Data in Brief 35 (2021) 106781

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# Open Data to Support Agricultural Diversification (version October 2020)

Nur Marahaini Mohd Nizar[a], Ebrahim Jahanshiri[a,*], Siti Sarah Mohd Sinin[a], Eranga M. Wimalasiri[a], Tengku Adhwa Syaherah Tengku Mohd Suhairi[a], Peter J. Gregory[a,b], Sayed N. Azam-Ali[a]

[a] *Crops For the Future UK, Chelmsford, Essex CM2 7PJ, England, UK*
[b] *School of Agriculture, Policy & Development, University of Reading, Earley Gate, Reading RG66AR, UK*

ARTICLE INFO

*Article history:*
Received 2 December 2020
Revised 4 January 2021
Accepted 18 January 2021
Available online 21 January 2021

*Keywords:*
Neglected and underutilised crops
Agro-ecological data
Potential yield
Crop nutrition composition

ABSTRACT

Following the development of a database that was specifically designed to store value chain information, particularly for underutilised crops, this article describes the data that are currently stored in the database and accessible through its web portal. The data includes various datasets on utilisation status, agro-ecological requirements and season lengths, potential yield and nutritional composition of crops. The data are stored in the form of tables with fixed data elements (column attributes). This article outlines the standard procedures (SOPs) that were developed in-house for data collection, metadata creation and data curation. These processes were used to ensure the quality and reusability of the data that is made available publicly through the database interface. Various statistics and example visualisations are provided to demonstrate the significance of such data for developing solutions for sustainable agricultural diversification.

DOI of original article: 10.1016/j.compag.2020.105920

[*] Corresponding author.
*E-mail addresses:* e.jahan@cropsforthefutureuk.org, ej@cropbase.co.uk (E. Jahanshiri).

https://doi.org/10.1016/j.dib.2021.106781

## Specifications Table

| | |
|---|---|
| Subject | Agricultural and Biological Sciences (General) |
| Specific subject area | Agricultural diversification, leveraging on the potential of neglected and underutilised crops. |
| Type of data | Table, chart |
| How data were acquired | Datasets are a compilation of secondary information that are collected from various sources including books, research articles, databases, website articles, experts and local communities growing underutilised crops, etc. |
| Data format | Tabulated and linked data (secondary data), analysed |
| Parameters for data collection | The focused parameters include agro-ecological requirements and crop season length, potential yield and nutritional composition (both macro- and micro-) and market data of crops at species and limited sub-species level. The compiled data definition can be accessed at the following link: https://doi.org/10.5281/zenodo.3988137 |
| Description of data collection | The primary way to conduct literature review from several sources that could potentially contain the data for needed parameters, following a Standard Operating Procedure. Data was then extracted and stored in a structured format using an online data manager that was developed in-house. |
| Data source location | Institution: Crops for the Future UK (CIC)<br>City/Town/Region: Chelmsford, Essex<br>Country: England, UK<br>The list of primary data sources is extensive* and the sources are mainly databases, research articles, website articles etc. However, large portion of the data were collected from:<br>Crop Ecological Requirements Database (ECOCROP) [1]<br>FoodData Central [2]<br>*Information about all primary data sources are compiled in a “Metadata table” of the database and can be accessed here: https://cropbase.co.uk/cropbasev5/metadata.php |
| Data accessibility | Repository name: Global Knowledge Base for Underutilised Crops<br>Direct URL to data portal:<br>https://cropbase.co.uk/cropbasev5/<br>Link to database code and data used to make figures in this article:<br>https://github.com/CFFRC-KST/CFFRC-Global-Knowledge-Base<br>link to DOI of data:<br>https://doi.org/10.5281/zenodo.3735694 |
| Related research article | Mohd Nizar, N. M., Jahanshiri, E., Tharmandram, A. S., Salama, A., Mohd Sinin, S. S., Abdullah, N. J., Zolkepli, H., Wimalasiri, E. M., Mohd Suhairi, T. A. S. T., Hussin, H., Gregory, P. J., & Azam-Ali, S. N. (2021). Underutilised crops database for supporting agricultural diversification. *Computers and Electronics in Agriculture, 180*, 105920. https://doi.org/10.1016/j.compag.2020.105920 |

## Value of the Data

- The dataset contains information about the major elements of agricultural species that are collected from various sources. The data can be potentially used to facilitate decision making in regard to diversification of cropping systems with focus on the neglected and underutilised crops.
- The analysis-ready dataset can be used to provide insights for agricultural policy and regulatory communities as well as other stakeholders including growers and scientific community that are interested in the development of alternative and complementary food systems worldwide.
- Linked data that cover different aspects of agricultural species can provide new solutions for the current or future issues such as selecting crops that contain high amount of a nutrient and in the meantime are adaptable to the excess or shortage of water.

- As change in markets and climates forces local communities to develop more sustainable solutions that can withstand future conditions, more interdisciplinary and linked data that cover the value chain of underutilised crops will be required. This species-centric dataset provides a solid basis for complex theories regarding alternative solutions to be tested.
- Data curation and transparent metadata validation processes allows the data to be credible to use in any type of analysis. The online system allows data to be downloaded and re-used for analysis.
- Linking data to other data types including environmental, remote sensing, genomic and nutrition data is straightforward and can be achieved by using location and crop identification numbers (IDs). This allows analysis to be performed for location-specific questions that could potentially lead to solving local problems.

## 1. Data Description

Currently 51 tables exist in the dataset where nine of them hold the focus variables for data collection, 17 are the tables designed for keeping data related to agricultural diversification for future uses and 25 of them are storing open-ended data options. These options include city names, cropping systems, nutrient elements, data accuracy flags, contributors' name, plant parts, uses, etc. In total, there are 422 attribute columns, 121 of which are columns for foreign keys that link the records to other records that exist in other tables. Variable naming follows standard conventions and for each variable, there is a specific definition. All data definitions can be accessed at https://doi.org/10.5281/zenodo.3988137. There are eight tables in the main dataset with the focused variables are:

### 1.1. Crop records

This table holds the basic information which represents the identity of crops i.e. the name, scientific name and family name. Out of 2748 crops recorded in the database, 2492 are the records of species, 125 are varieties, 76 cultivars, 51 subspecies and 4 are landraces. The crops may belong to one of 195 families recorded in the database. Among the crops, 20.9% belong to Leguminosae, 14.3% Poaceae, 5.8% Myrtaceae, 3.1% Compositae, and 2.5% Malvaceae (Fig. 1a).

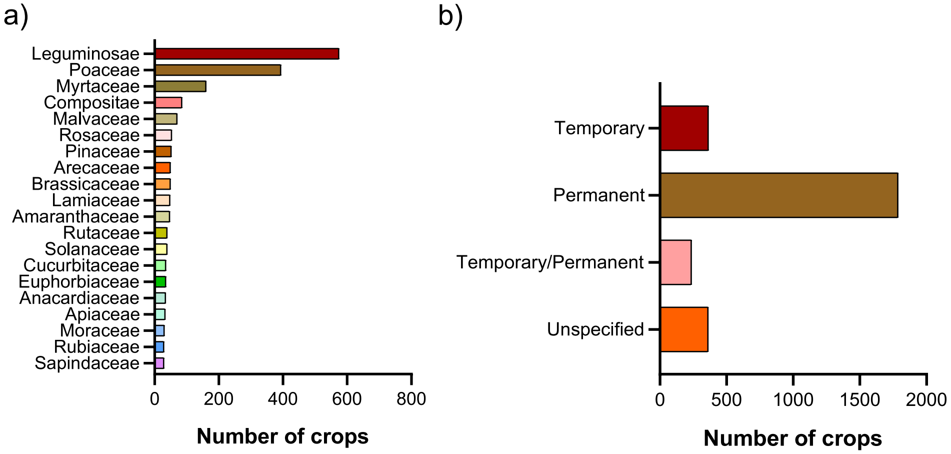

**Fig. 1.** a) Crop family record count, b) crop type record count.

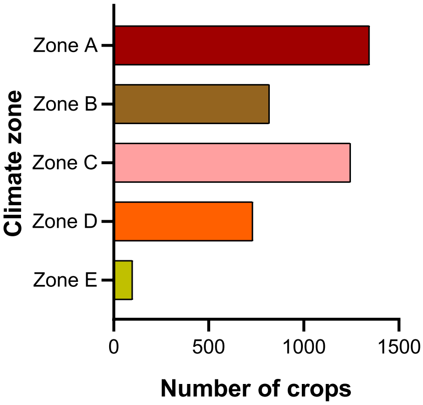

**Fig. 2.** Crops count according to different climate zones.

The rest of the families accounted for less than 2% of the total number of crops. The crops may belong to temporary, permanent or both crop types. Temporary crops are crops which are sown and harvested in the same growing year. Permanent crops are those that are planted once and remain at the planted location after every annual harvest. The terms and definitions are adopted from FAO crop statistics and defintions [3]. There are 364 crops recorded as temporary and 1786 crops recorded as permanent (Fig. 1b). 237 of them can be categorised as both temporary and permanent while the rest remain unspecified.

### 1.2. Agro-ecology

This table stores agro-ecological requirements of the crops where. It has more than 90 agroecological variables including temperature, rainfall, soil depth, soil pH and soil texture requirements of crops. Data on crops per climate zone is also available. We adopted Köppen climate classification which includes Zone A (Tropical), Zone B (Arid), Zone C (Temperate), Zone D (Continental) and Zone E (Polar) and their subcatgories [4]. In the current version of the dataset, most of the recorded crops are adaptable in climate Zones A and C (Fig. 2).

### 1.3. Seasonality

The season length is the duration taken from planting to harvesting for an annual crop or from the first harvest to its next harvest for a perennial crop. The season length comprises both maximum and minimum values and can be used to design location-specific cropping patterns.

### 1.4. Crop classification

The crop classification dataset contains main classification categories. The categories are mainly adapted from FAO Indicative Crop Classification (ICC) [5]. Out of the total records in this table, 19.3% of them are recorded as medicinal, 8.8% vegetables (leafy/stem), 7.5% ornamental/landscape crops, 7.2% fruits, 7% fodder crops, 6.3% forage crops (feed to the animal independently), 3.7% beverage crops, 3.6% oilseed crops and 2.1% fibre crops. Vegetables (root/bulb/tuber) and pesticidal crops both amount to 2.2% from the class (Fig. 3). The rest of the classification accounted for less than 2% of the total number of the class.

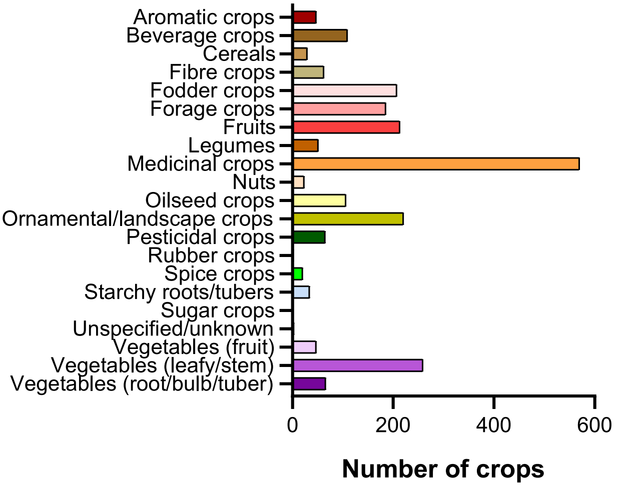

**Fig. 3.** Crops count based on crop classification.

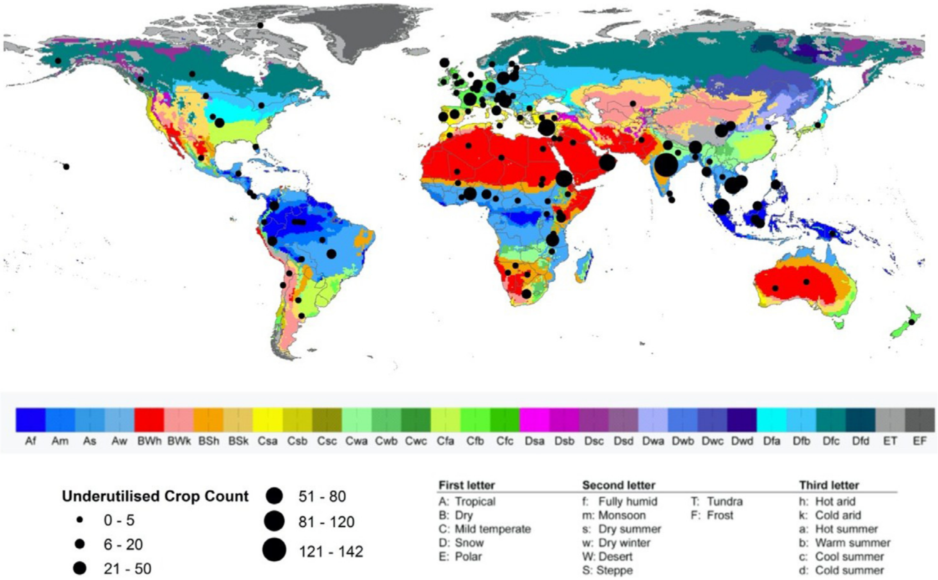

**Fig. 4.** Distribution of underutilised crops worldwide with added Köppen climate classification (obtained from http://hanschen.org).

### 1.5. Crop utilisation status

The current dataset distinguishes the underutilisation status of the crops. The crops will be flagged with the following labels; major crop, underutilised, grown/found in, or unspecified (we were unable to decide on previous categories or no data found in support of other categories) based on a specific location or region. Out of the total records in the table, 53% of them are marked as underutilised in specific locations, 24.2% are marked as grown/found and 22.7% are marked as major crops. Fig. 4 shows the distribution of crops that are considered underutilised worldwide.

### 1.6. Potential yield

The potential yield is defined as the best yield attainable in a given environment of cultivation [6]. The values are presented in maximum, mean and minimum.

### 1.7. Nutrition

This table comprises proximate composition, vitamins, minerals, amino acids, and fatty acids data of the crops. The value of each variable is mainly recorded as the mean value. The weight basis is also recorded to retain the integrity of the data and include as much information as possible from the source. Since the data are collected from different sources, a data flag is used to indicate if the data is calculated, roundup, estimated or taken directly from the source.

### 1.8. Metadata

Metadata records the contributor (name of the person who added the data), date (when the data is added), source details, location in the source and the accuracy flag (gives the indication of data reliability). Some of the fields were adopted from Dublin Core Metadata Initiative standard [7]. Each metadata record is connected to the data collected in the other tables through a foreign key called Metadata ID which has one unique identification number that connects both records. One metadata record may be linked to several records of other tables.

### 1.9. Supplementary data

All the data that were used to create figures in this article are included as a supplementary material to this article and in the reposityr. The data includes an .xlsx file that contains complied data from the database for the figures and a map file that contains the location and counts for underutilised crops.

## 2. Database code

The database was developed following relational format that allows data tables to be linked and complex queries to be performed on the data. The database code has undergone a few revisions to ensure its robustness in terms of data storage and retrieval. The database code is accessible at https://github.com/CFFRC-301 KST/CFFRC-Global-Knowledge-Base [8]. An implementation of the database as a data portal is available at https://cropbase.co.uk/cropbasev5/. This interface allows access to the data freely. Data can be downloaded in variety of formats including CSV and PDF. The same data portal is used to update and curate the data online.

## 3. Use-case

A use-case is provided for the database data [9] in Fig. 5 that shows the distribution of crops against major climate zones as heat maps for protein and carbohydrate content. Some of the crops did not have records for protein and carbohydrate content. Therefore, the total number of crops that are used for protein content were 210 in Zone A, 21 in Zone B, 50 in Zone C and 5 in Zone D (Fig. 5a). Similarly, 126, 9, 29 and 4 crops from respective categories were used for carbohydrate content (Fig. 5b).

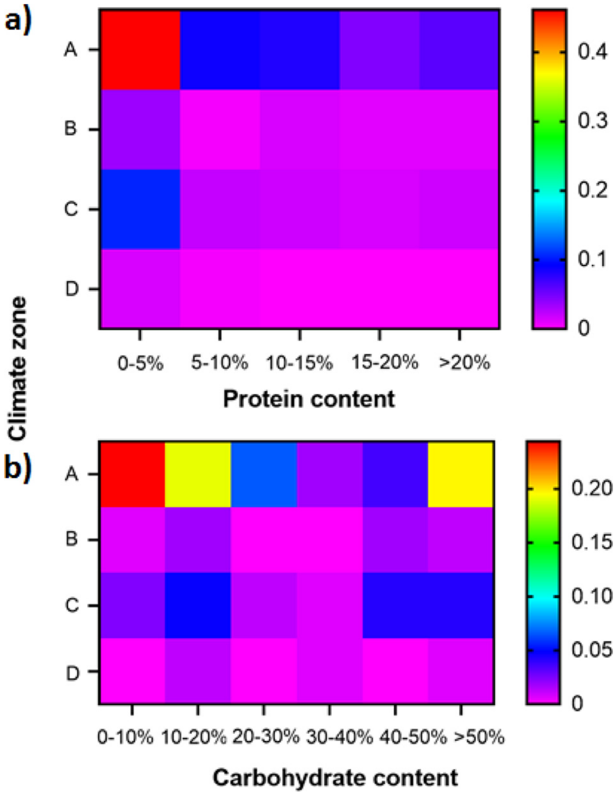

**Fig. 5.** Variation of a) protein and b) carbohydrate content among crops in different climatic zones.

## 4. Experimental Design, Materials and Methods

### *4.1. Data collection and data entry process*

A major list of crops was compiled from different sources such as research data, textbooks and databases such as ECOCROP [1]. The data collection started in early 2016 and during the time, multiple changes were made in the structure which caused some data loss but in better organization than before. The data collection was done manually with the help of 11 interns (estimated of about 10,000 h of work) who were trained in-house for data collection and entry into the database. Data entry was also mainly done manually through the graphical user interface that was created for the database. Some data, especially large datasets were automatically uploaded in bulk.

Because of the enormity of data collection task, a list of important tables was created to start the data collection. A budgeted timeline was created for the data collectors to collect data for all the crops for a set of specific parameters. Some data are naturally not available for some of the crops (particularly underutilised crops) due to limited accessible knowledge.

The general way of data collection was to search several sources i.e., research articles, website articles, books etc., which contain the relevant information for the parameters of interest for a specific crop. Traditional way of searching in different databases using scientific names, common names and keywords were used for data collection. We extracted data from the process and entered data into the database following standard operating procedures (SOPs) to ensure the data are in a standardised format (https://doi.org/10.5281/zenodo.3988378). We also ensured that the

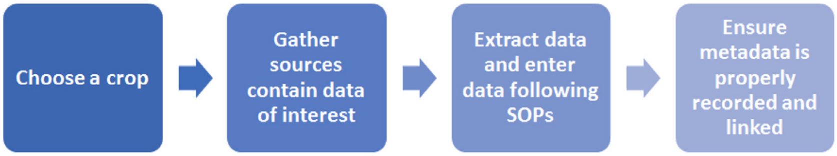
Choose a crop
Gather sources contain data of interest
Extract data and enter data following SOPs
Ensure metadata is properly recorded and linked

**Fig. 6.** Schematic diagram of data collection and data entry process.

data are properly linked to their respected metadata record. Fig. 6 shows the simplification of the data collection and data entry.

## Ethics Statement

Not applicable.

## CRediT Author Statement

**Nur Marahaini Mohd Nizar:** Methodology; Data Curation; Writing Original Draft; **Ebrahim Jahanshiri:** Conceptualization; Methodology; Writing Original Draft; Supervision; Project administration; Funding acquisition; **Siti Sarah Mohd Sinin:** Data Curation; **Eranga M. Wimalasiri:** Validation; Visualization; **Tengku Adhwa Syaherah Tengku Mohd Suhairi:** Validation; visualization; **Peter J. Gregory:** Writing Review & Editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## Acknowledgments

The authors would like to thank Crops For the Future Research Centre (CFFRC) colleagues for support and assistance in the development of this dataset: Ayman Salama, Anil Shekar Tharmandram, Sue Walker, Patrick O'Reilly, Yuveena Gopalan, Asha Karunaratne, Advina Julkifle, Tan Xin Lin, Ooi Gin Teng, Aryo Feldman, Goh Ee Von, Salvatore Viridis and Sean Mayes. The authors wish to thank the interns who contributed to data collection, and database development: Joshua Woodring, Thik Sukree, Anton Welling, Nur Irdina binti Muhammad Abir, Siti Nur Jelita Sabran, Nur Sakinah Ismail, Nurul Jannah Abdullah, Husna Zolkepli, Nurul Nadia Ya'akob, Maizatul Vanisha Binti Masril, Tharshinni Muruges and Muhd Syahiran Ahmad Shahar. This research was funded by European Union's Horizon 2020 research and innovation programme, grant agreement No. 774234.

## Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.5281/zenodo.3735694.

## References

[1] FAO, Crop ecological requirements database (ECOCROP) (2020) http://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1027491/. (accessed 04 August 2020).

[2] USDA-ARS, FoodData Central (2019) https://fdc.nal.usda.gov/. (accessed 04 August 2020).

[3] FAO, Crops statistics - concepts, definitions and classifications (2020) http://www.fao.org/economic/the-statistics-division-ess/methodology/methodology-systems/crops-statistics-concepts-definitions-and-classifications/en/. (accessed 04 August 2020).

[4] J. Grieser, R. Gommes, S. Cofield, M. Bernardi, New gridded maps of Koeppen's climate classification (2006) http://www.fao.org/nr/climpag/globgrids/KC_classification_en.asp. (accessed 04 August 2020).

[5] FAO, World programme for the census of agriculture, Classification of crops (2010) http://www.fao.org/fileadmin/templates/ess/documents/world_census_of_agriculture/appendix3_r7.pdf. (accessed 04 August 2020).

[6] L.T. Evans, R.A. Fischer, Yield potential: Its definition, measurement, and significance, Crop Sci. 39 (6) (1999) 1544–1551, doi:10.2135/cropsci1999.3961544x.

[7] Dublin Core Metadata Initiative, Using Dublin Core$^{TM}$ - The Elements (2006) https://www.dublincore.org/specifications/dublin-core/usageguide/elements/. (accessed 04 August 2020).

[8] A.S. Tharmandram, Mohd Nizar, N. M., E. Jahanshiri, A. Salama, Mohd Sinin, T.A.S. Tengku Mohd Suhairi, H. Hussin, P.J. Gregory, S.N Azam-Ali, Global knowledge base for underutilised crops, Zenodo (2020), doi:10.5281/zenodo.3735694.

[9] Mohd Nizar, N. M., E. Jahanshiri, A.S. Tharmandram, A. Salama, Mohd Sinin, S. S., N.J. Abdullah, H. Zolkepli, E.M. Wimalasiri, T.A.S.T. Mohd Suhairi, H. Hussin, P.J. Gregory, S.N Azam-Ali, Underutilised crops database for supporting agricultural diversification, Comput. Electron. Agric. 180 (2021) 105920, doi:10.1016/j.compag.2020.105920.