Publication and alignment of authoritative vocabularies for food: A GODAN <> IC-FOODS WG

Content alignment and technical platforms for alignment and publication

1. Background and rationale

1.1 Domain authorities

There are different authoritative institutions that set the standards for different aspects of the food value chain, primarily EFSA, FAO, WHO, WTO, USDA, FDA.
While these institutions all have a specific mandate and look at food from slightly different angles (human health / food safety, food trade, agronomy, food in the agricultural value chain), in their work they all need to refer to common concepts like food products types, food sources, food components, organisms… And indeed over the last decades each institution has developed and maintained official classifications to be used in their respective area of authority. Many of these classifications cover areas of common interest, like food product types, product / commodity coding systems, food composition.

- **EFSA** maintains FoodEx (now **FoodEx 2**), a comprehensive food classification and description system aimed at covering the need to describe food in data collections across different food safety domains. (There are also former European food classifications like the CIAA Food Categorization System or the Eurocode-2)
- The **UN system** maintains several classifications:
    - The UN Statistical Division maintains the **Central Product Classification** (CPC) and the **Harmonized Commodity Description and Coding Systems** (HS) (created by the WCO)
    - WTO still maintains their Harmonized System, periodically amended with the WCO one
    - FAO maintains the **FAO/WHO Codex Classification of Foods and Animal Feeds** part of the Codex Alimentarius, the **INFOODS Food Nomenclature, Terminology and Classification Systems** and Food Component Identifiers (Tagnames), the Definition and Classification of Commodities, and an extension of CPC for agriculture
- **USDA** maintains the USDA **National Nutrient Databank for Food Composition** and the USDA **Branded Food Products Database**.

- **FDA** has a database of food additives and developed an International Interface Standard for Food Databases.
- Big **research centers and universities** also have a role in adopting and validating classifications and in experimenting with new innovative types of semantic technologies for the representation of food and nutrition data.

Although there is collaboration between these institutions and reciprocal awareness of each other's work, at the moment there isn't agreement on a shared set of terms even for common aspects of the terminology, nor any agreed and formalized alignment between these classifications.

The consequence is that organizations, companies and government bodies use different classifications depending on the most relevant authority they have to comply with, or in some cases have to use more than one to comply with different authorities, or when there are no normative constraints, they use their own. This makes it very difficult:
a) for all actors in the food value chain, including the intermediaries that create hardware and software for them, to identify the relevant classifications and keep them up to date; and
b) to integrate food value chain data across geographic regions and across different parts of the value chain (for instance, ideally, in a global distributed food tracking system).

On the one hand, perfect alignment between existing vocabularies may not be possible, as each classification includes concepts relevant to specific organization directives and interests and not necessarily common to the work of others. On the other hand, shared perspectives and overlapping classification system "facets" provide opportunities for vocabulary alignment and data sharing across institutions. Examples are food product types and food components.

Besides, the identification of the aspects or "facets" that are very specific to the mandate of each institution and those that are of common interest could also help to create a more efficient environment where each institution focuses on its most specific mandate and contributes to a common interlinked pool of classifications / ontologies in which common or aligned concepts would be maintained collaboratively.

It is also worth noting that none of these classifications seems to have been published in makes the reuse of concepts from these classifications less easy for applications than it would machine-readable form and with Uniform Resource Identifiers (URIs) assigned to concepts. This be if the classifications were published following semantic technologies and the Linked Data approach.

While the domain authoritative bodies would have all the knowledge to align the contents of classifications and to decide which pieces if any could be maintained collaboratively, the technical aspects of classification alignment and maintenance, like editing platforms and

technologies to link concepts and publish vocabularies, would of course require a different type of expertise.

This is where another type of authoritative bodies may be of help.

1.2 Vocabulary authorities

General standardization bodies like ISO also publish classifications for food products. In the case of food products, ISO publishes technical specifications rather than vocabularies.
Authoritative bodies in the area of vocabularies are W3C for RDF and Linked Data vocabularies and the OBO Foundry for scientific ontologies.
Something interesting has already been done in the area of food classifications around:

- **Integrating different food classifications.**
  The most important experiment in this area is probably the **LanguaL Thesaurus**. Originally created in synergy by several organizations (among the original members of the International LanguaL Steering Committee were FDA, USDA, International Agency for Research on Cancer (IARC-WHO)), it is now maintained by Danish Food Informatics. It's a broad thesaurus for describing, capturing and retrieving data about food.
  In particular, it includes sub-classifications from different EU, US and international classification systems (like CIAA, Eurocode2, USDA, Codex Alimentarius).
  LanguaL facilitates links to many different food data banks and contributes to coherent data exchange. In total, more than 40000 European, North American foods and foods from other countries are now LanguaL indexed.

- **Reusing classifications in ontologies.**
  In the area of vocabulary development, some classifications have been transformed into simple RDF concept schemes (SKOS), others have been reused in or transformed into ontologies.
  Several scientific ontologies have been published in the OBO Foundry. The OBO Foundry is a collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate.

  An interesting ontological experiment in the area of food involving the reuse of existing classifications is **FoodON**, currently maintained by University of British Columbia. FoodON is a new ontology built to interoperate with the OBO Library and to represent entities which bear a "food role". The aim is to develop semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes. FoodON also formalizes as classes all the food product types that are in LanguaL.

Another interesting project on alignments in the broader agri-food area is the Global Agricultural Concept Scheme (GACS), which mapped a subset of concepts from three agricultural thesauri and published new common URIs as a new global core thesaurus.

The OBO Foundry and the owners of LanguaL and FoodON, a well as the GACS team, would have the expertise and tools to support the technical aspects of aligning classifications and publishing a common set of concepts using semantic technologies and following the Linked Data approach.

1.3 IC-FOODS

An initiative that is putting different actors together for building the future “Internet of Food” is IC-FOODS, led by UC Davis. The initiative is focused on vocabulary aspects and involves mostly Universities and vocabulary authorities.

“IC-FOODS brings together the brightest minds in ontological, computational, and mathematical modeling from around the world to work together to aggregate, design, and develop standardized, human and machine readable vocabularies and ontologies that advance the nascent fields of Food Systems, Food, and Health Informatics […] Enabling us to build the next generation, semantically enabled Internet of Food (IoF)"
Main partners are various Universities (UC Davis, Berkeley, Oxford…) and the OBO Foundry, but the first conference in November 2016 saw among its participants also representatives from the private sector, research bodies like the Max Planck Institute and national authorities like USDA.

GODAN would like to make the link between the domain authorities that maintain the most important classifications and the IC-FOODS group, in particular the partners who can provide the technical support and the platforms for vocabulary work.

2. Purpose

UC Davis is therefore proposing a new GODAN Working Group on “Alignment of authoritative vocabularies for food” under the coordination of UC Davis and the IC-FOODS initiative, bringing together the key domain authorities and other partners who can provide the platforms and methodologies to work on and publish vocabularies.

The WG will:

- **Survey** the existing classifications managed by the domain authorities who agree to participate in the WG; survey and collect mapping work already done, identify essential missing mappings.
- **Design**: decide which types of vocabularies are most appropriate for publishing the classifications (simple concept scheme, ontology…); agree on a methodology (quality criteria, curatorial pipelines…); discuss level of granularity and distribution of classifications/ontologies (one or several linked ones, distributed or centralized); identify vocabulary editing and linking platforms.
- **Governance**: decide on institutional responsibilities, ownership, and related technical issues like stable URIs, possible collaborative maintenance etc.
- **Publish** the classifications and their mappings following Linked Data best practices.

- In cases in which this has not been done already and where deemed essential, do the remaining **alignment** work.
- Identify **use cases** that demonstrate the usefulness of the RDF publication and alignment.
- Provide basic vocabulary **services** to reuse terms and mappings in tools and portals, primarily for the identified **use cases**.
- If deemed useful, agree on which concepts are common and can be published as an **authoritative common set of URIs** and which concepts should remain in different specialized vocabularies.

The technical approach is very similar to the one adopted for the above mentioned GACS. A link with the GACS group would be desirable, as well as with the GODAN / Research Data Alliance WG “AgriSemantics”, working on broader issues around semantics and shared vocabularies for agri-food data. Also the participation of representatives from the OBO Foundry is desirable to learn about their ontology environment and plans for the “OBO Food Foundry”.

The final objective of this work is to provide a set of semantically interlinked URIs for key food product concepts as an infrastructural component that will facilitate the development of both software for actors in the food value chain and added-value integrated services that need to track food products.

3. Participants

Proponent: UC Davis under the IC-FOODS project.

Members: teams from FAO (FAO Nutrition and FAO Statistics), EFSA, IFPRI, UK Food Standards Agency, USDA, University of British Columbia Hsiao Lab, OBO Foundry, GACS, Tufts University, Gent University Department of food technology, safety and health. .

Prospective partners: UN Statistical Division? WHO? Danish Food Informatics?.

4. Proposed Deliverables

We keep deliverables to the minimum. The actual content and extent of the first deliverable will depend on the priorities of the partners.

- The published set of linked URIs
- Essential web services for lookup and cross-walks, designed around the identified use cases.
- A report describing the methodology followed, the challenges, the solutions and the results, as lessons learnt and inspiration for other similar endeavours.