wild relatives). 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 or not recorded. 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 being selected for model runs. Each model is replicated twice using two methods: bootstrapping and substitution of 75% of the data. In each replication, 75% of the data are randomly 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 < 0.6 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 the range of tolerance to these 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 utilised 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 acknowledgement. The outputs from pDT and the modelling tools utilised 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 emphasising machine-actionability (European commission 2018Jacobsen et al. 2020). We are buiding on the RO-Crate approach (Soiland-Reyes et al. 2022) to implement lightweight packaging of the pDT's model description and output together with rich metadata. Structured metadata are provided by Schema.org and its Bioschemas extension (Gray et al. 2017) to facilitate both readability of data packages by humans and processability (i.e. machine-actionability, Weiland et al. (2022)) by software agents. In this way, RO-Crate opens up an implementation path for web-based or "webby" FDOs and enables mobilisation and reuse of the pDT CWR across the Destination Earth framework. This approach aids integration with European initiatives like the European Green Deal*[3], utilising 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 highly suitable for utilising parallel processing as the different model runs are independent. In preparation for executing the operation in parallel, the R environment has been containerised 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 parallelisation scheme is not fully implemented yet. The large parallel computing capacity of LUMI-C is expected to be advantageous 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 amongst 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 can also enable them to prioritise the populations 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 published in open access journals and 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 Catalogues (STAC*[5]) will be provided as FAIR Semantic Mappings to foster the reusability of all resulting data products (Broeder et al. 2021) and subsequently mobilised 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

amongst several populations of the 26,600 species. The outputs and impacts will grow with enhanced data availability and quality, improving future prospects.

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 re-run automatically on an annual basis, its results are continuously updated, offering real-time outputs. These outputs are available at global scale 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 utilise 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 also plans to integrate ecogeographic land characterisation (ELC) maps via the CAPFITOGEN tool (Parra Quijano et al. 2021). These maps illustrate adaptive scenario classes that can be overlaid on to 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.

Policy implications and recommendations

Crop wild relatives play a critical role in ensuring food security and agricultural resilience in the face of environmental challenges. However, just like other organisms, CWR are facing threats from climate change (Jarvis et al. 2008) and land-cover/land-use changes (Maxted et al. 2012). CWR are also data deficient and less represented in gene-bank collections showing less attention is given to both in-situ and ex-situ conservations.

To enhance the conservation and utilisation of CWR genetic resources, it is imperative to strengthen data management and collaboration amongst relevant stakeholders. Drawing from the recommendations by Arnaud et al. (2017) and the collaboration agreement between GBIF and FAO*[6], policy-makers should prioritise the integration of CWR data into existing platforms, such as GBIF, Genesys, EURISCO and FAO PlantTreaty. This entails enhancing data fitness for use in agrobiodiversity through quality standards outlined in the GBIF FAO collaboration agreement. Additionally, the establishment of a dedicated monitoring directive for CWR can streamline efforts in monitoring and managing CWR populations across Europe and beyond, ensuring their long-term conservation.

Moreover, in-situ conservation efforts for CWR should be supported through coordinated actions at the local, national and regional levels. Taking existing efforts, such as the Nordic CWR policy report and regional approach advocate (Fitzgerald et al. 2019), policy-makers should prioritise the development and implementation of comprehensive

conservation plans tailored to regional contexts. This includes using genetic reserves adhering to quality standards to ensure effective conservation outcomes (Iriondo et al. 2012). Furthermore, collaboration with initiatives like Biodiversa+ and EIONET can facilitate funding and monitoring programmes for CWR conservation. By adhering to detailed in-situ conservation guidelines, policy-makers can strengthen the resilience of agricultural systems and safeguard the invaluable genetic diversity harboured by CWR populations.

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.

We also thank Taimur Khan, Ingolf Kuhn, Jan Dick and one anonymous reviewer for reviewing and providing constructive comments, which have significantly improved the paper.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Arnaud E, Castañeda-Álvarez NP, Cossi JG, Endresen D, Jahanshiri E, Vigouroux Y (2017) Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity. https://www.gbif.org/document/82283/report-of-the-task-group-on-gbif-data-fitness-for-use-in-agrobiodiversity. Accessed on: 2024-2-10.
- Bouhssini ME, Street K, Joubi A, Ibrahim Z, Rihawi F (2009) Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. Genetic Resources and Crop Evolution 56 (8): 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. Zenodo https://doi.org/10.5281/zenodo.4651420
- Brown AHD (1989) Core collections: a practical approach to genetic resources management. Genome 31 (2): 818-824. https://doi.org/10.1139/g89-144

- Caradus JR, Forde MB, Wewala S, Mackay AC (2012) 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) 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/
- Fitzgerald H, Palmé A, Asdal Å, Endresen D, Kiviharju E, Lund B, Rasmussen M, Thorbjörnsson H, Weibull J (2019) A regional approach to Nordic crop wild relative *in situ* conservation planning. Plant Genetic Resources: Characterization and Utilization 17 (2): 196-207. https://doi.org/10.1017/s147926211800059x
- 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
- Gray AJ, Goble C, Jimenez RC (2017) Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), 4 pp. Bioschemas: From Potato Salad to ProteinAnnotation. 16th International Semantic Web Conference (ISWC 2017), RWTH AachenUniversity, October 23rd to 25th, 2017., CEUR, Vienna. 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/
- Iriondo J, Maxted N, Kell S, Ford-Lloyd B, Lara-Romero C, Labokas J, Brehm J (2012) Quality standards for genetic reserve conservation of crop wild relatives . In: Maxted N, Dulloo M, Ford-Lloyd B, Germany LF, Iriondo J, Carvalho MPd (Eds) Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces. CABI Books https://doi.org/10.1079/9781845938512.0000
- Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, Courtot M, Crosas M, Dumontier M, Evelo C, Goble C, Guizzardi G, Hansen KK, Hasnain A, Hettne K, Heringa J, Hooft RW, Imming M, Jeffery K, Kaliyaperumal R, Kersloot M, Kirkpatrick C, Kuhn T, Labastida I, Magagna B, McQuilton P, Meyers N, Montesanti A, van Reisen M, Rocca-Serra P, Pergl R, Sansone S, da Silva Santos LOB, Schneider J, Strawn G, Thompson M, Waagmeester A, Weigel T, Wilkinson M, Willighagen E, Wittenburg P, Roos M, Mons B, Schultes E (2020) FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2: 10-29. https://doi.org/10.1162/dint_r_00024

- Jarvis A, Lane A, Hijmans RJ (2008) The effect of climate change on crop wild relatives. Agriculture, Ecosystems & Environment 126 (1): 13-23. https://doi.org/10.1016/j.agee.2008.01.013
- 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. https://doi.org/10.1016/B978-0-323-91001-9.00009-8
- 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 B, Jury S, Kell S, Scholten M (2006) Towards a definition of a crop wild relative. Biodiversity and Conservation 15 (8): 2673-2685. 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 Agriculture112. URL: https://www.fao.org/3/i1500e/i1500e18a.pdf
- Maxted N, Kell S, Ford-Lloyd B, Dulloo E, Toledo Á (2012) Toward the Systematic Conservation of Global Crop Wild Relative Diversity. Crop Science 52 (2): 774-785. https://doi.org/10.2135/cropsci2011.08.0415
- 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/. URL: https://www.capfitogen.net/en/access/capfitogen3-local-mode/
- Phillips S, Anderson R, Schapire R (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling 190: 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.
- Soiland-Reyes S, P. Sefton MC, Castro LJ, Coppens F, Fernández JM, Garijo D, Grüning B, Rosa ML, S. Leo EÓC, Portier M, Trisovic A, Community RO, Groth P, Goble C (2022) Packaging research artefacts with RO-Crate. Data Science 5: 97-138. https://doi.org/10.3233/DS-210053
- 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 S, McCouch S (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 (2023) World Population Prospects 2022: Summary of Results. United Nations Department of Economic and Social Affairs, Population Division . https://www.un.org/development/desa/pd/content/World-Population-Prospects-2022

- Upadhyaya H, Wang Y, Gowda CLL, 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
- 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
- Weiland C, Islam S, Broder D, Anders I, Wittenburg P (2022) FDO Machine Actionability. Zenodo https://doi.org/10.5281/zenodo.7825649
- Wittenburg P, Schwardmann U, Blanchi C, Weiland C (2023) FDOs to Enable Cross-Silo Work. Proceedings of the Conference on Research Data Infrastructure 1 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 https://doi.org/10.3897/biss.7.111979
- Wood S (2010) 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

*6 Plant Treaty Global Information System Scientific Advisory Committee meeting May 2023 - GBIF Norway - Global Biodiversity Information Facility

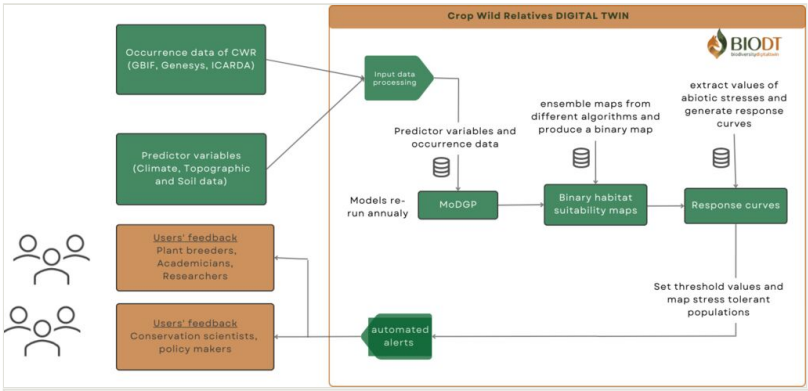

Figure 1.

Simplified workflow of the crop wild relatives prototypes digital twin. CWR - crop wild relatives; GBIF - Global Biodiversity Information Facility; Genesys - Global Information System on Plant Genetic Resources; ICARDA - International Center for Agricultural Research in the Dry Areas; MODGP - modelling the distribution of germplasms of interest.

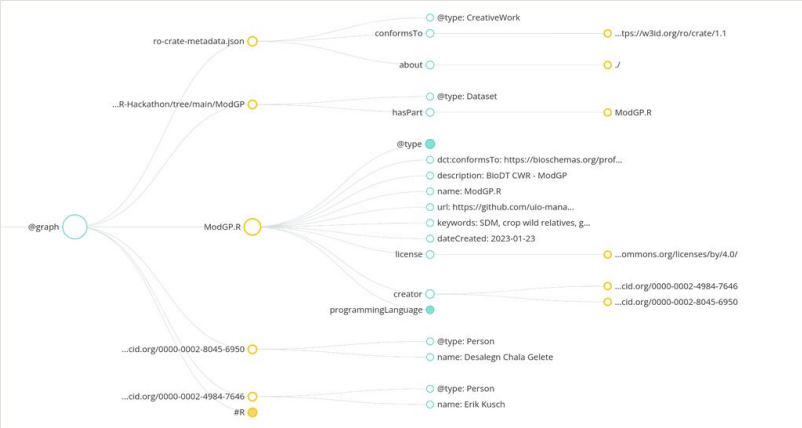

Figure 2.

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/).

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 gene-banks 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 gene-banks.
	Global Crop Wild Relative atlas	https://www.cwrdiversity.org/	Global catalogue 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 catalogue	https://nordic-baltic-genebanks.org/gringlobal/search.aspx	Nordic gene-bank 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-Lnad is a global climate re-analysis dataset produced by the European Centre for Medium-Range Weather Forecast. It simulates climate data using hourly weather information, providing dynamic data unlike many other climate data sources. This allows users to recompute climate data by incorporating the most recent weather updates.
Edaphic	Soil grids	https://soilgrids.org/	SoilGrids is a dataset that provides global map for soil properties at different depths (0-5 cm, 5-15 cm, 15-30 cm and 30-60 cm) with a spatial resolution of 250 m. These properties include organic carbon, pH, sand, silt and clay fractions, amongst 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.