The landscape of agrifood data standards: From ontologies to messages.

Christopher Brewster[1]
[1]Data Science Group, TNO, Kampweg 5, Soesterberg, 3769DE, The Netherlands;
Christopher.Brewster@tno.nl

ABSTRACT

This paper provides an overview of the landscape of agrifood standards including those for crop research, farming, food supply chain and food retail purposes. The research reported here is part of a wider attempt at developing a strategic analysis of the standards, in part to identify gaps and overlapping standards, but also to provide an evaluation (both objective and subjective) of the quality and utility of the standards available. We identify two types of standards (messaging standards and ontologies) and three communities (research, farming, supply chain). The three communities develop standards largely ignoring the activities of others and even of people working in their own community. Given the ever-growing digitisation of the agrifood sector we consider what barriers there are to greater integration of the standards. If the digitisation of agrifood is to be useful for all participants, far greater efforts at integration of standards are needed.

Keywords: Data standards, ontologies, messaging standards, data integration

1. INTRODUCTION

This paper describes work in progress[1] to provide a strategic assessment of the current landscape of agrifood data standards. The more conventional use of the word "standards" in agrifood refers to "food standards" (i.e. cleanliness, hygiene, "fit for human consumption") of the sort usually meant by the UK's Food Standards Agency[2] (for example), or else standards with regard to food or agricultural practices (e.g. standards and regulations concerning the use of pesticides or herbicides, or of food ingredients) such as those recommended by the European Food Safety Authority[3]. Another closely related category are standards concerning different types of certification e.g. Global GAP[4], organic food or Fair Trade. In all these cases, the standards mostly refer to ensuring that a *process* has been followed or that the amounts of certain ingredients do not exceed a threshold.

In this paper, we are focusing on data standards, and by this we mean standards which describe the format and meaning of data used in information systems by the food and agriculture sector. These standards are used to construct computer systems which (in theory) should make them

[1] This paper reflects work in progress on a longer analytical report which will be published in the future. The work is supported by the ICT-AGRI 2 project (http://ict-agri.eu/)
[2] http://www.food.gov.uk/
[3] http://www.efsa.europa.eu/
[4] http://www.globalgap.org/uk_en/

compatible with each other, but also enable the publication and sharing of data for other purposes. In the agrifood sector the most widely used and best known examples of standards are the ISOBUS standard for agricultural machinery data, the GS1 EPCIS standard for product data encoded in barcodes and RFIDs, and the AGROVOC vocabulary for the annotation of agrifood research.

2. Selection and sources of standards

There can be no "systematic" review of standards relevant to the agrifood sector because one can encounter many different standards in use both in the academic world and in business. Equally many standards are used which have no bearing originally with agriculture but are adopted by a specific subset of users. The sources used for this report include the following:

- Reports from previous research projects, including the SmartAgriFood Project (http://www.smartagrifood.eu) and the FIspace Project (http://www.fispace.eu), especially (Mietzsch et al. 2013).
- Internet searches for specific standards and the links and documents retrieved by these means. This included searching on Google Scholar for academic research concerning a standard as well on the open web for software that implemented or otherwise used a standard.
- Discussions with colleagues in the wider agrifood research environment.
- The VEST Agroportal web site (http://vest.agrisemantics.org) provides a long list of relevant standards mostly ontologies.
- Agroportal (http://agroportal.lirmm.fr/). This is a relatively new initiative which tries to bring together all *ontologies* for the agrifood domain, with the focus being on ontologies used for research data.

A full assessment of the resources available in the last two items remains to be undertaken.

3. Types of data standards

There are two fundamental types of standards: messaging and ontological standards. Messaging standards (the best example is EDIFACT) are standards which describe how to format syntactically (and sometimes semantically) a message, usually describing some event or time related information such as an invoice or a certificate. Messages are sent because an event has occurred (a product has arrived or is ready for collection), or because an action needs to be taken (e.g. payment for an invoice). Ontological standards include vocabularies, taxonomies, classification schemes and other means to "tag" or label an object as having a specific characteristic. Examples of these types of standards include AGROVOC, animal identification standards, and Schema.org. Ontological standards usually describe a feature of a product that is always true of that product (unless it is transformed into something else). Obviously ontological standards are often used within messaging standards, although an informal assessment indicates that there could be much more uptake of ontological standards within messaging standards.

One might want to consider identification schemes which provide a single unique identifier for some entity of interest as being a different type of standard. One example of such a standard is ISO 11784 as implemented in the EC for animal identification. But we could equally consider these kinds of standards as very pared down ontologies.

4. Communities of data standards

The agrifood world consists of (at least) three distinct communities which are aware of each other to some limited extent but have produced quite separate and distinct collections of standards to handle quite different types of data. We can consider these three collections of data, with a corresponding set of standards available to label or describe that data. One collection of data is that produced by the *research community* i.e. mostly academic researchers in universities and research organisations around the world (such as the CGIAR establishments). Their data concerns the development of new crops, the performance of pesticides and herbicides, the interaction of agricultural practices with other parts of nature etc., and they have developed such standards as AGROVOC, more recently GACS (Baker 2014), and the Crop Ontology (Shrestha et al. 2010). Another collection of data is that produced by farms in undertaking farming i.e. the *farming community* and by this we mean both farmers themselves and the suppliers of material and technology to this sector. The best-known example of a standard in this area is the ISOBUS standard (ISO11783 2011), and we would suggest that this dimension has been relatively under resourced compared to other parts of agrifood[5]. The third community of data is data concerning the supply chain and the description of products as they travel along the supply chain and reach retailers i.e. the *supply chain community*. The prime examples of standards for this dimension are the GS1 family of standards (GS1 2016) and EDIFACT. These different communities of data and standards are shown in Figure 1.

[5] The development of the AgGateway initiative means that progress is being made in this area as well.

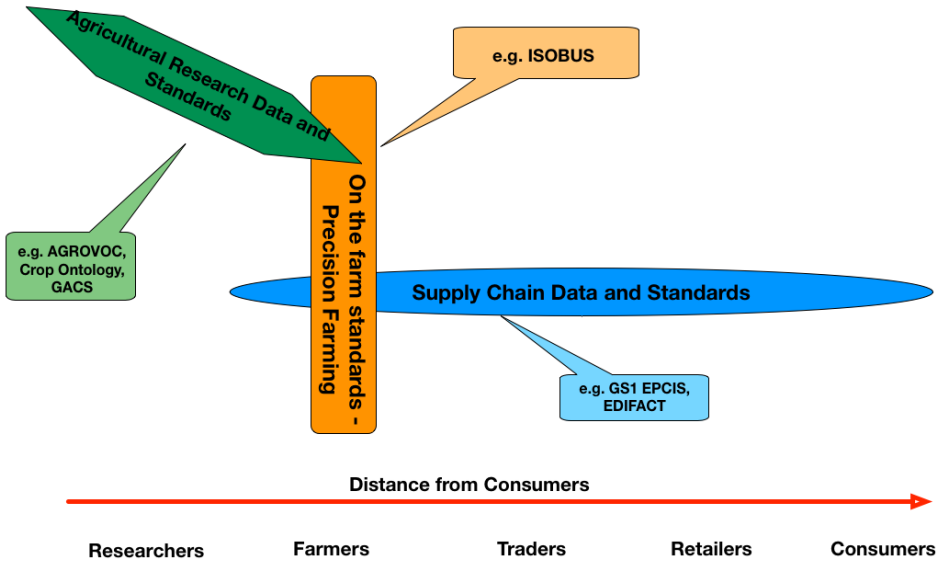

Figure 1Three communities of agrifood data standards

5. Examples of standards

In this section, we present a few examples of the standards we are referring to together with a strategic assessment of their direction of travel and likelihood for further adoption. Given the 300 ontologies identified by the GODAN "Map of Data Standards" initiative inevitably here we have only selected a few to consider for the sake of our more general point concerning the fragmentation of this area and the need for further effort at collaboration and integration. Most of these standards are well known to this community or to some parts of it.

5.1 Research community standards

AGROVOC: The AGROVOC thesaurus by the Food and Agricultural Organization of the United Nations (FAO) is nowadays the most comprehensive multilingual thesaurus and vocabulary for agriculture (Rajbhandari & Keizer 2012). It is an open resource widely used to annotate publications, more than data. The AGROVOC thesaurus contains more than 40 000 concepts in up to 21 languages covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. The AGROVOC standard continues to be used exclusively by the agricultural research world although there is obviously a considerable potential for its use in other areas outside research. There are many concepts in the ontology which are not so relevant to industry (i.e. more appropriate for the annotation of research) but the vocabulary is highly developed, strongly supported by an international community and would appear to have no danger of being abandoned or left to go out of date. There has recently been a

substantive effort to create an integration of AGROVOC with the CABI and NAL vocabularies called GACS[6].

Crop Ontology: The Crop Ontology project, a CGIAR sponsored project, was created specifically as an open resource for the digital annotation of data (Shrestha et al. 2010). It is intended to enable the annotation data about phenotype, breeding, germplasm, pedigree, traits, and other features of importance to plant breeders (Shrestha et al. 2012). This is in fact a set of ontologies based on the OBO format with the intention that they can be used for reasoning (most standards are not sufficiently formal so as to allow for reasoning). In effect, the Crop Ontology project contains a number of sub-ontologies for various plants (e.g. cassava, banana, chickpea) together with "trait" dictionaries. All ontologies are available in RDF and OBO formats. Within its own domain, this set of ontologies shave been very successful (near 200 publications cite it) but there is no indication the Crop Ontology has any impact or role beyond the plant breeding domain. Interestingly while Shrestha et al. (2010) mention a number of sources "used as references for building the CO" none include (for example) AGROVOC or any other external ontology for the definition of any concepts.

5.2 Farming community standards

AgroRDF: The standard is quite comprehensive when it comes to describing on farm operations (Steinberger et al. 2007; Martini et al. 2010). The XML standard is divided into modules which cover such topics as: Addresses; Agricultural analysis; Crops, including classes of crops; Farm data including contact persons etc.; Field including the geospatial coordinates of fields; Harvests. The AgroRDF is a semantic overlay model which references a number of existing semantic standards including GeoNames, vCard, PROV and QUDT but not to AGROVOC. The standards (agroXML and agroRDF) have had little further development since 2012. This does not mean they cannot be used but there are continuous changes particularly with regard to new sources of data that need to be integrated. Probably the most important weakness is the largely German focus, so although many concepts are labelled in English there are also many which are not. No account is taken of other languages, and a great deal more alignment and reuse/reference of other ontologies would be an improvement and increase its utility.

ISOagriNET: This standard is intended for data feeds relevant to sensors, automated feeders, milking machinery and equipment in animal husbandry. The ISOagriNet standard is in fact a collection of standards all of which have received ISO approval. The ISO 17532: 2007 is a protocol standard for M2M communication which depends on ISO 11787 (1995) for the syntax specification, and ISO 11788-1 (1997) for the data elements. The latter together with data dictionaries for specific domains provide the semantics (available from here the North Rhine Westfallen Milk Control website here http://ian.lkv-nrw.de).

This standard in spite of official ISO recognition has had limited uptake beyond a few German livestock management systems. (Tomic et al. 2015) provides a detailed description of the standard as well as a critique concerning the vulnerability of any FMIS data base to changes in the data dictionary, and the lack of "explicitly and uniquely reference descriptions of the used

[6] http://agrisemantics.org/

concepts". Tomic et al. propose an alternative approach based on turning the ISOagriNet standard into an ontology (currently not publicly available). This is a good example of a relatively isolated initiative with a limited geographical uptake.

5.3 Supply chain community standards

EDIFACT: EDIFACT is the core standard for Electronic Data Interchange (EDI). EDI has been in development since the 1980s and has achieved considerable success among large enterprises especially in such fields as the automotive industry. However, as has been long recognized EDI uptake has never been very great except for very specific sectors (e.g. Automotive industry) largely due to cost. In the agrifood sector, EDI has been used by large supermarkets to interact with major suppliers (i.e. industrial food industry) but supermarket chains also interact with large numbers of small suppliers who do not or cannot afford EDI based systems. EDIFACT is widely used in Europe according to a number of sources[7] and supermarkets have been ever more insistent that their suppliers are EDI capable. The UK for example had its own EDI standard (TRADACOMS) but more and more retailers are switching to EDIFACT[8].

From a standards perspective, the major weakness of EDIFACT is that it is largely a *syntax* with no precise semantics. A lot of elements are defined in natural language[9] but without a formal specification. According to Engel et al. "data elements transmitted in EDIFACT messages do not always have invariant semantics. Their semantics may be determined by the values of other data elements, so-called qualifiers. Hence, for the correct interpretation of a data element one has to take into account possible semantic relationships with other data elements. While these relationships are usually easy to identify from the EDIFACT standards for humans, this information is neither contained formally nor explicitly in the standards" (Engel et al. 2012). This paper presents an ontological fomalisation of EDIFACT as a so called *EDIonto* which the authors claim is built on OWL as opposed to earlier attempts which used WSDL (Foxvog & Bussler 2006). However, the work on providing as semantics for EDEFACT appears to have had little impact so far. The implications of an absence of semantics as stressed elsewhere in this report is that there is no guarantee different implementations will be able to interoperate. Note that some GS1 standards including GS1 EANCOM and GS1 XML are messaging standards compliant with the EDIFACT standard.

Overall EDIFACT is a major standard in the agrifood sector but does not integrate well with other standards (especially ontological standards). As matters stand currently there is no standardised way to interact with existing vocabularies (e.g. AGROVOC) and the syntax plus natural language explanation do not guarantee consistency and thus interoperability between systems. As a standard with a long history, it has a large installed base but does not fit with current approaches to data exchange.

[7] http://www.edibasics.co.uk/edi-resources/document-standards/

[8] E.g. https://www.webedi.co.uk/morrisons

[9] For example http://www.unece.org/trade/untdid/d99b/trmd/author_c.htm for the explanation of "Authorisation Message".

GS1 EPCIS: The GS1 organisation provides the barcodes to be found all commercial products. The EPCIS standard is intended to "enable disparate applications to create and share visibility event data, both within and across enterprises" (GS1 2016). In practice "visibility" means the ability to track objects (including all food products) along the supply chain. The standard specifies how "events" are captured so as to answer the basic questions of what, where, when and why. There are four event types (object events, aggregation events, transformation events, transaction events). The standard had its origins in the design of RFID but all modern barcode systems underneath are using EPCIS. The standard however is not formally specified so different implementations cannot be guaranteed to be compatible. The standard is integrated with the GS1 CBV (Core Business Vocabulary) which provides a classification for all products. A formalisation of this standard and the CBV was developed in Solanki and Brewster (2015) and is available on line[10]. While this is obviously one of the most used standards in the world, there is no integration of this community with other groups working towards the digitisation of the food system.

EFSA: The EFSA standard is used to describe food samples in order to track the prevalence of biological risks, contaminants and chemical residues (from pesticides and herbicides). It is freely downloadable as a set of XML files. The intention is to ensure harmonised data collection across the EU. While the EFSA standard is very specific in its design and purpose, it is a systematic well designed standard with both a syntax and semantics. The standard appears to be used exclusively for the transmission of test results from member states to EFSA, and no mention is made in the publications of any software for data exchange (although there is a web interface for uploading and validation of XML messages). The expected data format is either Excel or XML, although there are extensive XML Schema files to provide examples and documentation. Particularly significant in this standard are controlled vocabularies for product types (down to types of vegetables), agricultural methods of production (e.g. Battery production, Free range production, Organic production, Wild or gathered or hunted). Also very interesting is the concept of "Sampling Point" with extensive possibilities here (e.g. Wholesale, Retail, Import activities, Packing centre, Re-wrapping centre, Rail transport, Water transport, Catering, Take-away or fast-food outlet) in effect mapping all possible stages in the food system. Finally there is a list of over 2000 parameters (or analyte) that can be tested for (e.g. various types of Salmonella). This is an impressive, well designed standard, with a major organisation with legal authority backing it. The extensive use of hierarchical controlled vocabularies makes this a standard which could easily be formalised as RDF/OWL and aligned with other vocabularies.

6. Discussion

The core distinction we have made between ontological and messaging standards sharply divides the research community from the other two communities we have identified. This division reflects a deeper philosophical difference in perspective. Fundamentally research data tends to be static, taking a snapshot of a set of phenomena and thus the important perspective is to classify the "state of the world". If we trace most agricultural and biological research to its intellectual forebears in the work of John Ray and then Linnaeus (to mention just two names among many), the fundamental purpose was one of classification. In contrast, in farming and in the supply

[10] http://users.ox.ac.uk/~coml0597/ontologies.html

chain, the main concern is the recording of events – that a field has been plowed, that a shipment of tomatoes has been sent, that a test has been made on an animal. In such a context, the main purpose of communication and of data is to convey events and thus standards for messages have predominated in this area. The one significant exception to mention here is Schema.org which began life as a way to markup websites, especially e-commerce websites so as to make the task of search engines more effective. For some reason, describing cameras on e-commerce sites is seen as a simple classification task.

There are a number of barriers to greater integration in this area. The ontological or semantic technology approach has not made a great deal of headway to integrate with the other types of standards. This is not due to technical challenges. Classificatory or identifier standards such as GS1's EPCIS codes are used in completed other messaging standards such as those of UN/CEFACT, while the messaging standards of GS1 have been largely ignored. So in a similar manner technically, AGROVOC terms could be used by the horticulture sector to describe their consignments of tomatoes. The main barriers are psychological, lack of a business case, cultural isolation and quite simple competition:

- The *psychological* barriers are due to a combination of "not invented here syndrome" and lack of awareness and communication. These are well-known phenomena from other areas of human activity.
- *Business incentives*: In the agrifood world, there has been (until recently) limited perceived business benefit in the sharing of data or the use of standards between actors in a sector or between sectors (by which we mean everything from research to end consumer). This is in spite of considerable academic research and consultants' whitepapers envisioning use cases where data sharing would be beneficial. Thus major initiatives for the "digitization" of agriculture reflect more technology push than user demand. Perhaps one of the more glaring exceptions to this is the floriculture sector in the Netherlands which has achieved high levels of data integration in part facilitated by the Floricode coding and messaging standards (mentioned above).
- *Cultural isolation:* In spite of a highly interconnected world, both digitally and through our supply chains, there are many standards which are proposed, promoted and succeed on in certain countries and regions. Examples include AgroXML/RDF and ISOagriNET. Another form of cultural isolation is that which has created the gulf between research standards and supply chain standards.
- *Competition:* There is considerable competition between standardization bodies and a tendency to ignore existing work by another standards body. Examples of this include the recent work on a UN/CEFACT standard for laboratory results which appears to cover a similar domain as the EFSA standard. Both are messaging standards but the EFSA one is highly developed. Another example is the development of GS1 messaging standards in competition with UN/CEFACT messaging standards.

We are, however, now entering new era with regard to both the quantity and availability of data, and the demands put upon the participants in the agrifood system to make use of data in effective ways. The obvious development is the growth of smart farming or precision agriculture, and the application of IoT technology to the whole agrifood sector from farm to fork. The concomitant streams of data potentially are flowing from the various sensors deployed on the field and in farm machinery, in the food processing and transportation sector are a huge opportunity. The

potential is there for this data to be useful to the research community just as parts of the research data might be useful for on-farm or supply chain practitioners. Equally the growing demand for more detailed information about food grown sold both from regulators and consumers means that the relatively light-weight messaging standards may need to be augmented with the kind of rich data fields found in the ontological standards.

7. Conclusion

With the growth of the current generation of information technology and the development of data driven innovation in the agrifood sector (Wolfert et al. 2017), there is a fundamental need for a more joined up approach to the integration of the agrifood sector, its data streams and consequently the data standards which underpin these data flows. We believe that if effective use is to be made of data developed at the research stage for new agricultural products (seeds, crop varieties), then this has to be integrated with the food production stage for ongoing feedback, and even the performance of the food product in the supply chain and with consumer. Equally data from the farm to the final consumer needs to be able to be integrated and with such a variety of incompatible data standards this currently is excessively difficult.

8. REFERENCES

Baker, Thomas. 2014. 'Global Agricultural Concept Scheme (GACS): A Multilingual thesaurus hub for Linked Data'. Available from: http://aims.fao.org/sites/default/files/GACS_Integration_Proposal_1%200.pdf

Berge, J., 1994. *The EDIFACT Standards 2nd ed*., Blackwell Publishers, Inc.: Cambridge, MA, USA

GS1. 2016. *EPC Information Service (EPCIS) Standard.* Available from: http://www.gs1.org/epcis

ISO 11783. 2011. *ISO 11783-11:2011 Tractors and machinery for agriculture and forestry -- Serial control and communications data network -- Part 11*. Available from: https://www.iso.org/standard/57849.html

Martini, D. et al., 2010. Fitting information systems to the requirements of agricultural processes: a flexible approach using agroXML and linked data technologies. In *Proceedings AgEng 2010 -- International Conference on Agricultural Engineering -- Towards Environmental Technologies, Clermont-Ferrand September 6-8, 2010*. Clermont-Ferrand.

Rajbhandari, S. & Keizer, J., 2012. 'The AGROVOC concept scheme--a walkthrough'. *Journal of integrative agriculture*, 11(5), pp.694–699.

Rosemary Shrestha, Elizabeth Arnaud, Ramil Mauleon, Martin Senger, Guy F. Davenport, David Hancock, Norman Morrison, Richard Bruskiewich, Graham McLaren. 2010. "Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature". *AoB PLANTS* 2010; plq008. doi: 10.1093/aobpla/plq008

Solanki, Monika and Christopher Brewster. 2015. Enhancing visibility in EPCIS governing Agrifood Supply Chains via Linked Pedigrees." International Journal on Semantic Web and Information Systems 10, 45-73.

Steinberger, G. et al., 2007. Integration von agroXML in eine landwirtschaftliche

Geodateninfrastruktur. *Schweizer Landtechnik,* 62(2), pp.114–115.
Tomic, D. et al. (2015) Experiences with creating a Precision Dairy Farming Ontology (DFO) and a Knowledge Graph for the Data Integration Platform in agriOpenLink. *Journal of Agricultural Informatics*. 6 (4), 115–126. [online]. Available from: http://real.mtak.hu/30173/1/213_1071_1_PB_u.pdf.
Vest. 2016. VEST/AgroPortal Map of Standards. Available from: http://vest.agrisemantics.org/
Wolfert, S., Ge, L., Verdouw, C., Bogaardt, M.-J. 2017. Big Data in Smart Farming – A review. Agricultural Systems, 153, pp. 69-80. https://doi.org/ 10.1016/j.agsy.2017.01.023