Computer Standards & Interfaces 36 (2014) 592–597

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

Towards harmonized data interchange in food consumption data

H. Pakkala [a,*], T. Christensen [c], K. Presser [b], I. Martínez de Victoria [d]

[a] Division of Welfare and Health Promotion, Department of Lifestyle and Participation, National Institute for Health and Welfare (THL), P.O. Box 30, FI-00271 Helsinki, Finland
[b] National Food Institute, Division of Nutrition, Technical University of Denmark, Mørkhøj Bygade 19, DK-2860 Søborg, Denmark
[c] Institute of Information Systems, Department of Computer Science, ETH Zurich, Universitätstrasse 6, CH-8092 Zürich, Switzerland
[d] University of Granada, Granada, Spain

ARTICLE INFO

Article history:
Received 25 October 2012
Received in revised form 2 September 2013
Accepted 15 October 2013
Available online 23 October 2013

Keywords:
Food consumption
XML
Data interchange

ABSTRACT

Food consumption data are collected and used in several fields of science. The data are often combined from various sources and interchanged between different systems. There is, however, no harmonized and widely used data interchange format. In addition, food consumption data are often combined with other data such as food composition data. In the field of food composition, successful harmonization has recently been achieved by the European Food Information Resource Network, which is now the basis of a standard draft by the European Committee for Standardization. We present an XML-based data interchange format for food consumption based on work and experiences related to food composition.
The aim is that the data interchange format will provide a basis for wider harmonization in the future.

1. Introduction

Food consumption data are collected for a variety of purposes. They are used, for example, for clinical studies that analyze the diets of patients, for epidemiological research, public health studies, and in food-oriented risk assessment [1–4].

Food consumption data can be collected by paper questionnaires, interviews, use of computers, and sometimes over the Internet. In most cases, the collected data are stored in electronic format for further analysis. Food consumption datasets are often used with datasets of other kinds, such as with food composition data, medical records or other information relating to the study participant. Sometimes these datasets are stored in a dedicated data management system, although it is quite common that food consumption data are linked with data from other information systems or stored within other information systems. In some cases, the food consumption data are collected decentrally or in multiple study centers, and sometimes this information is collected with slightly different methods or software. In addition, it is common to link the same food consumption data for example with food composition datasets from different countries, from different years, or with datasets from different areas of science. These are just a few examples to illustrate how food consumption datasets are commonly linked, split, combined and transferred. One may even find that important consumption data are underutilized for other purposes than collected due to the apparent cumbersome processes involved in merging data from different non-comparable formats; for example, food risk assessment could benefit significantly from harmonized intake of data that has been collected for nutritional analysis. To summarize: there is a constant need for food consumption data interchange. However, there is no harmonized interchange format publicly available that is supported by the food consumption community. The data interchange is often carried out by ad hoc procedures using spreadsheets, leading to manual work that is tedious, cumbersome, error prone and expensive.

Since 2005, harmonization efforts in the field of food composition have been intensified through the European Food Information Resource Network (EuroFIR) [5]. Currently, it seems that EuroFIR has managed to achieve a critical momentum that has enabled a breakthrough in harmonizing food composition data and food composition databases and their maintenance [6–10]. In addition, an international food data interchange standard has been developed together with the food industry [11]. As food composition data are commonly used with food consumption data, perhaps a similar harmonization scheme could be adopted for the harmonization of food consumption data — especially in the data interchange. There are also other standards regarding food information, but they are focused on food production [12], food manufacturing and trade [13], and recipes used for cooking [14]. Moreover, there are several existing food description and classification systems available [15]. However, in the area of food consumption there are no commonly used data interchange formats. The existing formats have been created by separate institutes (such as by the European Food Safety Agency [16]) or the formats are used within specific information systems (such as EPIC-Soft in Europe [17] and the system of the Nutrition Coordinating Center in U.S.A [18]). There have already been some efforts [19–21] towards increasing harmonization of food consumption data, but to date they have not led to wider action.

* Corresponding author. Tel.: +358 29 524 8593.
E-mail addresses: heikki.pakkala@thl.fi (H. Pakkala), karl.presser@inf.ethz.ch (T. Christensen), tuchr@food.dtu.dk (K. Presser), igmdvc@gmail.com (I. Martínez de Victoria).

0920-5489/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csi.2013.10.004

The aim of this paper is to present a general structure for a transfer package that could be used for the data interchange of food consumption data. The aim is to be very general: the transfer format should be versatile and it should be suitable for many different tasks, while the format should not limit the purpose of use. This means that we are trying to harmonize only the structure and format — not the content itself or the rules needed to create the content or interpret it. The making of these content rules is outside the scope of this paper. Moreover, this paper does not define or take a stand on which classifications and descriptions were needed — the interchange format should not be limited to any specific classification system or description scheme. We will make use of some classifications, however, to illustrate the usage of the transfer package. We do acknowledge that harmonizing at least some of the classifications would make the data interchange even more effective when taken together with the transfer package, but again, this is beyond the scope of this paper.

2. Requirements

The starting point for the harmonized food consumption data transfer package was that it should be as general as possible and it would be possible to use it for all kinds of food consumption data, i.e. it should be generic. Moreover, the transport package should be based on existing standards and concepts.

The main methods used for the collection of the food consumption data are 24-Hour Dietary Recall, Dietary Record, the Food Frequency Questionnaire (FFQ), and the Food Propensity Questionnaire (FPQ) [22,23]. The different data collection methods produce data of different kinds and structures. The data structures needed for storing and transporting dietary recalls and dietary records are rather similar. In addition, the structures for the data from FFQ and FPQ are almost identical. Thus, the transport format may be reduced to handling two main structures: Dietary recall/record and FFQ/FPQ. In addition, it is possible to use more than one method for collecting food consumption data and across multiple days, or to make use of personal interviews, postal questionnaires or web queries. Consequently, the interchanged data should be applicable with the different methods and their requirements should be taken into account. To summarize, the aim of the transport format is to include all information and documentation needed for the interpretation of the data (without any additional information) but not to include unnecessary information, such as the layout of the questionnaires.

Extensible Markup Language (XML) is a general-purpose markup language that is commonly used in sharing structured data, especially over the Internet [24]; for example, the EuroFIR Food Data Transport Package was built using XML [10]. Consequently, XML was chosen as the markup language for the food consumption data transport package.

2.1. Data model and key concepts

The key concepts are presented in Table 1, while the data model is presented with the help of a Class diagram using the Unified Modeling Language (UML) [25] in Fig. 1. A class is a concept of object-oriented modeling and it can be used for presenting the main concepts. In addition, the associations between the classes define the concept and its relation to other concepts. The Class diagram is a schema of associations between classes. The lines between the classes present the associations.

These two should be interpreted together. The data model is quite simplified and consists of only the main elements. A black lozenge with the association is used to describe composition. The multiplicities of the associations are presented with numbers and symbols — the most common cases are 0 (zero, optional), 1 (exactly one) and * (zero or more, many). In Fig. 1, for example, every Record/Recall is described by exactly one Method (such as 24 h-recall). Moreover, a Record/Recall consists of zero or more Days. Consequently, every Day belongs to exactly one Record/Recall. It should be emphasized that the Class diagram is a rather abstract modeling construct. In the data model, we describe that every Record/Recall has zero or more days. However, in real life the number of days (per study person) was defined as part of the method, and so recall without any interviews would mean a failed interview.

Table 1
The key concepts for the food consumption data interchange.

Concept	Explanation
Study	Characteristics of the study. May include information about the sample and the target population
Study Person	Participant of the Study. Includes information for the identification of the Study Person which can be used for the linking of the Study Person to other datasets such as health records or physical activity. May include other characteristics such as sex, age, height and weight. All Records/Recalls belong to some Study Person.
Record/Recall	Dietary record or (24)h recall. The Records/Recalls belong to one Study Person and there may be several Records/Recalls. The Record/Recall includes more specific information about the method used as the same dataset may include different Records/Recalls for the same Study Person.
Day	The recorded day in some Record/Recall. There may be several days in one record. The day may include information on the nature of the day (e.g. vacation)
Meal	One eating occasion during some day. The meal is linked to a Day and through that into some Record/Recall and further into Study Person. The meal may include information about the time and the type of the meal and the place of the meal.
Entry	The smallest recordable unit in a Record/Recall. The Meal consists of one or several entries or meal rows. Entry includes the linking information to Food and the amount of consumed Food.
Food	Food item consumed in one or several Entries. Food includes food description.
Questionnaire	FFQ or FPQ questionnaire. There may be several (different) Questionnaires for one Study Person. The Questionnaire consists of Question Sets.
Question Set	The Question Set connects together several of the same kinds of Questions. It is common that several questions use the same scale and some general information for the study person such as 'How often do you usually eat the following foods?'
Question	The smallest recordable unit in a Questionnaire. The Question includes a label or a title (i.e. the question row presented for the study person such as 'Meat products'). The Question includes information about the frequency and in the case of the FFQ, also quantification information such as portion size.
Scale	The Scale is used with the frequencies and with the quantification of the Question. The scale defines the interval and potential values of a specific question.
Method	The Method can be used for the specifying of the Record/Recall or the Questionnaire e.g. '24 h food recall'

The data model has three major classes: Study Person, Record/Recall and Questionnaire. The Study-class simply keeps the Study Persons together and every Study Person may have Records/Recalls or Questionnaires. The Record/Recall has an inner structure (from Day to Entry) and consequently the Questionnaire–Question Set–Question forms another main structure.

2.2. Food consumption data transport package

The XML implementation of the data transport package follows the data model (Fig. 2). In addition, it has a section (Sender Information) that includes information about the dataset provided, such as the name of the organization and the contact information.

Fig. 3 shows an example of the XML schemata, the XML schema for Meal. The XML schema has its own grammar for defining the elements, their order and data types. Moreover, it is possible to define whether some element is mandatory or not and whether it can occur several times and these features can be used for the validation of the XML data file (it is validated against the XML schema). The example from the Meal shows that Time is mandatory and the Meal can include