Bridging the Food Security Gap: an information-led approach to connect dietary nutrition, food composition and crop production

Short running title: Bridging the Food Security Gap

Razlin Azman Halimi¹, Bronwyn J. Barkla¹, Liliana Andrés-Hernández¹, Sean Mayes^{2, 3},
Graham J. King^{*1, 2, 3}

¹ Southern Cross Plant Science, Southern Cross University, Lismore, NSW2480, Australia

² School of Biosciences, University of Nottingham, UK

³ Crop Improvement and Production, Crops For the Future, Malaysia

*corresponding author graham.king@scu.edu.au

Abstract

Background:

Food security is recognised as a major global challenge, yet human food chain systems are inherently not geared towards nutrition, with decisions on crop and cultivar choice not informed by dietary composition. Currently, food compositional tables and databases (FCT/FCDB) are the primary information source for decisions relating to dietary intake.

However, these only present single mean values representing major components.

Establishment of a systematic controlled vocabulary to fill this gap requires representation of a more complex set of semantic relationships between terms used to describe nutritional composition and dietary function.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/jsfa.10157

Results:

We carried out a survey of 11 FCT/FCDB and 177 peer reviewed papers describing variation in nutritional composition and dietary function for food crops in order to identify a comprehensive set of terms to construct a controlled vocabulary. We used this information to generate a Crop Dietary Nutrition Data Framework (CDN-DF), which incorporates controlled vocabularies systematically organised into major classes representing nutritional components and dietary function. We demonstrate the value of the CDN-DF for comparison of equivalent components between crop species or cultivars, for identifying data gaps, as well as potential for formal meta-analysis. The CDN-DF also enabled us to explore relationships between nutritional components and functional attributes of food.

Conclusion:

We have generated a structured Crop Dietary Nutrition Data Framework that is generally applicable to the collation and comparison of data relevant to crop researchers, breeders and other stakeholders, and will facilitate dialogue with nutritionists. It is currently guiding establishment of a more robust formal ontology.

Keywords:

nutritional security, cultivar variation, genetic resources, databases, knowledge representation, food supply

1. Introduction

Food security is recognised as a major global challenge, underlined by the need to meet sustainable calorific and nutritional requirements of a world population projected to reach nine billion within the next 15 years¹. Whilst food security is a pre-requisite for achieving nutritional security, it is accepted that securing food supplies alone does not guarantee optimal nutritional status of a population^{2,3}. Human nutrition encompasses the energy and essential nutrients required to fulfil the dietary needs of the body⁴. It has been suggested that most human food chain systems are inherently not geared towards nutrition⁵. Within most current food systems, decisions tend to be made independently by those involved in agricultural production, plant breeding, and human nutrition, with little connectivity⁶. In practice, production decisions determining crop and cultivar choice by farmers are primarily driven by price, yield and market preference^{7,8}, and thus tend to be poorly aligned with dietary needs^{9,10}. Multiple examples exist, including a shift in wheat cultivars grown in W. Australia driven by export demand for cultivars meeting market specifications for Udon noodles¹¹, cultivar preference in Ethiopia driven by environmental stability and adaptability¹², and breeding selection for quinoa in the Peruvian Andes driven by mildew resistance, shorter maturation and other yield parameters¹³. In Central India, increased price has resulted in

cultivation of rice, which is of inferior nutritional value when compared with sorghum, maize and millets¹⁴.

Within breeding programmes, traits affecting dietary nutritional composition tend to be of less importance than those affecting yield, biotic resistance and appearance^{15, 16}. Plant breeders are required to produce varieties that meet market requirements for uniformity, production efficiency and product quality, including specifications from the food industry⁷. This is true for private and public sector breeding^{17, 18}, as well as farmer breeding co-operatives^{13, 19}. This is also reflected in predictive models developed for farmers to identify and rank valuable crop traits for cultivar choice²⁰, where value is primarily attached to production system or market preference, rather than nutritional content.

At present, decisions by nutritionists and consumers for dietary intake tend to be made at the level of individual crops or food products²¹. This has led to a recognition that the lack of cultivar-specific nutritional composition data presents a significant obstacle to wider adoption of crop varieties with improved nutritional value^{9, 22, 23}. Food compositional tables and databases (FCT/FCDB) are currently the primary sources of information for formulating guidelines for food intake, such as Recommended Daily Allowances (RDA) and Dietary Reference Intake (DRI)^{24, 25}, as well as for food labelling^{26, 27} and marketing²⁸. However, in FCT/FCDB, each nutritional component is presented as a single numerical concentration that represents a mean (not median) value. Moreover, there is no indication of variation attributable to either cultivar or growing environment. Although there has been an increased

Appreciation of functional foods with potentially positive effects on human health beyond basic nutrition^{29, 30}, such data are not widely managed within FCT/FCDB³¹.

It is interesting and perhaps ironic that in contrast to human nutrition, decision making for dietary intake of livestock feed formulation is controlled with high precision for distinct developmental stages of each animal species, including identification of crop cultivars that meet specific dietary needs^{32, 33}. In addition, numerous software applications have been developed with increasing sophistication to automate feed production, most of which incorporate relative cost and nutritional content, and more recently details of feed mix and livestock dietary outcomes³⁴.

There appear to be tangible benefits to harmonising access to crop nutritional data for different decision makers in the food supply chain, such as plant breeders and nutritionists. However, this requires a common understanding of concepts to facilitate dialogue, as well as increased interoperability of data sources. Developing such infrastructure starts with shared terminology and can be facilitated by formal systems of knowledge representation³⁵. For example, adoption of common language based on controlled vocabularies has been shown to facilitate access and use of comparative data sources in metabolomics³⁶. However, in the case of crop nutritional composition and dietary function data, this is likely to require re-evaluation of the information processing pipelines and standards used for their generation, organisation and dissemination (Figure 1).

Accepted Article

It has been recognised that a systematic and formal framework for describing and organising relevant nutritional components with controlled vocabularies would add value to crop nutrition research, by increasing the interoperability of data sources between plant breeders and nutritionists³⁷. However, no comprehensive formal vocabulary has yet emerged for nutritional composition of food crops. Although some attempts have been made to establish controlled vocabularies for nutritional components such as ‘protein’ and ‘lipid’, these are incomplete and inconsistent. For instance, neither Crop Ontology³⁸, FoodOn³⁹ or OntoFood⁴⁰ represent a comprehensive set of nutritional components or dietary functions, nor provide sufficient detail in terms of structured relationships between terms⁴¹. This limits their utility for managing and comparing data within- and between-crops. Establishment of a systematic controlled vocabulary to fill this gap requires representation of a more complex set of semantic relationships between terms used to describe nutritional composition and dietary function.

In this paper, we compare the quality of information available in FCT/FCDB with more detailed sources of data that better reflect variation due to crop type, cultivar and interaction with production environments. We propose and outline development of the Crop Dietary Nutrition Data Framework (CDN-DF), based on controlled vocabularies systematically organised into simple hierarchal branching trees. Two main classes of terms represent nutritional components and dietary function. We include a use case of grain legumes, in order to understand the practical challenges faced when comparing nutritional information from different major and minor crops.

2. Materials and methods

2.1 Sources of nutritional composition data

Food composition data values were sourced from nine national FCT/FCBD and two specialised FCBD; Food Composition Database for Biodiversity version 4.0 (BioFoodComp4.0) and the Global Food Composition Database for Phytate version 1.0 (PhyFoodComp1.0) (Table 1). The reference website for each FCT/FCDB was used to identify the criteria for comparison (Table 1). Supporting documents and files were downloaded and used to provide context on FCT/FCDB data provenance, compilation and coverage. Proximate composition data of Kabuli-type chickpea cultivars grown in the USA between 2011 and 2016 were sourced from the *US Pulse Quality Survey*^{42, 43}.

2.2 Development of Crop Dietary Nutrition Data Framework

The USDA Nutrient Database for Standard Reference (NDSR)⁴⁴ was used as the initial data source representing a well-developed national FCDB, from which 144 nutritional components were identified and allocated unique entity terms within the data framework (Figure 2). A further 401 nutritional components were identified from defined search of peer-reviewed literature (Supplementary Table 1). Where synonyms of a given entity occurred in the literature, the term allocated was that closest to being of relevance to nutritionist (e.g. the fatty acid term ‘oleic acid’ was fixed, although it may appear in different literature and in ChEBI as ‘(9Z)-Octadec-9-enoic acid’ or ‘cis- Δ^9 -octadecanoic acid 18:1 cis9’)

Accepted Article

The CDN-DF was designed and presented in Microsoft ExcelTM (2013) spreadsheets for two major classes representing nutritional components and dietary function. The vocabulary within each class consists of unique entity terms arranged within a hierarchical branching tree implemented using the 'Group' function within Excel (Supplementary Tables 2 & 3). Within the nutritional component class, entity terms were allocated to six primary categories, each representing the root node with one or more branches. The class tree was extended with three additional levels representing unique terms for progressively specific sub-categories, and a final level that was intended to correspond to the smallest bioavailable molecule (Supplementary Table 2).

The dietary function class tree was designed based on a separate search of peer reviewed literature, in which terms describing specific aspects of dietary function were identified (Figure 2). A preliminary yet non-exhaustive set of primary categories were identified: anti-nutritional factors, food toxins, phytonutrients and antioxidants. At the second level more specific functional sub-categories were allocated. At the third level sub-categories of nutritional components are indicated, based on initial interpretation of journal articles which have quantified nutritional components associated with various aspects of a dietary function (Supplementary Tables 1 & 3).

Finally, corresponding ontology terms from Chemical Entities of Biological Interest (ChEBI)⁴⁵ were assigned to entity terms within both class trees (Figure 2).

2.3 Grain legume nutritional data

Nutritional composition data were collated within the CDN-DF for a subset of five grain legume (syn. pulse) species of the *Fabaceae*: soybean (*Glycine max*), chickpea (*Cicer ariteneum*), cowpea (*Vigna unguiculata*), mungbean (*V. radiata*), and a taxonomically related underutilised species, bambara groundnut (*V. subterranea*). This included data for 196 compounds derived from 177 different sources, including FCT/FCDB, peer reviewed journal articles, industry reports and trade bodies, selected based on inclusion of i. replicated data values, reported mean and standard deviation, and ii. detailed description of analytical method. For each nutritional component, a minimum of three data sources that met these selection criteria were used. From this compiled dataset, a subset obtained from 29 peer reviewed journal articles was selected for comparison of total starch, resistant starch, starch amylose concentrations and available reports of *in vitro* and *in vivo* glycaemic index (GI) (Supplementary Table 4). Data from six sources were extracted to assess the relationship between total phenolic content (TPC) and anti-oxidative capacity as determined by diphenyl-1-picrylhydrazyl (DPPH) scavenging activity (Figure 3).

3. Results

3.1 Comparison of different sources of nutritional data

We evaluated a range of different data sources in order to identify suitable terms for a controlled vocabulary representing nutritional components. Analysis of the compiled datasets (Table 1) indicated that minimum information held in FCT/FCDB is categorised as ‘proximate’ composition⁴⁶, which quantifies the amount of carbohydrate, protein, fat, fibre

and moisture per 100g for each food item reported. The granularity of information within FCT/FCDB varies, and appears to be a function of available funding, with those from developed economies having better data coverage and depth in comparison to those from less developed economies. For example, the Composition of Foods Integrated Dataset (CoFID, UK) includes data for 186 compounds analysed from 2,898 foods, whereas the Mozambique FCT includes data on 35 compounds from 53 foods (Table 1).

Each nutritional component within a FCT/FCDB is presented as a single numerical concentration that represents a mean (not median) value. For example, we found that the beta carotene content of broccoli is reported as a single value in the NDSR, although derived from over 124 samples based on unspecified sampling strategies⁴⁴.

Data reporting nutritional components with dietary functional roles such as plant secondary metabolites are often limited, even for well-established, robust FCT/FCDB (Table 1).

However, the Canadian Nutrient File (CNF) and Australian Food Composition Database (AFCD) contain data for a narrow set of carotenoids, phytosterols and vitamins^{47, 48} but no other functional categories. PhyFoodComp1.0 is the sole example of a FCDB that has emerged following a systematic effort to collate the content of a food component with a well-established dietary function in the human diet. This was achieved by documenting the phytate content of 3,377 raw food materials, products and recipes from 648 data sources⁴⁹.

Crop cultivar-specific data that reflect variation in nutritional composition, growing season or cultivation practice are absent in all nine national FCT/FCDB (Table 1). BioFoodComp4.0 is an initiative to compile cultivar-specific nutritional data⁵⁰ from a range of international sources. It comprises a set of two-dimensional tables presented in spreadsheets that collate data for 793 crops and 7,941 cultivars derived from peer-reviewed publications and some FDCDBs. The 451 nutritional components appear to include redundancy of terms due to use of multiple reporting units, and are not organised with any hierarchical or other structure.

The *US Pulse Quality Survey* represents an example of open source cultivar-specific data generated by collaboration between academia, industry bodies and growers association. Evaluation of data extracted from the 2011⁴² and 2016⁴³ editions indicate that in the USA, two out of the seven major cultivars of Kabuli chickpea have been replaced, representing a cultivar turnaround of ~30% -typical of many arable crops. Amongst all cultivars, Bronic had a significantly higher protein content (20.7%), and CDC Frontier and Sierra had starch concentrations over 42%. Further information of value to the food processing industry within the *Survey* included physical parameters such as 1,000 seed weight, hydration and swelling capacity, and cooked firmness, along with physiochemical properties such as starch characteristics (peak viscosity, peak time, pasting temperature)⁴³.

Another issue that hinders the direct comparison, inter-operability and sharing of crop nutrition data is the considerable variation in units used to report nutritional composition data within FCT/FCDB and research literature. In some cases concentrations of compounds are

presented in up to four different units. For example, amino acid content was variously reported as g/16gN, % protein, % dry matter and g/kg protein; monosaccharide content as either mg/g dry matter basis, g/100g sample or % of sugar; and fatty acid content as % of total fatty acid, % in oil and mg/100g of sample. The use of different units appears primarily dependent on the analytical approach taken, but may also reflect available equipment, historical adoption of specific methods, or development of in-house methodologies.

3.2 Crop Dietary Nutrition Data Framework

Here we propose the establishment of the CDN-DF as a structured controlled vocabulary, organised within two major classes representing nutritional components (Supplementary Table 2) and dietary function (Supplementary Table 3). The vocabulary within each class consists of unique entity terms that are arranged within a hierarchical branching tree. (Figure 2). A maximum of five levels were defined for the nutritional component class and three levels for the dietary function class, corresponding to a progressively granular representation. A key property of this organisation is that the sum of component values corresponding to entity terms described in lower branches/levels may be used as proxies for the level above.

Within the nutritional component class, 545 entity terms were allocated to six primary categories, each representing the root node with one or more branches. These closely correspond to the major proximate components⁴⁶: carbohydrate, protein and lipid as well as mineral, vitamin and secondary metabolite. The class tree was then extended with three additional levels representing unique terms for progressively specific sub-categories, and a

Accepted Article

final level that was intended to correspond to the smallest bioavailable molecule. (Figure 2 & Supplementary Table 2). As an example, from the primary node of <1:carbohydrate> the secondary branches include the terms <2:digestible_carbohydrate> and <2:non-digestible_carbohydrate>. The second-level term <2:digestible_carbohydrate> would then have a third-level sub-branch of <3:digestible starch> and subsequent sub-branches of <4:amylose> and <5:glucose> (Supplementary Table 2). For the dietary function class a similar approach was taken, with a set of preliminary yet non-exhaustive list of primary categories identified: anti-nutritional factors, food toxins, phytonutrients and antioxidants. At the second level more specific functional sub-categories were allocated.

In practice, the CDN-DF is available to be used for facilitating literature and database searches followed by the recovery, collation and curation of data from multiple sources. For any particular dataset, individual data records and values should be allocated to a unique term within the CDN-DF, dependent upon the appropriate level at which it has been described. The use of the controlled vocabulary and associated term may then be incorporated within a curation database underlying comparative or meta-analyses. It is important to recognise that a key step in the collation of data from multiple sources involves ensuring that the reporting units associated with each entity term are consistent or undergo appropriate conversion.

3.3 Use case: grain legumes

Concentrations of total starch, resistant starch, starch amylose and reports of *in vitro* and *in vivo* glycaemic index (GI) showed intra and inter species ranges (Supplementary Table 4).

We also assessed the relationship between the nutritional component TPC and the functional attribute anti-oxidative capacity, as quantified by DPPH scavenging activity (Figure 3). The results indicated that anti-oxidative capacity is effected by both crop and cultivar selection, with mungbean cultivars showing the greatest DPPH scavenging capacity, followed by cowpea, soybean and chickpea (Figure 3). Our analysis also showed that mungbean and cowpea cultivars have a greater range of TPC concentration and DPPH scavenging activity in comparison to soybean and chickpea cultivars. No adequate data sources were identified for bambara groundnut.

4. Discussion

We carried out a survey of 11 FCT/FCDB and 177 peer reviewed papers describing variation in nutritional composition and dietary function for food crops in order to identify a comprehensive set of terms that could be used to construct a controlled vocabulary. We used this information to generate the CDN-DF, comprised of a systematic allocation of entity terms to two major classes organised as simple branching hierarchical trees.

From our evaluation of different nutritional data sources, we conclude that FCT/FCDB are limited in their capacity to inform decisions on cultivation and consumption of specific crops or cultivars that would have positive outcomes affecting dietary nutrition. This was clear from the presentation of compositional ‘crop mean values’ (Table 1), which may mask variance due to different sampling and analysis protocols⁴⁶ as well as the actual range, variance and skewness in nutritional components due to cultivar⁵¹ or growing environment⁵².

The reporting of a single crop compositional value may also fail to reflect variation associated with regional crop production and markets, which has been recognised as limiting the impact of crop biodiversity on food systems and nutrition⁵³. Increasing the availability of cultivar-specific nutritional data is particularly relevant for regional decision-making by farmers or processors for production and sale into specific markets^{8, 9, 54}, and indeed may help in the development of markets sensitive to nutritional composition.

At the consumption end of the food system, poor awareness of crop compositional variation may distort estimates of nutrient intake, particularly in distinguishing between micronutrient deficiency and adequacy. For example, grain protein concentration of rice cultivars has been reported to range 2.8 fold (5g - 14g/100g)⁵⁵, whilst banana beta-carotene content can vary dramatically from 1 µg to 8,500 µg/100 g fresh weight between varieties⁵⁶. The limited availability of comparative dietary function data within FCT/FCDB also limits the ability to manage human diet at the level of crop or cultivar^{57, 58}. However, there are ongoing efforts led by the team responsible for the NDSR to widen the number of secondary metabolite compounds included in FCT/FCDB^{25, 44, 59}. In addition, reports from crop-specific trade bodies such as the *US Pulse Quality Survey*^{42, 43} may provide nutritional composition data at cultivar level that affects market price and farmer cultivation decisions^{8, 14}.

There are also notable exceptions in the use of crop and cultivar-specific nutritional data in the nutraceutical and functional food sector where cultivar development has been vertically

integrated with marketing³⁰, including instances of biofortification⁶⁰, and recent increased consumer interest in 'nutrient dense' foods⁶¹. In addition, there are other examples of change in crop end-use market to meet global demand, such as the concerted effort over the past sixty years to modify the nutritional attributes of soybean as a protein source. This has led to the large scale cultivation and breeding of high protein varieties^{62, 63}.

A systematic and formal framework for describing or organising crop nutritional components has not previously emerged, although establishment of BioFoodComp4.0 provided a first step towards documenting nutritional composition in the context of cultivar and environment⁵⁰. This unstructured and non-hierarchical compilation was established to promote food biodiversity within nutrition projects and programmes⁶⁴, and was nominally designed for use by other database compilers to enable easy access and incorporation of the available data into national FCT/FCDB⁵⁶. However, this is limited by the lack of systematic organisation or capacity to navigate a hierarchal framework of nutritional composition or dietary function.

In order to address these gaps we have proposed the CDN-DF which incorporates controlled vocabularies systematically organised into major classes representing nutritional components and dietary function (Figure 2, Supplementary Tables 2 & 3). The development of the CDN-DF identified 401 nutritional chemical components not reported in national FCT/FCDB. When implemented in spreadsheets, we have found that this structured vocabulary facilitates navigation and exploration of nutritional terms, as the hierarchical tree branches may be collapsed or expanded (Supplementary Tables 2 & 3). Once populated with data curated from

multiple referenced sources, this schema allows rapid comparison of equivalent components between crop or cultivar, as well as identifying data gaps (Supplementary Table 4). This also provides a potentially valuable tool for formal meta-analysis, which relies on referenced data collation and management⁶⁵.

The organisation of entity terms in the nutritional component class is based on systematic definition of the hierarchical relationships. These reflect well-established groupings of molecules as described in the crop nutrition literature (Supplementary Tables 1 & 2). In most cases there are analytical methodologies available that are able to quantify the entities described at the different levels within the tree. This class tree greatly extends the number of terms within a controlled vocabulary available to crop scientists or dieticians, compared with those described in ontologies such as Crop Ontology³⁸ and OntoFood⁴⁰. In addition, the CDN-DF includes a provisional dietary function class (Supplementary Table 3). Although preliminary, we recognise that there is considerable scope to extend and refine this vocabulary, and represent the various complex relationships that exist between dietary function and relevant nutritional components⁶⁶. Ideally this effort should involve relevant expertise and be informed by systematic evaluation of the literature, including canonical sources that represent the evidence base in relation to human physiology. In the future, there may be value in defining additional class trees within the CDN-DF, to encompass physiochemical and processing properties of crop products that relate to dietary nutrition and for example, may vary in terms of bioavailability and shelf-life^{67, 68}.

Accepted Article

As a use case we analysed grain legume datasets organised using the controlled vocabulary and hierarchical relationships within the CDN-DF. Although this indicated that bambara groundnut was under-represented compared with the other four crops, we were able to establish that reported variation in the concentrations of protein, fatty acid and minerals covered a similar range as for major crops such as chickpea, cowpea and mungbean⁶⁹. However, gaps in data available for starch digestibility, vitamins, and the majority of phytochemicals and anti-nutritional factors highlight where additional datasets could be generated. Notwithstanding these gaps, the data suggest that there is sufficient variation within the global bambara groundnut genepool to develop high protein cultivars, and improve concentrations of unsaturated fatty acids.

The use case also demonstrated the value of combining compositional and functional data within the same structured framework. For example, we presented a positive correlation between TPC and anti-oxidative capacity (Figure 3), in agreement with well-established findings that phenolic compounds in grain legumes contribute to their antioxidant capacity^{70, 71}. The variation observed also suggests that TPC is a valid target for selection within breeding programs. In contrast, the relationship between functional attribute GI and food composition is more complex, and has been associated with resistant starch, typically attributed to higher concentrations of starch amylose^{72, 73}. For the five grain legumes, we were able to mine the available data to illustrate the considerable intra-species variation reported for starch amylose and resistant starch concentration, as well as for *in vivo* and *in vitro* GI values (Supplementary Table 4).

This comparison also highlighted the lack of cohesiveness at the crop cultivar level that otherwise would permit inference of any valid conclusions with respect to the interaction between these parameters. Given the growing need to manage diabetes in global populations⁷⁴, and reports of grain legumes being ‘low GI’ foods^{72, 75, 76}, more comprehensive surveys of crops and cultivars are required to establish functional interactions between specific food components and GI, with the latter determined using standardised *in vivo* methods.

The CDN-DF represents a first step in facilitating the harmonisation of data sources and navigation of datasets for comparative analysis both within and between crops. To extend this further and increase access, sharing and re-use of datasets requires development of a formal ontology, able to be machine and human readable. Such features are notably lacking from FCT/FCDB. The structured vocabulary we have defined here is under-pinning the establishment of a Crop Dietary Nutrition Ontology (CDNO)⁴¹, which is expected to increase interoperability of data sources between breeders and nutritionists.

It is timely to develop standardised frameworks for knowledge representation relating to crop nutrition that adhere to the principles of F.A.I.R. (Findable, Accessible, Interoperable and Re-usable) data management⁷⁷. Initiative such as the Breeding API (BrAPI)⁷⁸ and MIAPPE⁷⁹ are enhancing the ability of pre-breeding scientists and plant breeders to compare and make use of data from diverse sources. Likewise, development of formal systems of knowledge representation including ontologies have contributed to progress in the sophistication of

Accepted Article

nutritional epidemiology research, leading to the recent development of the Ontology for Nutritional Epidemiology (ONE)⁸⁰.

The particular value of the CDN-DF lies in its ready implementation and immediate availability to assist in collation of diverse datasets. The framework includes a hierarchical structure with controlled vocabularies both for nutritional composition data and for dietary function. We have demonstrated its value by compiling data for grain legumes, and deriving valuable information relating composition and functional nutrition. We anticipate the CDN-DF will play a role in wider endeavours to add value from F.A.I.R data exchange such as the Divseek International Network^{81, 82}, and increase the ability of researchers, breeders and other stakeholders to compare data. This may include supplementing current FCT/FCDB with reciprocal data links and should allow for a more robust understanding of how crop type and cultivar contribute to dietary nutrition.

Acknowledgements

The authors wish to thank Dr Ben Liu of Southern Cross Plant Science, Southern Cross University for his advice and assistance in allocating entity terms within the nutritional components and functional dietary class trees.

Conflict of interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

This work was supported by Southern Cross University (SCU) and Crops For the Future (CFF).

References

1. FAO. The future of food and agriculture – Trends and challenges. 2017.
2. Ghattas H. Food security and nutrition in the context of the nutrition transition. 2014.
3. Hwalla N, El Labban S, Bahn RA. Nutrition security is an integral component of food security. *Frontiers in Life Science*. 2016;9(3):167-72.
4. WHO. Health topics. Nutrition 2018 [Available from: <http://www.who.int/topics/nutrition/en/>].
5. GLOPAN. Food Systems and Diets: Facing the Challenges of the 21st Century. 2016.
6. Béné C, Oosterveer P, Lamotte L, Brouwer ID, de Haan S, Prager SD, et al. When food systems meet sustainability – Current narratives and implications for actions. *World Development*. 2019;113:116-30.
7. Lammerts van Bueren ET, Struik PC, van Eekeren N, Nuijten E. Towards resilience through systems-based plant breeding. A review. *Agron Sustainable Dev*. 2018;38(5):42.
8. Lee H, Bogner C, Lee S, Koellner T. Crop selection under price and yield fluctuation: Analysis of agro-economic time series from South Korea. *Agricultural Systems*. 2016;148:1-11.
9. Dwivedi SL, Lammerts van Bueren ET, Ceccarelli S, Grando S, Upadhyaya HD, Ortiz R. Diversifying Food Systems in the Pursuit of Sustainable Food Production and Healthy Diets. *Trends Plant Sci*. 2017;22(10):842-56.
10. Khoury CK, Bjorkman AD, Dempewolf H, Ramirez-Villegas J, Guarino L, Jarvis A, et al. Increasing homogeneity in global food supplies and the implications for food security. *Proceedings of the National Academy of Sciences*. 2014;111(11):4001.
11. Elliot P, Kingwell R, Carter C, Yamamoto M, White P. Western Australia's noodle wheat industry. Current status and future challenges. Australia: Australian Export Grains Innovation Centre; 2015.
12. Asrat S, Yesuf M, Carlsson F, Wale E. Farmers' preferences for crop variety traits: Lessons for on-farm conservation and technology adoption. *Ecological Economics*. 2010;69(12):2394-401.
13. Gamboa C, Van den Broeck G, Maertens M. Smallholders' Preferences for Improved Quinoa Varieties in the Peruvian Andes. *Sustainability*. 2018;10(10):3735.
14. DeFries R, Mondal P, Singh D, Agrawal I, Fanzo J, Remans R, et al. Synergies and trade-offs for sustainable agriculture: Nutritional yields and climate-resilience for cereal crops in Central India. *Global Food Security*. 2016;11:44-53.

- Accepted Article
15. Burchi F, Fanzo J, Frison E. The Role of Food and Nutrition System Approaches in Tackling Hidden Hunger. *Int J Environ Res Public Health*. 2011;8(2):358-73.
 16. Sands DC, Morris CE, Dratz EA, Pilgeram A. Elevating optimal human nutrition to a central goal of plant breeding and production of plant-based foods. *Plant Sci*. 2009;177(5):377-89.
 17. Acheampong Patricia P, Owusu V, Nurah G. How does Farmer Preference matter in Crop variety Adoption? The case of Improved Cassava varieties' Adoption in Ghana. *Open Agriculture* 2018. p. 466.
 18. Laborte AG, Paguirigan NC, Moya PF, Nelson A, Sparks AH, Gregorio GB. Farmers' Preference for Rice Traits: Insights from Farm Surveys in Central Luzon, Philippines, 1966-2012. *PLoS ONE*. 2015;10(8):e0136562.
 19. Christinck A, Weltzien E. Plant breeding for nutrition-sensitive agriculture: an appraisal of developments in plant breeding. *Food Security*. 2013;5(5):693-707.
 20. Kaine G, Lengley S, Seymour E. A model of farmers' identification of valuable traits and trait selection decisions. Australian Government Department of Primary Industries; 2011.
 21. Mozaffarian D, Angell SY, Lang T, Rivera JA. Role of government policy in nutrition—barriers to and opportunities for healthier eating. *BMJ*. 2018;361:k2426.
 22. FAO. Report on the Commissions on Genetic Resources for Food and Agriculture. Rome, Italy; 2013. Contract No.: CGRFA-14/13/Report.
 23. Smith KW, Bronselaer A, De Baets B., De Tre G, Van Camp J, Van Damme P, et al. Scoping study to determine the data sources on biodiversity in diet and food intake. Final report. Belgium; 2014.
 24. Penington JAT, Stumbo PJ, Murphy SP, McNutt SW, Eldridge AL, McCabe-Sellers BJ, et al. Food Composition Data: The Foundation of Dietetic Practice and Research. *J Am Diet Assoc*. 2007;107:2105-13.
 25. Haytowitz DB, Pehrsson PR. USDA's National Food and Nutrient Analysis Program (NFNAP) produces high-quality data for USDA food composition databases: Two decades of collaboration. *Food Chem*. 2018;238:134-8.
 26. Clancy AK, Woods K, McMahon A, Probst Y. Food Composition Database Format and Structure: A User Focused Approach. *PLoS ONE*. 2015;10(11).
 27. Finglas PM, Berry R, Astley S. Assessing and Improving the Quality of Food Composition Databases for Nutrition and Health Applications in Europe: The Contribution of EuroFIR. *Adv Nutr*. 2014;5(5).
 28. Marconi S, Durazzo A, Camilli E, Lisciani S, Gabrielli P, Aguzzi A, et al. Food Composition Databases: Considerations about Complex Food Matrices. *Foods*. 2018;7(1):2.
 29. Roberfroid MB. Global view on functional foods: European perspectives. *Br J Nutr*. 2002;88(S2):S133-S8.
 30. Siró I, Kápolna E, Kápolna B, Lugasi A. Functional food. Product development, marketing and consumer acceptance—A review. *Appetite*. 2008;51(3):456-67.
 31. Spence JT. Challenges related to the composition of functional foods. *J Food Compos Anal*. 2006;19:S4-S6.
 32. Miller EL. Protein nutrition requirements of farmed livestock and dietary supply Rome; 2004.

- Accepted Article
33. Nutritional Requirements of Pigs [Internet]. MSD. 2019.
 34. Saxena P. Feed formulation software: A comparative study. 2010;1:20-1.
 35. Lange MC, Lemay DG, German JB. A multi-ontology framework to guide agriculture and food towards diet and health. *J Sci Food Agric*. 2007;87(8):1427-34.
 36. Spasić I, Schober D, Sansone S-A, Rebholz-Schuhmann D, Kell DB, Paton NW. Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC bioinformatics*. 2008;9 Suppl 5(Suppl 5):S5-S.
 37. Lemay DG, Zivkovic AM, German JB. Building the bridges to bioinformatics in nutrition research. *Am J Clin Nutr*. 2007;86.
 38. Matteis L, Chibon P-Y, Espinosa H, Skofic M, Finkers R, Bruskiewich R, et al. *Crop Ontology: Vocabulary For Crop-related Concepts* 2013.
 39. Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*. 2018;2(1):23.
 40. OntoFood [Internet]. NCBO. 2015 [cited 26 October, 2018]. Available from: <https://bioportal.bioontology.org/ontologies/OF>.
 41. Andrés-Hernández L, Baten A, Azman Halimi R, Walls R, King GJ. Knowledge Representation and Data Sharing to Unlock Crop Variation for Nutritional Food Security. *Crop Sci*. 2019.
 42. Thavarajah D, Thavarajah P. 2011 U.S Pulse Quality Report. 2011.
 43. Hall C. 2016 U.S Pulse Quality Survey. 2016.
 44. USDA Food Composition Database [Internet]. USDA. 2019 [cited 26 May, 2017]. Available from: <https://ndb.nal.usda.gov/ndb/>
<http://fdc.nal.usda.gov>.
 45. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*. 2016;44(D1):D1214-9.
 46. Greenfield H, Southgate DAT. *Food Composition Data. Production, Mangement and Use*. 2 ed: FAO; 2003.
 47. Canadian Nutrient File [Internet]. Health Canada. 2015 [cited 1st February 2018]. Available from: <https://food-nutrition.canada.ca/cnf-fce/index-eng.jsp>.
 48. National Health and Medical Research Council AGDoHaA, Health. NZMo. 2017 Nutrient reference Values for Australia and New Zealand. Australia: Australian Government; 2017.
 49. FAO/INFOODS/IZiNCG Global Food Composition Database for Phytate Version 1.0 - PhyFoodComp 1.0. [Internet]. FAO. 2018.
 50. FAO. BioFoodComp Food Composition Database for Biodiversity. In: FAO/INFOODS, editor. 4.0 ed 2017.
 51. Burlingame B, Charrondiere R, Mouille B. Food composition is fundamental to the cross-cutting initiative on biodiversity for food and nutrition. *J Food Compos Anal*. 2009;22:361 - 5.
 52. Litaladio N, Burlingame B, Crews J. Horticulture, biodiversity and nutrition. *J Food Compos Anal*. 2010;23.

53. Lachat C, Raneri JE, Smith KW, Kolsteren P, Van Damme P, Verzelen K, et al. Dietary species richness as a measure of food biodiversity and nutritional quality of diets. *Proceedings of the National Academy of Sciences*. 2018;115(1):127-32.
54. Allen S, de Brauw A. Nutrition sensitive value chains: Theory, progress, and open questions. *Global Food Security*. 2018;16:22-8.
55. Kennedy G, Burlingame B. Analysis of food composition data on rice from a plant genetic resources perspective. *Food Chem*. 2003;80.
56. Charrondière UR, Stadlmayr B, Rittenschober D, Mouille B, Nilsson E, Medhammar E, et al. FAO/INFOODS food composition database for biodiversity. *Food Chem*. 2013;140.
57. Martin CL, Murphy SP, Au DLM. Compiling glycemic index and glycemic load values for addition to a food composition database. *J Food Compos Anal*. 2008;21(6):469-73.
58. Dahdouh S, Grande F, Espinosa SN, Vincent A, Gibson R, Bailey K, et al. Development of the FAO/INFOODS/IZINCG Global Food Composition Database for Phytate. *J Food Compos Anal*. 2019;78:42-8.
59. Sebastian RS, Enns CW, Goldman JD, Martin CL, Steinfeldt LC, Murayi T, et al. A New database facilitates characterization of flavonoid intake, sources, and positive associations with diet quality among us adults. *The Journal of Nutrition*. 2015;145(6):1239.
60. Bouis HE, Saltzman A. Improving nutrition through biofortification: A review of evidence from HarvestPlus, 2003 through 2016. *Global Food Security*. 2017;12:49-58.
61. Drewnowski A, Dwyer J, King JC, Weaver CM. A proposed nutrient density score that includes food groups and nutrients to better align with dietary guidance. *Nutr Rev*. 2019;77(6):404-16.
62. Nafziger ED. Soybean Agronomy. Reference Module in Food Science: Elsevier; 2016.
63. Carter J, T. E. , Rzewnicki PE, Burton JW, Villagarcia MR, Bowman DT, Taliencio E, et al. Registration of N6202 Soybean Germplasm with High Protein, Favorable Yield Potential, Large Seed, and Diverse Pedigree. *Journal of Plant registrations*. 2010;4(1).
64. FAO. International Food composition tables/ database directory 2017 [Available from: <http://www.fao.org/infoods/infoods/tables-and-databases/en/>]
65. Rotundo JL, Westgate ME. Meta-analysis of environmental effects on soybean seed composition. *Field Crops Res*. 2009;110(2):147-56.
66. Tapsell LC, Neale EP, Satija A, Hu FB. Foods, Nutrients, and Dietary Patterns: Interconnections and Implications for Dietary Guidelines. *Adv Nutr*. 2016;7(3):445-54.
67. Mahajan PV, Caleb OJ, Singh Z, Watkins CB, Geyer M. Postharvest treatments of fresh produce. *Philos Trans A Math Phys Eng Sci*. 2014;372(2017):20130309-.
68. Platel K, Srinivasan K. Bioavailability of Micronutrients from Plant Foods: An Update. *Crit Rev Food Sci Nutr*. 2016;56(10):1608-19.
69. Azman Halimi R, Barkla BJ, Mayes S, King GJ. The potential of the underutilized pulse bambara groundnut (*Vigna subterranea* (L.) Verdc.) for nutritional food security. *J Food Compos Anal*. 2019;77:47-59.
70. Xue Z, Wang C, Zhai L, Yu W, Chang H, Kou X, et al. Bioactive compounds and antioxidant activity of mung bean (*Vigna radiata* L.), soybean (*Glycine max* L.) and black bean (*Phaseolus vulgaris* L.) during the germination process. *Czech J Food Sci*. 2016;34(1):68-78.

- Accepted Article
71. Yao Y, Cheng X, Wang L, Wang S, Ren G. Biological potential of sixteen legumes in China. *Int J Mol Sci*. 2011;12(10):7048-58.
 72. Sandhu KS, Lim S-T. Digestibility of legume starches as influenced by their physical and structural properties. *Carbohydr Polym*. 2008;71(2):245-52.
 73. Birt DF, Boylston T, Hendrich S, Jane J-L, Hollis J, Li L, et al. Resistant starch: promise for improving human health. *Advances in nutrition (Bethesda, Md)*. 2013;4(6):587.
 74. WHO. Global Report on Diabetes. 2016.
 75. Jenkins DJA, Wolever TMS, Jenkins AL, J Thorne M, Lee R, Kalmusky J, et al. The glycaemic index of foods tested in diabetic patients: A new basis for carbohydrate exchange favouring the use of legumes 1983. 257-64 p.
 76. Ratnaningsih N, Suparmo EH, Marsono Y. In vitro Starch Digestibility and Estimated Glycemic Index of Indonesian Cowpea Starch (*Vigna unguiculata*). *Pak J Nutr*. 2017;16(1):1-8.
 77. Wilkinson MD, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3.
 78. Selby P, consortium oboTB, Abbeloos R, consortium oboTB, Backlund JE, consortium oboTB, et al. BrAPI—an application programming interface for plant breeding applications. *Bioinformatics*. 2019.
 79. Ćwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, Chen D, Cornut G, et al. Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods*. 2016;12:44-.
 80. Yang C, Ambayo H, Baets BD, Kolsteren P, Thanintorn N, Hawwash D, et al. An Ontology to Standardize Research Output of Nutritional Epidemiology: From Paper-Based Standards to Linked Content. *Nutrients*. 2019;11(6):1300.
 81. DivSeek. Divseek International Network: Global Institute for Food Security , University of Saskatchewan Canada; [Available from: <http://www.divseek.org/>].
 82. Meyer RS. Encouraging metadata curation in the Diversity Seek initiative. *Nat Plants (London, U K)*. 2015;1(7):15099.
 83. Xu BJ, Yuan SH, Chang SKC. Comparative analyses of phenolic composition, antioxidant capacity, and color of cool season legumes and other selected food legumes. *J Food Sci*. 2007;72(2):S167-S77.
 84. Xu B, Chang SK. Comparative study on antiproliferation properties and cellular antioxidant activities of commonly consumed food legumes against nine human cancer cell lines. *Food Chem*. 2012;134(3):1287-96.
 85. Lee JH, Jeon JK, Kim SG, Kim SH, Chun T, Imm JY. Comparative analyses of total phenols, flavonoids, saponins and antioxidant activity in yellow soy beans and mung beans. *Int J Food Sci Technol*. 2011;46(12):2513-9.
 86. Zia-Ul-Haq M, Iqbal S, Ahmad S, Bhangar MI, Wiczowski W, Amarowicz R. Antioxidant potential of Desi chickpea varieties commonly consumed in Pakistan. *J Food Lipids*. 2008;15(3):326-42.
 87. Zia-Ul-Haq M, Ahmad S, Amarowicz R, De Feo V. Antioxidant activity of the extracts of some cowpea (*Vigna unguiculata* (L) Walp.) cultivars commonly consumed in Pakistan. *Molecules*. 2013;18(2):2005-17.

Figure Legends

Figure 1: Schematic representation of the generation, organisation and dissemination of nutritional data relating to crop production and human diet. The Crop Dietary Nutrition Data Framework (CDN-DF) and the Crop Dietary Nutrition Ontology (CDNO) will serve as the basis for a future Crop Comparative Database, which can be used to supplement current FCT/FCDB. This will allow for a more robust understanding of the contributions of crop type and cultivar to nutritional composition of food crops.

Figure 2: Schematic workflow process used for development of the Crop Dietary Nutrition Data Framework (CDN-DF). The CDN-DF consists of two hierarchal class trees: nutritional components and dietary function. Workflow for development of the nutritional component class tree is indicated within orange box, workflow process for construction of dietary function class tree is indicated within green box. List of keywords used for the literature searches and data sources used to define entity terms are listed in Supplementary Table 1.

Figure 3: Scatter plot of chemical composition (total phenolic content, TPC) and functional property (antioxidant activity measured using diphenyl-1-picrylhydrazyl (DPPH) scavenging activity) for four grain legume species, soybean (*Glycine max*), chickpea (*Cicer ariteneum*), cowpea (*Vigna unguiculata*) and mungbean (*V. radiata*). TPC expressed as mg of gallic acid equivalents (GAE) per gram of samples, DPPH scavenging activity expressed as μmol Trolox

equivalent. Data were collated using the CDN-DF from a range of independent sources: soybean^{71, 83-85}, chickpea^{71, 83, 84, 86}, cowpea^{71, 84, 87}, and mungbean^{71, 84, 85}. Linear regression analysis across the complete dataset resulted in a R^2 of 0.51, and a dotted regression line plotted.

Table 1: Summary of food compositional tables and databases (FCT/FCDB) from the USA, UK, Australia, Canada, Denmark, Brazil, Mozambique, Thailand and Bahrain. Table provides comparison of nutritional data available, coverage, method of compilation and granularity. NDSR= USDA National Nutrient Database for Standard Reference; CoFID= McCance and Widdowson's Composition of Foods Integrated Dataset; AFCD = Australian Food Composition Database; CNF= Canadian Nutrient File; FRIDA= Frida Food Data Denmark; TBCA-USP= Brazilian Food Composition Tables; FCT Mozambique= Food Composition Tables for Mozambique; FCT Thailand= Thai Food Composition Database; FCT Bahrain= Food Composition Tables for Kingdom of Bahrain.

Name of database	Current version	Country of origin	Publisher	Number of foods analysed	Number of components analysed	Purpose	Data collection method	Online search option	Downloadable information	Cultivar specific information
NDSR (1)	2018	USA	United States Department of Agriculture (USDA)	7,793	150	Major source of food composition data for USA	Compilation [†]	Yes	Yes	No
CoFID (2)	2019	UK	Public Health England (PHE)	2,898	186	Major source of food composition data for UK	Compilation [†]	No	Yes	No
AFCD [¶] (3)	2019	Australia	Food Standards Australia New Zealand (FSANZ)	1534 (First release)	256	Reference nutrient database for	Primarily analysis	Yes	Yes	No

CNF (4)	2015	Canada	Health Canada	5807	150	Australia. Major source of food compositio n data for Canadians	Compilation [†] ‡	No	Yes	No
FRIDA (5)	2019	Denmark	National Food Institute, Technical University of Denmark (DTU)	2299	46	Major source of food compositio n data for Denmark	Analysis & compilation [†] , ‡	Yes	Yes	No
TBCA-USP (6)	2015	Brazil	University of Sao Paulo & BRASILFOOD [§]	1720	Not stated	Major source of food compositio n data for Brazil	Analysis & compilation [†] , ‡	Yes	No	No
FCT Mozambique (7)	2011	Mozambiqu e & Finland	University of Helsinki, Finland	53	35	First compilation of nutrient content of food in Mozambiqu e	Analysis & Compilation [†]	No	Yes	No
FCT Thailand (8)	2016	Thailand	Institute of Nutrition, Mahidol University [§]	1700	27	Major source of food compositio n data for Thailand	Analysis & Compilation [†]	Yes	Yes	No
FCT Bahrain (9)	2011	Bahrain	Arab Centre for	187	61	First	Analysis &	No	Yes	No

			Nutrition [§]			attempt to generate nutrient data for foods and products consumed in Bahrain	compilation [†]			
BioFoodComp4. 0 (10)	2017	Global	FAO/INFOODS	793	451	First attempt to compile crop cultivar nutritional data.	Compilation †	No	Yes	Yes
Global Food Composition Database for Phytate 1.0 (11)	2018	Global	FAO/INFOODS/IZiNC G	3,377	35	First global repository of analytical data on phytate, data on moisture, phytate, zinc, iron and calcium	Compilation †	No	Yes	No

† Indicates that the data presented is compiled from numerous sources including laboratory analysis, peer reviewed journals and governmental laboratories

‡ Contains data from other food compositional database

§ Publishers are also Regional Food Data Centre of INFOODS (FAO).

¶ Previously called NUTTAB (Nutrient Tables)

Data sources:

1. USDA Food composition Database [Internet]. USDA. 2018 [cited 26 May, 2017]. Available from: <https://ndb.nal.usda.gov/ndb/>.
2. Pinchen H, Powell N, Weiner D, , Finglas P. McCance and Widdowson's Composition of Foods Integrated Dataset. In: England PH, editor. 7 ed. UK2019.
3. Australian Food Composition Database (formerly NUTTAB) [Internet]. Commonwealth of Australia and Food Standards Australia New Zealand (FSANZ). 2019 [cited 8 January 2019]. Available from: <http://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/default.aspx>.
4. Canadian Nutrient File [Internet]. Health Canada. 2015 [cited 1st February 2018]. Available from: <https://food-nutrition.canada.ca/cnf-fce/index-eng.jsp>.
5. Frida Public Foods Database [Internet]. 2019 [cited 12 January, 2019]. Available from: <http://frida.fooddata.dk/index.php?lang=en>.
6. Brazilian Food Composition Tables [Internet]. USP. 2017 [cited 30th November, 2017]. Available from: <http://www.fcf.usp.br/tbca/>
7. Lisa L, Hauta-alus H, Mutanen M. Food Composition Tables for Mozambique. 2011 October 2011.
8. Thai Food Composition Database 2015 [Internet]. Mahidol University 2018 [cited 22nd June, 2018]. Available from: <http://www.inmu.mahidol.ac.th/thaifcd/home.php>.
9. Musaiger AO. Food Composition Tables for Kingdom of Bahrain. Bahrain; 2011.
10. FAO. BioFoodComp Food Composition Database for Biodiversity. In: FAO/INFOODS, editor. 4.0 ed2017.
11. FAO/INFOODS/IZiNCG Global Food Composition Database for Phytate Version 1.0 - PhyFoodComp 1.0. [Internet]. FAO. 2018.