You Can't Grow Ontologies from Bad Soil

Rob Warren[1] and Jessica Singer[1]

Myra Analytics
Toronto, Canada
{warren,singer}@myraanalytics.ca

Abstract. The use of ontologies in organizations requires an increase in data discipline and quality. This paper explores some of the challenges that were encountered during the implementation of an ontology-based Farm Management Information System in active use. Overall ontologies were a great help to documentation and a great support to data integration within different software modules. The difficulties lie in that many organizations function on an exception basis or ah-hoc manner that does not lend itself to ontological documentation. This in not unlike the growing pains that organizations face when trying to conform to complex industrial standards or implementing sophisticated software packages.

Keywords: Ontology Curation, Farm Information Management System, Agricultural Informatics, Agricultural Data

1 Introduction

We report here on our experiences building an ontologically-based Farm Management Information System. Besides modeling farm operations, the system modeled the specific provenance of each crop unit, which proved to be a great deal more difficult than recording simple crop inventories. While the use of ontologies ultimately proved to be an advantage over a traditional database approach, different organizational views of farm operations highlighted some weaknesses in current ontological engineering methodologies. This paper explores the challenges with the operational uses of ontologies.

1.1 Nomenclature and Labeling: Vernacular, Agriculture, Trade and Scientific

"What is it *exactly*?' Nomenclature is as much about the community of practice as it is about the context in which it is used. With respect to crops, farmers are not always aware of (or care about) the scientific names or even the specific variety of the crops that they are planting. Consumers care little about the scientific names of their food as much as it meets their needs. The scientific community does care about proper scientific nomenclature for cataloging and identification purposes. Ontologies are able to link these different worlds, by providing specific labels for specific groups, while ultimately pointing to the same thing, in that the thing and the name of the thing are different things.

Beyond simple taxonomical uses, more complex cases dealt with situations that broke the taxonomical paradigm, such as vernacular and/or trade names. In these cases, the brand, the cultivar and variety have all been concatenated in a single product label such as "Green Giant Peaches and Cream Corn" cans. Here, the formal name for Peaches and Cream Corn is Zea Mays (SE Bi-Colour Type) 1560 which is not used on the product label. Hence, these classes of labels concurrently encompass multiple different pieces of information such as brand name, registrations numbers, vernacular and scientific terms; where a single term entails multiple other terms.

Conversely, ontologies can also help with common industrial practices where crop grades are used interchangeably with crop names. For example, the Canadian Wheat Commission makes use of Canada Western Red Spring[1] as a crop name, but in actuality it is a blend of a number of wheat varieties which change over time. The ability to map different nomenclatures to sets of crop varieties is desirable so that utility-specific blends can be referenced while keeping a reference to the original cultivar.

1.2 Context Sensitive Description

Different users have different needs and requirements and optimize their own labeling to their specifications with a loss of precision over the original meaning. Occasionally, additional ontological structures are necessary as in the case of the Harmonized System that is occasionally used to reference crops, but actually references to a tariff to be used for importation by a certain country. Ongoing changes in trade relationship necessitate an additional ontology, through a cursory review would not make this apparent.

This uneasy relationship between term, intent and description is due to food and food production being an integral part of human society since before verbal communication was established. As crops and languages have evolved around different cultures, it is often necessary to not only have multilingual labeling but multicultural labeling. Take for example the ambiguity of the words corn and maize in English speaking countries. Ontologies permit us to localize crops in a way that would otherwise not be possible

1.3 Interoperability with the Scientific Community: Genomics and Phenotyping

Another issue was the difference between the actual identification of a crop and its reported instance, which has become common with increase in specialty crops. Previously, limited demand and widespread farming consistency meant that variety identification was a non-issue. Static seed suppliers, established commodities markets and seasonal consistency means that grain inventory variety was narrow enough to be assumed or tracked in an employees head. Only gross identification can be done on an ad-hoc basis, detailed field examinations by agronomists or inspectors have a certain limit and a full laboratory grade identification is costly. Often, the variety of a crop is recorded through its provenance, most often as a hearsay from a farm employee. Ontologically, the identification of a crop becomes complex because certain observations establish it to be part of a class or grains, but only a reported variety could be attributed.

This especially pertains to the rise in popularity of heritage grains and organic grains that have never been genetically modified in the lab. Wheats are extremely adaptable as theyve grown, thrived and been selectively bred for cultivation all over the world for millennia. Many landraces (wheats native to a certain area or that have been cultivated there for centuries), have emerged over the years and raw grain consumption volumes are hopelessly disproportionate to our capacity to test the product flow. Lastly, a number of current prominent North American varieties of wheat were introduced into the ecosystem by immigrants bringing over wheat from their countries of origin, therefore biosecurity controls were put in place. Now those varieties, such as Red Fife, make up the primary portion of North American wheats. OWL ontological axioms are attractive in these situations as we can record the provenance and quality of the wheat identification in order to better manage crop varieties.

The same issues are ongoing in the fishing industry where mislabeled fish can be marketed at a higher price to consumers unable to identify the fish. A report from Oceana Canada[5] sampling consumer fish found that 47% of samples were mislabeled when identified through DNA testing. Ontologies can provide the flexibility to record these issues by creating genetic and phenotypic databases that interact with each other and can give more accurate descriptions of what exactly, or likely, in the product being purchased. This allows for tighter quality controls on best usage, growing practices or breeding experiments.

1.4 Irreconcilable Perspectives

Interoperability is not always possible and, on occasion, work-arounds that are ontologically suspect are necessary. Two examples are reviewed here dealing with machinery organization and clerical standards. In agriculture, a variety of vehicles and machinery is used to transport crops, occasionally with confusing nomenclature. A frustrating example of these cases was the terminology for Semis, grain transporters, which proved to be irreconcilable in some situations and made communications with end-users difficult.

Both farm tractors and the traction unit of a tractor-trailer are commonly called tractors. Because grain transportation can involve multiple tractor-trailer-trailer configurations from a fleet, data entry becomes tedious. To add to the confusion "grain-trucks", permanent tractor-trailer combinations are also used as well as "grain-carts", a farm tractor with a cart attachment. Disambiguation has previously been dealt with through farm specific shorthand and task-specific forms that would be self-evident in contexts that were nearly impossible to enumerate ontologically. The fact that even regulatory agencies would deal with these situations through a great deal of "common sense" was not helpful implementation wise.

The easygoing nomenclature for documentation was a similar problem where depending on the context scale tickets, bills of lading, waybills and manifests were all used interchangeably during everyday work. Local lore can take precedence even though these documents have very different legal implications that were never before materialized because they would never be observed outside of a local context[3,4]. Strong attachments to local operating nomenclature became problematic: not only was the "name of the thing" no longer a descriptor for "a thing", but the visibility of local nomenclature in tracebility systems caused confusion and suspicion for supply chain observers.

Ontological and user interface flexibility was required to accommodate different levels of sophistication and organizational cultures. In effect, in dealing with vastly different actors, the labeling of the thing must occasionally be decoupled from its meaning. This is difficult to accept by most ontologists and agricultural professionals because in attempting to support dramatically different vues, the knowledge system can appear to be erroneous and/or inconsistent.

2 You Can't Grow Ontologies From Bad Soil

This difficulty is furthermore compounded by the literature on ontological construction and quality as it has focused on creating processes for the creation and evaluations of ontologies based on existing domain experts and authoritative data sources. As with earlier work on relational database design, there is always an underlying tension between a theoretically correct design and one that is operationally useful. After decades of strict 3NF relational database design, a number of "nosql" database alternatives are now in active commercial and industrial use with lax consistency checking that would have been previously unacceptable, signaling a change in community beliefs about this topic. Everyday reality can best be described by paraphrasing George Box: "All models are wrong, but some are useful".

The development of ontologies for operational use experiences a similar tension. Not only is making an ontology operationally useful difficult from a modeling standpoint, but established domain processes, practices and knowledge sources can be unhelpful to ontological development. The deployment of ontologically driven tools suffers from difficulties similar to that of software engineering and deployment in large corporations, perversely due to its tight coupling to the subject matter.

2.1 Organizational Dynamics And Ontologies

Non-ontological methodologies make use of a great deal of implicit knowledge that is internalized within the people and the organizations involved. This also includes cognitive shortcuts that serve as both informal optimizations or proxies for unknown information. Creating ontologies under these conditions is not unlike implementing Enterprise Resource Planning application or ISO series standardization efforts: they expose knowledge gaps, attempt to enforce best business practices and pit re-engineering efforts against institutional folklore which requires as much organizational leadership as ontological engineering.

This can also threaten some local power structures and causes discomfort because of the scrutiny [4] that is imposed over data. Warts with data quality are always present in many datasets but also oddities about how data is collected and by whom: one organization would collect paper forms titled "Crop Cleaning". When asked why the first question was "Clean Crop: Yes, No, N/A" the manager responded that no one knew. A follow-on question about whether "N/A" on all other questions stood for "Not Applicable [for this crop]" or "Not Available [at this time]" was dismissed since the checkboxes were never used and always ignored.

2.2 Human-Centric Data Collection

Interestingly, this lost context was discovered within data as well as organizations during an early attempt at creating a soil ontology for the monitoring of farm fields, an incredibly complex topic. This project was soon abandoned because the existing data sources and practices were too tightly coupled to human observations and behavior. Historically, soil slopes were reported using a relative score of 1 to 4, while more recent standards reported a 7 value quantization of grade, including an Unclassified label[2]. These are at best a gross indicator of the fields slope (with wide room for interpretation) that lacks the survey-specific data required to align different observations.

Soil composition surveys based on the “soil triangle” rank soil components based on their proportions, resulting in abridged entries such as “township loam”, possibly a scientifically acceptable value of “ditto”. Chemical analyses of soil, which should be ideal candidates for standardization, are often reported without thought as to what lab test was used or what soil sampling methodology used. The standard operating procedure no doubt specified one, but was not recorded during result aggregation limiting our ability to interpret the result.

Often an unspoken assumption of ontological development is that the ontology is “truth” when it really represents one of many truths, even when based on external authorities. Implementation experience has shown that organizations, while eager to adopt “best business practices”, are deeply culturally entrenched in their own terminology. At times this can cause frustration because this attachment is stronger than gross commercial self interest. To this end it is sometimes necessary to accept that a separate presentation layer is needed to better communicate with an end user to translate to their truth. It is an interesting comment that ontologies and their development processes are equally useful as cultural discovery tools as they force us to perform a deep self-reflection on ourselves.

3 Conclusion

The use of technology is lowering the cost of acquiring agricultural information and storing it. While at face value it would seem to be an advantage, it also creates an embarrassment of riches: there are about 7,000 varieties of apples in existence. It would take someone about 20 years to sample each type of apple but the average person will only experience the dozen or so types of apples that are grown in their region. Current ontological practice expects designers to enumerate all varieties to fulfill ontological completeness requirements (at great cost and effort), even though the majority of them will likely never be encountered.

As the amount of data increases, so does the number of potential relationships. Annotations, provenance, sampling process and normalization are now expected to be recorded for each new data point. Ontological and graph approaches can help record this additional (and necessary) information and provide a structure for technical specialists. However, the requirement can be seen by domain specialists as an unnecessary effort who take their own standard operating procedures for granted.

Thus, ontology use within an organization requires strong technical management as well as company buy-in because of the tension between ontological correctness and the everyday contradictions, exceptions and ambiguity of operational efficiency. Occasionally, unpalatable changes can be hidden from end users, but an in-depth knowledge of the business is required as ontological objects transcends normal reporting lines. In closing, the desirability of OWL-style ontologies to handle complex, multilingual information is not in question, however that same complexity requires a strong company culture for it to be leveraged properly.

References

1. Canadian Grain Commission. *Official Grain Grading Guide*. Queen’s Printer, August 2019.
2. J. S. Clayton, W. A. Ehrlich, D. B. Cann, J. H. Day, and I. B. Marshall. *Soils of Canada*. Government of Canada, Canada Department of Agriculture, Research Branch, Canadian Soil Information Service, 1977.
3. J. Esteves, J. Pastor, and J. Casanovas. Monitoring user involvement and participation in erp projects. *International Journal of Technology and Human Interaction (IJTHI)*, 1(4):1–16, 2005.
4. R. Meissonier, E. Houzé, and N. Belbaly. The it culture as an obstacle to the adoption of an erp: Case of a high-tech sme. In *Advances in Information Systems Development*, pages 279–289. Springer, 2007.
5. S. Thurston and L. Wilmot. Mislabled: Montreal investigation results and how to fix canada’s fraud problem. Technical report, Oceana, 2019.