Open Data in electronics industry

John Walker, Tim Nelissen
February 28, 2013

In the electronics industry, the ability to get accurate, timely product data in front of the customer is a very important factor in the overall business process. Furthermore, enabling the customer to easily compare and select the right product for their application from the choice of literally hundreds, or even thousands, of candidates can reduce the overall time and costs involved in the purchasing process.

Typically this product data has it's source at the manufacturer where it is stored in multiple systems in a variety of structured and unstructured formats. Often the data is duplicated in multiple places via manual processes leading to additional work and huge inconsistencies. Eventually the product data is published in formats like PDF and HTML.

Typically the data is then scraped or manually captured by data aggregation companies who align the data from different manufacturers, then sell this information on to distributors who use the data on their own websites, paper catalogues, etc. Often the distributors also do additional data capture to supplement the purchased product data.

Our goal is to simplify the overall process to reduce the time, effort and complexity required to manage, publish and use the product data, thereby reducing the costs of doing business and allowing manufacturers to get the latest information to the customer more quickly.

The approach is to provide a single, trusted source of product and product-related data in semantically rich formats that can be used to communicate the data and generate the multiple publication deliverables. Opening up access to the data is a key component, whether this is to free the data from existing silos for use within the organization, or making the data available to third parties. Also to facilitate the aggregation of data from multiple parties, it is very important to agree on a common schema that can be used to describe the products and enable easy mapping between schemata.

A key part of the approach is to use a Component Data Dictionary based on the ISO 13584 data model and IEC 61360 standard. This dictionary is basically an ontology and provides a set of classes and properties that can be used to describe instances of electrical/electronic components. The dictionary then acts as a schema that can be used to validate, but crucially also describes and defines the meaning. This highly structured data can then be used to generate publications such as PDF data sheets, web pages, selection tables and mobile apps. For less structured natural language content we use the DITA XML standard from OASIS.

For the past years we have been mainly using XML-based technologies (XSLT, XPath, XQuery, XSL-FO) as a way to store and publish the data. We have had a great deal of success in our approach, but have realized that using XML is not always the most ideal way to represent the data as the model is essentially a graph. Also working with proprietary XML schema is a

barrier to the access and understanding of the data by third parties. As such we have begun experimenting with RDF and Linked Data.

So far, most of our success has been within the enterprise, but now we would like to put more focus on the broader ecosystem with data flowing in both directions. Basically how can the parties involve provide, and make use of, more open access to the data? As we are beginning to use Linked Data, we can make use of the basic principles to allow the data to be access over the web. However, this raises a number of interesting questions:

- What formats are preferred? Our experience so far is that knowledge of RDF is quite limited, also formats like XML and JSON are not used that extensively. Many people seem more comfortable using Excel friendly formats like comma- or tab-separated.
- To what extent is RDF technology commonly used and accepted for dynamic content publication (content on demand) for instance a corporate website? So far we experienced that relational models where you need to stick to 1-n relations in which you lose a lot of semantic meaning and XML which is a tree-based hierarchical model have their limits. The real world seems to be more complex for which flexibility of the data model and use of (HTTP) URIs as global unique identifiers are important enablers. Therefore for the publication site we believe RDF offers best fit for purpose. Is this also recognized by other parties?
- How do organizations make sure the quality of data they publish is validated in an efficient way to be able to support high-quality, efficient publication? We believe this can be covered via a combination of methods: use of international standards and schema, publish content with it's data model, add linguistic checking methods and make sure you publish only fit for purpose content including business rule embedding in the publication environment. Are there any other methods that can be used or which are available to enable this quality validation process?
- What are the security and access implications? Providing totally open access to the data makes a lot of managers nervous, so how can we ensure that only public data is made public. Also in many cases products are customer-specific, so how can we manage access control to give these specific customers access to data?
- How to manage semi-structured (natural language) and unstructured (images) content and be able to easily combine all these different types of content in publications? Obviously for some content types, XML can be a better choice than RDF due to the nature of that content, but how can we make use of the best that both have to offer in the processing pipelines.
- Not another new technology? Introducing a new technology into an existing complex landscape raises many worries such as additional complexity, availability of knowledge, etc. How to convince management of the benefits to get buy in.
- Aren't we giving away a key asset? The content is the lifeblood of an organization and giving it away (for free) makes many people nervous. How can we build a compelling story that opening up not only makes sense, but can actually bring big benefits. Are there other success stories we can reference? Rather than the goal being open data, how can we position open data as an enabler to other business benefits.

- How to drive standardization within the industry? This will require the use of a common vocabulary that can be used to describe products. Could we base this on existing standards such as Good Relations or eCl@ss. What body could be responsible for the maintenance of the vocabulary and how might this work in practice. Would companies be happy to use a standard vocabulary as-is, or would there be a need to extend the vocabulary? How could data from existing systems be mapped onto such a vocabulary?

As a presentation we would like to give some examples of the type of data and content that occur, deeper insights into the data flows and some examples of the publication deliverables where this information is used. It would be great to get some insights into if this is a specific issue for the electronics industry, or if other industries face similar challenges. Is there a broader demand in this area that could act as a driver for software companies to provide standard ‘off the shelf’ solutions?