**TAB Review of « Wheat data interoperability » WG proposal**
By F. Genova and B. Plale
Delivered to Council 6 Sep 2013

**Summary**

This WG is fully included in an international framework with high potential economic and societal value since it is part of the "Wheat initiative" proposed by research and funding organisations from several countries, and supported by the G20 agriculture ministers. It is also in the context of the "New Alliance for Food Security and Nutrition" in which G8 leaders agreed to share their relevant agricultural data with African partners. It must be noted here that this is an excellent occasion to give a "political" visibility to the RDA.

The problem undertaken by this group is important. The case statement as it is currently written confuses a paper product deliverable from a true Linked Data solution. TAB suggests that the case statement be revised to clarify the nature of the deliverable, that the group revise its timeline to include a 4-month digest/assimilate/discuss phase after the survey is done to allow for discussion and agreement on how to move from a set of unrelated ontologies and vocabularies to a Linked Data framework. That is, build towards a true Linked Data solution. Finally, the WG should view itself as a 2 stage WG (i.e., 2 WGs with membership continuity), with current deliverable as stage 1, and the second stage that takes the results to demonstrably useable Linked Data solution.

*Criteria*

There are measurable outcome: a Wheat linked data framework specification ("cookbook") on how to produce wheat data that are easily sharable, reusable and interoperable, which will have been evaluated in the Wheat initiative Information System. It will be taken up by the relevant communities and will foster data sharing and exchange. In view of the participant expertise and of their motivation (including the fact that the work is driven by the Wheat initiative) this seems do-able.

*Impact*

This WG is fully included in an international framework with high potential economic and societal value since it is part of the "Wheat initiative" proposed by research and funding organisations from several countries, and supported by the G20 agriculture ministers. It is also in the context of the "New Alliance for Food Security and Nutrition" in which G8 leaders agreed to share their relevant agricultural data with African partners. It must be noted here that this is an excellent occasion to give a "political" visibility to the RDA.

The WG is part of the Wheat initiative Information System (WheatIS), which has been specified through a survey of the relevant scientific community. It aims at "providing a common framework for describing, representing, linking and publishing Wheat data with respect to open standard". This is a key aspect of data sharing which do require specific actions, especially when, as in the case of the WheatIS, the initiative involves several data type and involves different scientific communities.

Participation of INRA, a leading partner in the Wheat initiative, of FAO, which has the topic of food security at the core of its international mandate, and of CIMMT, a maize and wheat improvement centre with headquarters in Mexico and offices in the developing world, ensures that the results will be used and disseminated. One target in the medium term is to adapt the "cookbook" produced by the WG to other important crops such as rice and maize. It can be expected that the following steps will be discussed in RDA Agriculture IG.

*Suggestions for Strengthening Effort*

A wheat interoperability framework could facilitate data exchange for the wheat initiative by serving as a semantic layer that links analogous terms, defines a vocabulary (metadata and data), and captures best practice in file formats. It is not clear from reading the case statement whether the intended users of the cookbook are human readers or are software tools; the current language in the case statement suggests it is either or both. This is discussed further below.

The proposed deliverable of a library of vocabularies and ontologies, however carefully identified, is a useful first step, but it will have minimal impact on truly advancing data interoperability because it contains and reflects no effort on tying the vocabularies and ontologies together. The WG should have, at a minimum, some demonstration of interoperability across different vocabularies and/or ontologies they have identified.

The proposed deliverable of a decision tree for describing data/metadata recommendations and file formats is presumably a document written for human consumption. The WG could (and should) spend a good bit of time identifying the minimal data and metadata needed to be written into the recommendations. The WG's final deliverable as a document is a useful contribution, however the case statement alludes to something larger.

The case statement suggests that the two deliverables, decision tree and library, are both cast as Linked Data, and it is the representation of the semantic information as Linked Data that will give the effort its greatest return. But for Linked Data solution to be advanced, the products must be represented in way that has coherence and consistency, and supports navigation by tools.

Ontologies' strongest use is as a consultation source by client (software) tools that would consult it, for instance, to determine if two terms are analogous or to figure out what data format to suggest. Similarly, a decision tree could be navigable by a client tool to recommend a particular format. For the framework to have powerful ontology and vocabulary features, work must be done to either create links between the vocabularies and ontologies in the proposed library, or settle on a definitive set of ontologies (and corresponding vocabularies). This effort in digesting and assimilating the results of the survey, and identifying a path forward on the ontology will need time – an estimated 4 months of time if the group is operating productively.

After the month 1-6 effort spent surveying existing standards, we recommend that the group take 4 months to digest and assimilate the results of the survey and reach consensus on an approach to representing the semantically linked information needed for wheat data interoperability that integrates and narrows to a definitive ontological and vocabulary representation for the framework that can be queried and navigated automatically by data repositories and client tools This design phase will save a lot of trouble later both because mistakes will be minimized, and more voices will be heard in the design process. Again, this could be achieved through mediating across all of the best practice (cookbook) multiple ontologies and vocabularies, starting from scratch, or narrowing to a smaller subset. The latter should be considered as the first choice is likely intractable, and starting from scratch rarely helps any longer in ontological work.

To summarize, the problem undertaken by this group is important. The case statement as it is currently written confuses a paper product deliverable from a true Linked Data solution. TAB suggests that the case statement clarify the nature of the deliverable, that the group revise its timeline to include a 4-month digest/assimilate/discuss phase after the survey is done to allow for discussion on how to move from a set of unrelated ontologies and vocabularies to a Linked Data framework. That is, build towards a true Linked Data solution. Finally, the WG should view itself as a 2 stage WG (i.e., 2 WGs with membership continuity), with current deliverable as stage 1, and the second stage that takes the results to truly useable Linked Data solution.