A Proposal for a European Large Knowledge Repository in Advanced Food Composition Tables for Assessing Dietary Intake

Oscar Coltell[1,2], Francisco Madueño[1], Zoe Falomir[1], and Dolores Corella[2,3]

[1] Department of Computing Languages and Systems, Universitat Jaume I, Castellón, Spain
{oscar.coltell, francisco.madueno, zfalomir}@uji.es
[2] CIBER Physiopathology of Obesity and Nutrition (CIBEROBN),
Institute of Health Carlos III, Madrid, Spain
[3] Department of Preventive Medicine and Public Health, University of Valencia,
Valencia, Spain
dolores.corella@uv.es

Abstract. A proposal for designing and developing a European Repository of Knowledge on Advanced Food Composition Tables (FCTs), based on the existing national FCTs, is proposed in this paper. The requirements of the system, the interoperability strategies, and the cooperation of each national FCT for maintaining and updating the repository are discussed.

Keywords: Knowledge repositories, Food Composition Tables (FCTs), Joint Programming Initiative in A Healthy Diet.

1 Introduction

The study of the interaction between diet and the genome is crucial to prevent and treat cardiovascular diseases, some cancers, type 2 diabetes, etc. The assessment of a person's diet is a tedious task, and in practice, a portion of the intake information is evaluated and then the habitual participants' intake is extrapolated. In order to obtain enough statistical power to avoid measurement errors and changes in diet, it is necessary to obtain repeated measures of dietary information from a large number of participants over time. For extracting information regarding participants' diet, nutritionist use Food Frequency Questionnaires (FFQ), 24 hour dietary recalls (24HDRs), dietary records or dietary histories [1]. These surveys collect consumed foods or dishes, which can be transformed into energy and nutrient intake using Food Composition Tables (FCTs).

When conducting large multicenter studies in which individuals from several countries are involved, one limitation is the difficulty of data acquisition, harmonization and standardization in the different populations. In 2008, one pioneer initiative on this regard was carried out by the “European Food Information Resource AISBL” (EuroFIR

AISBL)[1], an International non-profit association (AISBL), whose aim was: "*the development, management, publication and exploitation of food composition data, and the promotion of international cooperation and harmonization through improved data quality, database searchability, standards development, dissemination and training for all users and stakeholders*". The research objective approached here is a proposal of a knowledge network repository, with four basic types of knowledge (food composition, dish composition, dietary patterns and diet-disease effects) which can enhance the EuroFIR project with new methods and techniques in the fields of large knowledge repositories, data mining, and ontology engineering.

Last June 14 in The Hague, the Joint Programming Initiative[2] (JPI) in "A Healthy Diet for a Healthy Life" conference was held and the 2010-2020 roadmap for harmonizing and structuring research efforts in the area of food, nutrition and health was presented. The goal of the JPI conference was to define the Strategic Research Agenda for the period 2011-2020 and beyond[3], which main aims are to provide a holistic approach to: (i) identify the key factors that affect diet-related diseases, (ii) discover new relevant parameters and mechanisms and (iii) define strategies that contribute to the development of actions, policies and innovative products suitable to reduce the burden of diet-related diseases. The JPI Agenda developed the corresponding subroadmap for each one of the three key interacting research areas that were identified and described in the previous Vision Document[4] of the JPI. The Research Areas (RA) are the following: RA1-Determinants of diet and physical activity; RA2-Diet and food production; and RA3-Diet-related chronic diseases.

Each research area roadmap in the Agenda presents two prime initiatives: for 2012-2014 and 2015-2019. The prime initiative for RA1 (2012-2014) is "*Establish a European transdisciplinary research network on determinants of dietary and physical activity behaviors and the relation with health and best practice implementation strategies for sustainable changes*". This initiative is a research challenge where the preparatory work is the collection, integration and assessment of monitoring systems, databases, determinants and outcome assessments. And one of the research needs to face the challenge is to establish and maintain an integrated trans-disciplinary database, with potential for secondary analysis by interested researchers with specific research hypothesis, assuming the initial data are collected according to best practice in biological, behavioral, socio-economic and environmental science traditions.

---

1 EuroFIR. http://www.eurofir.net/. (Last access in August 6, 2012).

2 JPI Conference: https://www.healthydietforhealthylife.eu/hdhlconference/ (Last access in August 6, 2012).

3 The JPI Strategic Research Agenda for the period 2011-2020 and beyond. https://www.healthydietforhealthylife.eu/index.php?index=25. (Last access in August 6, 2012).

4 The JPI Vision Paper (September 2010) https://www.healthydietforhealthylife.eu/ index.php?index=24. (Last access in August 6, 2012).

Technically speaking, the research challenge of creating a European FCT (EFCT) involves a technological challenge in the field of large databases and large repositories. The Scientific Advisory Board of the JPI, called DEDIPAC, claimed that the EFCT should not be a "data" or "information" database, but a knowledge network repository with contributions of at least 27 European countries. The specific challenge to face is to organize the existing knowledge, their supporting infrastructures and their associated management requirements of the databases containing national Food Composition Tables (FCT) and their integration in a large knowledge repository. Traditionally, FCTs were tables where a portion of each single food was decomposed in energy, macronutrients and other components that are not nutrients. The standard size of the portion is 100 g, but some FTCs take the edible part of the food (i.e., discarding the peel in oranges; in this case, 100 g of edible orange), and other FTCs take the whole food (i.e., the whole 100 g of orange, including the peel). Moreover, macronutrients are grouped in families, as lipids, proteins, carbohydrates; and no nutrients are minerals, vitamins and aminoacids. Usually, each FTC register contains around 50 components. However, the number of components may vary in each FTC. Regarding national and private (academic or enterprise) FCT creation, although they can be standardized and biochemically proved, they are usually different from country to country (or depending on the academic organization or enterprise aims and resources).

With the evolution of the information and communication technologies, FCTs were converted in databases and, later, Web services were added to allow on-line access to them. But the drawbacks of the traditional FTC were inherited by the FCT databases and emerged some specific problems as, for example, the lack of service due to site saturation or network breakdowns, the restricted access only to active members (who have paid the corresponding fee), the lack of programmed access (a set of procedures to manage queries coming from applications), the native language, and so on. That is the situation of the European FCT provided by the FAO[5] or EuroFIR[6].

The aim of this paper is to discuss a proposal for designing and developing a European Repository of Knowledge on Advanced FCTs and related knowledge (food composition, dish composition, dietary patterns and diet-disease effects, and semantic connections between them) based on the existing national FCTs, their system interoperability strategies, and the cooperation of each national FCT for maintaining and updating the repository.

For achieving this aim, the following strategies are discussed in this paper: (i) a process for retrieving data from the different national resources and populate the Repository (Section 2); (ii) the viability of the current software resources and protocols that can be used to integrate the different FTC databases (Section 3); and (iii) new methods and

---

[5] FAO. Food Composition Tables–Europe. http://www.fao.org/infoods/tables_europe_en.stm. (Last access in August 6, 2012).

[6] EuroFIR How to access FCDBs. http://www.eurofir.net/food_information/food_composition_databases /how_access_fcdbs. (Last access in August 6, 2012).

techniques for generating and extracting knowledge form the Repository (Section 4). Finally, some conclusions are provided.

2 Designing a Process for Retrieving Data and Populate the Repository

The process for retrieving data from the different national resources and populate the Repository can be very complex because the national FTC databases has been developed according to each country objectives, culture, funding and interests. Thus, data structures, nomenclatures, number of food components included or, even, formats and units (English or International Metric systems: e.g. quantities in grams vs. quantities in ounces) are not shared. Moreover, each database has different access protocols and restrictions (i.e., public vs. private access, human interface vs. programed interface or both, etc.) Therefore, before starting to discuss how we could apply the technical approach, previous political work should be done searching agreements for data sharing, open access protocols and medical and nutritional interests. Despite the above mentioned complexity, the process outlines can be described in a workflow composed by four steps:

STEP1: defining a Minimal Set of FCT data (MS-FCT). The MS-FCT is the common data that holds every FCT database in the same or approached format (no need of transformation or conversion). On the other hand, the Standard Set of FCT data (SS-FCT) must be defined. The SS-FCT is the standardized data that every FCT database should contain according strategic objectives of the knowledge repository (homogeneity, integration, interoperability).

STEP2: defining the knowledge levels in the repository. Initially, we have defined the following levels (see Fig. 1):

1. **Level 1: Food Composition**. Basic knowledge about the composition of each food but with the following variations: national FCT source, determination methods for each component, local and regional variations of the food, and original language.
2. **Level 2: Dish Composition**. Knowledge about the composition of dishes in single food, the standard portions (in Metrical and English measures) and their corresponding images, the corresponding recipes (the same food mixture is different ac-cording the cooking process), and the local and regional variations in recipes and portions.
3. **Level 3: Dietary patterns**. Knowledge about discovered dietary patterns in nutritional studies using data mining strategies. From dietary patterns, it would be possible to generate dietary models to apply in the kind of studies described in the JPI research areas prime initiatives.
4. **Level 4: Diet-disease effects**. Knowledge about associations and interactions between diet and disease (via genetic and phenotypic factors), recommendations for specific populations (i.e., celiac), high risk food for specific diseases, lowering risk food for specific diseases, etc.

All together should run in cooperation with every national FCT database trust, providing full access to authorized sources, level of service and frequent updates to guarantee the quality and accuracy of the provided knowledge in the repository.

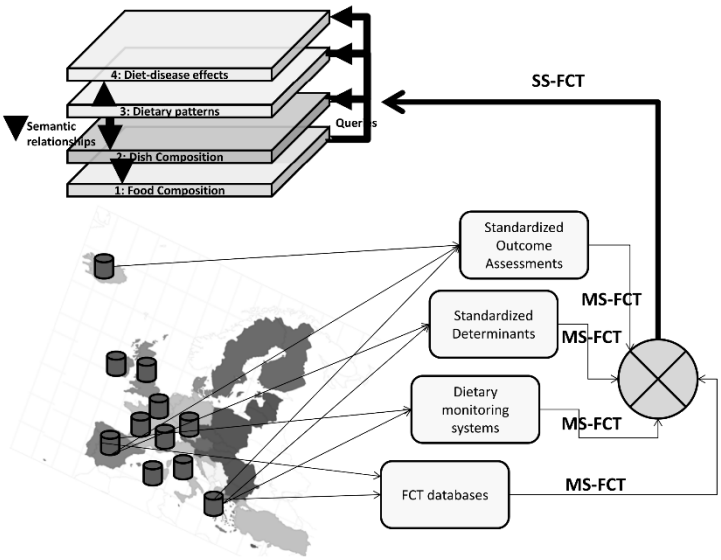

Fig. 1. Repository environment and functional structure. From each EU partner database, or set of databases (FCT, Dietary monitoring systems, Standardized Determinants and Standardized Outcome Assessments), a MS-FCT is provided. After some homogenization and integration processes, a SS-FCT is generated to update the Repository. The Queries path shows how queries between levels flow. Semantic relationships are defined only to immediate levels and show how to extract knowledge from the repository.

STEP3: studying, designing, developing and applying current software resources and protocols to integrate the different EU partners' FTC databases and other data (Fig.1), generating the corresponding sets of MS-FCT, for retrieving data from the different national resources.

STEP4: populating and maintaining the Repository, mainly injecting standardized data from the different national resources under the SS-FCT approach, but also using direct built-in methods and interfaces. It should be noted that the information is generated on national resources and not in the Repository.

3 The Viability of the Current Software Resources and Protocols

FCTs allow mapping foods or dishes with their corresponding energy and nutrients. In Nutritional Epidemiology, this is crucial due to the proved relation that exists between diet and some diseases [2], as for example, cardiovascular diseases [3-5], diabetes [6-7], and obesity [8-10], whose study requires large amounts of data for a statistical analysis. Then, the development of the proposed Large Knowledge Repository is certainly a colossal and challenging task evolving current technology and new technologies that undoubtedly have an initial cost but may pay off in the long term.

Previous works by our group [11], developed some medium scale projects in the area of medical informatics for automatizing nutritional questionnaires and calculating the nutritional composition of meals using several FCTs which used an ontology for translating the components in different FCTs to a common name. That ontology, named Nutriontology (NO), is running on an independent platform, which also contains all FTCs physical databases, applying interoperability strategies to manage the database access. Moreover, NO is part of a set of ontologies managed by an upper level ontology named NutriGenOntology (NGO). Other independent generic Web platform, named "Project", manage the set of automatized nutritional questionnaires and the participant's (and other data) database corresponding to one nutritional study. Thus, the communication between NO and a project are performed by Web services. Really, Project is a template which is instantiated in a particular platform as new nutritional studies are started and, then, the platform adopts the study name or acronym (i.e., Fituveroles, Obenutic, Obenomics, etc.) Therefore, we consider that this pilot system carried out by our group, which combines ontologies and web services in the appropriate manner, can be a start-up for achieving an integrated European FCT.

Besides, currently information repositories technology is rendered as insufficient for accomplish the integration and interoperability levels expected in such repositories, and the heterogeneity in the data is not efficiently managed. For example, the Semantic MediaWiki[7] do already consider the unit conversion problem at a very basic level. Another option, taking in account the very large scale of our proposal, is to define two wide strategies in both levels (Fig.1): level 1 with integration and interoperability; level 2 with homogenization. To integrate the different FTC databases, one suitable solution is combining semantic mappings for modelling FTC structures and semantic operations for retrieving data from the different national resources, and then, generating the corresponding MS-FTCs. Homogenization in the second level, under the SS-FCT approach, could foster the enhancement and specialization of existing data mining methods and techniques. Other solutions may be considered since some intelligent systems can cope with heterogeneity and interoperability in all levels. Then, it is too early for

[7] Semantic MediaWiki repository. http://semantic-mediawiki.org/wiki/Help:Custom_units#Converting_between_proportional_units. (Last access in August 5, 2012).

comparing the cost of addressing heterogeneity and interoperability versus the cost of homogenization in the proposed repository.

4 Developing new methods and techniques for generating and extracting knowledge form the Repository

It is necessary to define a standard language (i.e., XML-based language) for representing the Minimal Set of FCTs data and Standard Set of FCTs data, both including the basic four types of knowledge the Repository has to manage: food composition, dish composition, dietary patterns and diet-disease effects. But, the characteristics of these types of knowledge and the challenges derived from them must be identified.

The food composition knowledge tell us what elements are in one standard portion (100 g. of edible portion or net intake) of each food: macronutrients (proteins, fat and carbohydrates), micronutrients (aminoacids, minerals and vitamins), other components (water, alcohol, caffeine, etc.), and the corresponding total energy of the whole portion. In the biochemical analysis made for composing the FCT, each sample is taken from raw food, wherever possible with minor exceptions, to avoid nutrient alterations in cooking processes. Therefore, the primary source of the information is the food composition biochemical analysis performed by each national food authority. This kind of analysis is make once unless a new and better biochemical technique appears in market. The secondary source of information is the own FCT. It could be subject to change due to adding new food entries (the most usual) or reviewing the existing ones (very rarely). Moreover, there are some standards about FCT structure and organization. The derived challenge is, firstly, to homogenize FCT entries in a common set of components, nomenclatures and formats/units under the MS-FCT approach but keeping national differences; and secondly, to integrate and combine all national FCT entries in a maximal concept as it is the SS-FCT. The last one would cover lacks of data for each individual food in a FCT combining data from the rest of FCTs.

The dish composition knowledge describes the three main aspects of each dish: what food contains and in which quantity/proportion contributes each individual food, what cooking process has been applied, and what is the size of the portion. The proportion of each individual food determines the calculations of edible portions for obtaining the food composition from the FCT. The list of each individual food is not static due to national, regional, local and, of course, home variations, but keeping the main components (i.e., apple pie will not be more apple pie when apple is replaced by peach). Each kind of cooking process alters the properties of the food (i.e., vitamin or fiber degradation, fat substitution, etc.). Then, FCTs cannot be applied directly, but with cooking revisions. The size of the portion is the description of how big is and what quantity of food contains a dish. Here, a specific problem arises from the term “dish”, because we can have solid, liquid and semi-liquid food. Then, when we are describing a portion of solid food, we are using the traditional meaning of physical dish (or similar) and

measures in grams or ounces/pounds. However, when we are describing a portion of liquid and semi-liquid food, we have to use different container as glass or cup, and measures in milliliters or liquid ounces/pints. Usually, portions are categorized as small, medium and big, where each category has assigned one quantity in weight or volume, but the quantity depends of the nature of food itself. Moreover, there are not any standard (or the facto standard) about dish structure and portions, but the cooking alterations are well studied and weighted. Therefore, the primary source of the information is composed by, in one hand, published tables of cooked food proprieties; and, on the other hand, published collections of recipes in books, journals, Web, etc. The derived challenge in this case is to define a Minimal Common Recipe Catalog (MCRC) which can be used in the scientific environment for assessing dish composition in the Repository. The MCRC should include the "official" composition of each dish plus cooking variants, standardized portions and units according the food state (solid, liquid, semi-liquid).

The dietary patterns knowledge show us common profiles of food intake in persons to whom dietary assessment questionnaires were administered. Dietary patterns usually are inferred from the participants in nutritional studies and, later, can be reviewed and organized to have well-established patterns. Therefore, the primary source of the information is the set of discovered dietary patterns, and the second source is the collection of scientific publications describing other patterns. The derived challenge in this case is to achieve a standard catalog of well-established patterns for making comparisons in each nutritional study.

The diet-disease effects knowledge show us the associations and interactions between diet and diseases, when diet may act as risk or protector factor over individuals with (genetic) susceptibility to particular disease. Really, associations and interactions are not analyzed taking in account a particular meal or food, but specific dietary patterns. So, dietary patterns and disease are strongly related. Therefore, the main source of the information is the set of statistically significant diet-disease associations and interactions discovered in the nutritional studies and published in journals. The derived challenge in this case is having the maximum and accurate knowledge as possible about diet-disease associations and interactions.

5 Conclusions

A framework for designing and developing a European repository of Knowledge for Food Composition Tables is proposed with in this paper and the scenarios and the steps for constructing this repository are also described. The main outline is to construct the knowledge base in a scalable way, moving from standardized knowledge towards population-dependent knowledge. The main challenge is to integrate repositories belonging to different national states (many issues due to the use of different data structures, different nomenclatures, and different formats and units). Moreover, FCTs are extended with three additional types of knowledge, dish composition, diet patterns and

diet-disease effects, coming from other biomedical/biological data sources, for mining associations and interactions between diseases and food by means of dietary patterns.

A pilot approach was carried out by our group, which developed some medium scale projects in the area of medical informatics for automatizing nutritional questionnaires and calculating the nutritional composition of meals using several FCTs which used an ontology for translating the components in the different FCTs to a common name. Based on the success of this approach, we propose a solution to the integration of all European FCTs based on ontologies and web services, and asynchronous web technologies for assuring the minimal response time in knowledge queries, and for providing modular services, and the maximal underlying data organization.

Acknowledgements. This work has been partially funded by grants GEWIMICS (SAF2009-12304, MICINN), AGL2010-22319-C03 (MICINN), BEST/2011/261 (GVA), ACOMP/2011/145 (GVA), and CIBER "Physiopathology of Obesity and Nutrition" (ISCIII-FIS). CIBERobn is an initiative of the ISCIII.

References

1. Falomir Z., Arregui M., Madueño F., Coltell C., Corella D.: Automation of Food Questionnaires in Medical Studies: a state-of-the-art review and future prospects. Comp. Biol. Med. (in press, accepted on 25/07/2012 with DOI 10.1016/j.compbiomed.2012.07.008) (2012)
2. Feart C., Alles B., Merle B., Samieri C., Barberger-Gateau P.: Adherence to a Mediterranean diet and energy, macro-, and micronutrient intakes in older persons. J. Physiol. Biochem. (Epub ahead of print. PubMed PMID: 22760695) (2012)
3. Ganguly R., Pierce G.N.: Trans fat involvement in cardiovascular disease. Mol. Nutr. Food. Res. 56(7), 1090-1096 (2012)
4. de Oliveira Otto M.C., Mozaffarian D., Kromhout D., Bertoni A.G., Sibley C.T., Jacobs D.R. Jr, Nettleton J.A.: Dietary intake of saturated fat by food source and incident cardiovascular disease: the Multi-Ethnic Study of Atherosclerosis. Am. J. Clin. Nutr. 96(2), 397-404 (2012)
5. Hansen-Krone I.J., Enga K.F., Njølstad I., Hansen J.B., Braekkan S.K.: Heart healthy diet and risk of myocardial infarction and venous thromboembolism. The Tromsø Study. Thromb Haemost. 108(3). (Epub ahead of print. PubMed PMID: 22739999) (2012)
6. Rivellese A.A., Giacco R., Costabile G.: Dietary Carbohydrates for Diabetics. Curr. Atheroscler. Rep. (Epub ahead of print. PubMed PMID: 22847773) (2012)
7. Guldbrand H., Dizdar B., Bunjaku B., Lindström T., Bachrach-Lindström M., Fredrikson M., Ostgren C.J., Nystrom F.H.: In type 2 diabetes, randomisation to advice to follow a low-carbohydrate diet transiently improves glycaemic control compared with advice to follow a low-fat diet producing a similar weight loss. Diabetologia. 55(8), 2118-2127 (2012)
8. Corella D., Arnett D.K., Tucker K.L., Kabagambe E.K., Tsai M., Parnell L.D., Lai C.Q., Lee Y.C., Warodomwichit D., Hopkins P.N., Ordovas J.M.: A high intake of saturated fatty acids strengthens the association between the fat mass and obesity-associated gene and BMI. J. Nutr. 141(12), 2219-2225 (2011)
9. Bulló M., Garcia-Aloy M., Martínez-González M.A., Corella D., Fernández-Ballart J.D., Fiol M., Gómez-Gracia E., Estruch R., Ortega-Calvo M., Francisco S., Flores-Mateo G.,

Serra-Majem L., Pintó X., Covas M.I., Ros E., Lamuela-Raventós R., Salas-Salvadó J.: Association between a healthy lifestyle and general obesity and abdominal obesity in an elderly population at high cardiovascular risk. Prev. Med. 53(3), 155-161 (2011)
10. Foster G.D., Shantz K.L., Vander Veur S.S., Oliver T.L., Lent M.R., Virus A., Szapary P.O., Rader D.J., Zemel B.S., Gilden-Tsai A.: A randomized trial of the effects of an almond-enriched, hypocaloric diet in the treatment of obesity. Am. J. Clin. Nutr. 96(2), 249-54 (2012)
11. Fabregat A., Arregui M., Barrera E., Portolés O., Corella D., Coltell O.: NutriGeneOntology: A Biomedical Ontology for Nutrigenomics. In: Proceedings of the 2008 International Conference on Biomedical Engineering and Informatics; 2008, vol. 1, pp. 915-919. IEEE Computer Society, New York (2008)