What the open data movement means for the life cycle management community

Chris Davis[1],*, Gerard P.J. Dijkema[1]

[1]Delft University of Technology, Faculty of Technology, Policy and Management, Delft, The Netherlands.

*c.b.davis@tudelft.nl

Abstract Life Cycle Management can be challenging due to the need to incorporate large amounts of information of a diverse nature, which can require a significant effort to compile. Because of this, some have argued that the act of performing an LCA is not necessarily eco-efficient itself, since the cost of the effort required may exceed the environmental benefits realized. Since this argument was first brought forward, we have seen the emergence of the World Wide Web dramatically change the economics of information gathering via lowering collection costs, and allowing for large-scale peer review and availability at an unprecedented scale. We investigate these trends, and explore how the same technologies and philosophies behind these projects can also be applied to the benefit of Life Cycle Management practitioners. This also highlights the need for researchers skilled in the multitude of means useful for managing and finding insights in large data sets. Open data is not just about using certain Web technologies, but has social dimensions that must be addressed as well. It will require actions such as rethinking of data licenses in light of the potential benefits, and it will likely lead towards a shift in business models from a focus on collection and maintenance of data to that of interpretation. Given this, we discuss some of the barriers to implementation, highlight promising initiatives already starting to emerge, and finally propose the next steps that are needed.

1 Introduction

Life Cycle Management can be challenging due to the need to incorporate large amounts of information of a diverse nature, which can require a significant effort to compile and maintain. As a result, it has been argued that performing a Life Cycle Analysis (LCA) may not always be eco-efficient due to the high economic costs that may be involved [1]. While almost a decade and a half has passed since this claim was made, and people can argue about its validity in different cases, this is still an important concern that should guide the development of LCA. Simply put, if we want to achieve widespread environmental benefits, then we need to actively drive down the costs of achieving them.

One way of reducing the cost of an LCA is to reduce the amount of effort and time needed to complete it. To an extent, the creation of several Life Cycle Inventory (LCI) databases has already led to important progress towards this goal. Other progress has been made with tools such as hybrid analysis techniques [2], which help to provide data on the background processes included in an LCA by using aggregated data about monetary flows between industrial sectors combined with aggregated emissions data. Although these are very positive developments, we still face a problem of the scale of the system that we are trying to understand and evaluate. For example, estimates have been made that a “big-box” store such as a Wal-Mart may have 100,000 different items in stock, and that there may be over 10 billion distinct products available worldwide [3]. On-line tools such as Goodguide attempt to provide consumers with LCA-based information if not indicators to select their pick out of many flavors of essentially the same product[4].

Supply-chain improvement and product selection supported by LCA, however, not only requires data but also that this data be correct, complete and up-to-date to lend the analysis results and indicators presented credibility and authority. And while we use tools such as LCA to get a systems view on sustainability issues, we need to remember “the real challenge posed by the systems idea: its message is not that in order to be rational we need to be omniscient but, rather, that we must learn to deal critically with the fact that we never are” [5,6].

This is not a call to give up on data collection, but it is rather a call to rethink how we do it. We address this challenge in this paper by exploring the supply-chain of data, the role of governments and official organization who have gone Open Data, and the possibilities of the World-wide web for crowdsourcing data. We conclude by briefly discussing barriers to and opportunities of bringing this further.

2 Making the supply chain of data more eco-efficient

We are actually facing two issues: the nature of the problems we are trying to solve, and also how we organize ourselves to solve those problems. If we want to increase the availability of data, then it is useful to consider that data is produced through supply chains just as much as physical products are. By this, we mean is that data is produced through several stages, each of which may involve different people. For example, data about a particular industrial process may come from a fact sheet published by a company, that was read by a researcher who used this as a base for additional calculations, and then published an article on a case study in her own country, which was then read by another researcher and adapted for his own work. A similar example is given by Hammond et al. [7], who in discussing environmental indicators, illustrates an information pyramid starting from primary data, which is processed into analyzed data, then used to create indicators, and finally indices.

Just as with a physical supply chain, the flow of data can be constrained at any one of these stages. This may happen deliberately through the need to keep some data proprietary, but also unintentionally through a lack of awareness of enabling tools or philosophies. As collecting and compiling data is recognized as being one of the most time consuming steps in the process of creating an LCA [8], and we should as much as possible seek to remove bottlenecks.

This analogy with supply chains leads to deeper implications when we consider that data, just like a raw resource, can flow through multiple supply chains and end up in a variety of products. This is evident since in managing the life cycle of products, parts of the same dataset may flow through various tools such as Life Cycle Assessment, Environmental Impact Assessment, Cost Benefit Analysis, Material Flow Analysis, and Ecological Footprints [9]. A diversity of tools is necessary since they each cover different scales, perspectives, and assumptions.

The last stage in making the supply chain of data more eco-efficient is that of recycling. Essentially, what happens to data when we are done with it? Does it get recycled, reused, and upgraded, or does it get lost in the landfill? If we want to make tools like LCA more eco-efficient, then we need to get better at closing our own loops. Recycling of data may equate to maintaining, refreshing and updating data, finding new purposes and applications for it.

What should be clear from this discussion is that the gathering and processing of data is not the sole endeavor of a single researcher, but rather an effort in which we are linked together in a larger interdependent system. What happens at various stages in this network affects others further down the chain.

We believe that a way forward involves leveraging Web technologies to make our research more eco-efficient. Over the past two decades, we have seen the emergence of the World Wide Web dramatically change the economics of information gathering via lowering collection costs, and allow for large-scale peer review and availability at an unprecedented scale. For example, while traditional academic publishing still runs on cycles of several months involving a hand full of reviewers, we are starting to see what some have called “Trial by Twitter” [10], where scientists are receiving feedback in real time by any of their peers who wish to join in the discussion. One can argue the merits of each approach, but through this and other means described below, the way in which we do science is changing.

In Davis and Dijkema [11], we discuss the role of Web technologies for managing LCA data through a Semantic Wiki platform. This type of platform opens up opportunities for managing the structured data of LCA (i.e. data on processes and their quantities of flows), while also allowing users to create an online record of discussions around various topics such as the quality of the data, similar processes, and details that may need to be considered for the same process existing in other geographic regions. This type of platform facilitates the recognition that data sets are never finalized, but thrive on constant discussion and review. In this paper we will discuss the larger scale developments that are occurring across the Web with regard to the availability of data. While not all of the data content of these examples is immediately relevant to managing the life cycles of products, the point is to show the types of possibilities enabled by having this content more easily available.

3 Data maintained by government agencies and official organizations

What we see happening now is a growing push towards open data, most notably by high-profile initiatives such as http://data.gov in the US, http://data.gov.uk in the UK. These government sites act as a hub for datasets generated by various government agencies. Through these sites, users are able to quickly search by terms and retrieve the relevant datasets that may have been produced by different agencies. The pages for the datasets list metadata such as the responsible agency, the release date, and available file formats. RSS feeds are sometimes available, so that people can subscribe to them and be alerted whenever updates are made to a particular dataset. Additionally, these sites also support a two way conversation where users are able to rate the dataset and provide comments on it. This model of publishing data has spread to different cities as well, such as New York City[12] and San Francisco[13]. In all these cases, it is about empowering citizens with better knowledge about what is happening in their regions, based on data that they have already funded the collection of.

Another notable example, is the World Bank's http://data.worldbank.org which has data and visualizations of 2000 indicators covering various economic, social, and environmental topics. This is taken a step further by the website Gapminder.org, which takes many of these indicators and then integrates with other indicators from the World Health Organization (WHO), Food and

Agriculture Organization (FAO), International Energy Agency (IEA). While this site is a very good example of how effective use of information technology can help to make data more accessible, it also highlights a situation where a bottleneck is not in gathering the data, but in figuring out how to sieve and extract data for different uses.

For the LCA community, there are already positive efforts such as the UNEP/SETAC database registry. The aim behind this is to serve as a hub that connects both producers and consumers of data. Producers do not have to release all of their data, but can specify the level of access they wish to provide people by default, such as only listing the name of a process and providing a link to their own site where the data may be purchased. Users can utilize the search engine on the site to search over multiple datasets. An interesting philosophy behind the project is the idea that "data quality is result of properties of a data set and what users require from the data set" [14]. We find this to be an important concept, since it shifts the administrative burden of managing data quality from a top-down to a bottom-up process. This is greatly enabled by the proposed functionality of letting users comment on and flag information that needs attention.

4 Crowdsourced data

A growing and perhaps under-appreciated amount of data comes from various crowdsourcing efforts. In other words, this is data that is compiled and managed by members of the general public. The classic example is Wikipedia, and below we will highlight several examples. We do not claim that these sources are a replacement for proper academic research, but we do demonstrate that there are some strong signals amongst the noise.

For example, within Google Earth, one of the available layers is that of the "Google Earth Community", which contains millions of user-contributed descriptions of various points on the globe. The coverage of this layer is so extensive that it is not unusual to find points in even the most remote of locations. Since anyone can contribute to this layer, the content is varied, although it is not difficult to track down useful user-contributed content such as maps of national electricity grids, 3D models of proposed wind farms, and collections highlighting areas where deforestation is occurring. In Europe, when one turns on this layer and looks at industrial facilities, they will find that someone has added the 11,000 sites recorded in the 2004 European Pollutant Emission Register [15]. This is not just simply someone posting official government data in a different file format. What has happened is that the person who posted the data has fixed it up by first setting missing coordinates to that of the nearest city, and then setting up a website where people can fix incorrect coordinates themselves. As mentioned, the content contained within this layer is of varying quality, and the point here is that just like with Wikipedia, it can often give a surprisingly useful start to investigations, but verifications and further research must still be performed.

Within the realm of LCA, there are relevant examples such as http://sourcemap.org, which is an open platform for documenting supply chains. For common consumer products, users can specify the components, the regions that they were sourced from, and the amount of CO_2 resulting from each stage of the product's life cycle. In its current state, the data is too limited for use in LCA, due to its focus only on CO_2 , and due to the fact that it lacks references to the underlying data. However, this is a positive start for educating the public about supply chains, and the desire to overcome its limitations has inspired similar efforts, such as our own work described by Davis and Dijkema [11].

The cost of recording and transmitting information has plummeted, which is leading to new opportunities and innovative uses. These trends apply to technologies such as sensor networks as well. For example, the TrashTrack project by MIT, used miniature transponders to follow 3000 pieces of trash, and other projects such as Mobile Individual Measurements of Air Quality (MIMAQ) are exploring how air quality observations can be made through wearable sensors[16].

5 Implementation and steps forward

This discussion is not ultimately about technology, but it is also about how we organize ourselves to better harness our efforts and collective knowledge. We are the creators of the infrastructure that can support this, and there are many barriers that can prevent this from happening.

Getting the data describing the life cycle of products has been very difficult due to the substantial costs involved, along with the fear of revealing strategic information or being shamed for poor environmental performance [17]. The availability of data on the life cycle of products is also influenced, due to the ability to have a competitive advantage by simply having data and acting as a gatekeeper to it. While this is an understandable strategy, it leads to a "Paradox of data quality": if we can't see it, how do we know that it is accurate? Also, this strategy increasingly faces competition from those who are moving beyond simply possessing data. Their strategy is to figure out how to effectively filter through and extract actionable information from an every growing flood of information. Continuing the analogy discussed above, value may be migrating to other stages in the supply chain of data. And beyond single supply chains, data may be linked and migrate to other supply chains, enabling multidisciplinary science and solutions.

A further question relates to how much we can achieve by working individually versus by networking our efforts. This is a serious question which has already arisen in fields such as Alzheimer's research [18]. In this case, researchers realized that due to the scale of their research, the only way to move the field forward was to open data and enable collaboration on a large scale. We are seeing similar issues within the field of life cycle management. While quality and consistence of life cycle data are important, it has been noted that "within specific databases, these are ensured to some extent, [although] across databases there can exist significant differences"[9]. This is not a technical problem, but rather one of the challenges that must be faced in better organizing ourselves.

While some of the tension of these different concerns will not go away, we should still be aware of how the growth of open data can open up new opportunities. For example, Fung and O'Rourke [19] highlight how the United States Toxics Release Inventory has brought about reductions in emissions, not through traditional command-and-control regulations, but by providing relevant information to private citizens, interest groups, and firms, which then leads to an alternative style of emissions regulations driven by pressure of grassroots organizations. This is not just about having data, but it is about the mechanisms that the availability of and access to data enables.

6 Conclusion

Through this paper we have described new developments that may help to make the practice of life cycle management more eco-efficient. These developments are leveraging the Web as a platform and may provide us with the means to reduce, both in terms of time and money, the cost of achieving environmental benefits. These trends are not occurring due to technology alone, but rather in concert with new types of thinking about how we can use these technologies to better harness our efforts and collective knowledge. Notable examples of this are occurring not just with traditional organizations such as government agencies, but also through bottom-up initiatives by members of the public who have a concern in these topics. This points towards a changing environment where traditional business models based on that of possessing data face increasing competition from those who have the skills to filter, interpret, and effectively use this data to solve environmental problems.

7 References

[1] Schaltegger, S. Economics of life cycle assessment: Inefficiency of the present approach. *Business Strategy and the Environment* 6(1), 1–8. 1997.

[2] Haes, H. A. U. de, Heijungs, R., Suh, S., and Huppes, G. Three strategies to overcome the limitations of Life-Cycle assessment.*Journal of Industrial Ecology* 8(3), 19–32. 2004.

[3] Beinhocker, E.D. *The origin of wealth: Evolution, complexity, and the radical remaking of economics*, Harvard Business School Press, 2006.

[4] <http://www.goodguide.com/>, (Accessed 13.04.2011).

[5] Ulrich, W. C. West Churchman-75 years. *Systemic Practice and Action Research.* Vol. 1, No. 4, 1988, pp. 341–350.

[6] Ryan, A.J. What is a systems approach? *eprint arXiv:0809.1698*, 2008

[7] Hammond, A., Adriaanse, A., Rodenburg, E., Bryant, D., and Woodward, R. *Environmental indicators: a systematic approach to measuring and reporting on environmental policy performance in the context of sustainable development*, World Resources Institute. 1995

[8] Rebitzer, G., Ekvall, T., Frischknecht, R., Hunkeler, D., Norris, G., Rydberg, T., Schmidt, W.P., Suh, S., Weidema, B.P., and Pennington, D.W., *Environment International* 30(5), July 2004, pp. 701–720.

[9] Finnveden, G., Hauschild, M.Z., Ekvall, T., Guinée, J., Heijungs, R., Hellweg, S., Koehler, A., Pennington, D., and Suh, S. *Journal of Environmental Management.* Vol. 91 No. 1, October 2009, pp. 1–21.

[10] Mandavilli, A. Peer review: Trial by twitter. *Nature* Vol. 469, No. 7330, 2011, pg 286.

[11] Davis, C. and Dijkema, G. P. J. A semantic wiki approach for peer review and collaborative curation of LCA data. Life Cycle Management 2011 Conference, Berlin, Germany. 2011 .

[12] <http://www.nyc.gov/html/ops/nycstat/html/home/home.shtml>, (Accessed 13.04.2011).

[13] <http://datasf.org/>, (Accessed 13.04.2011).

[14] Ciroth, A. *UNEP/SETAC database registry: Current status, expected outcomes, timeline.* Technical Report Life Cycle Initiative. September 2009.

[15] <http://bbs.keyhole.com/ubb/ubbthreads.php?ubb=showflat&Number=1028648>, (Accessed 13.04.2011).

[16] <http://mimaq.org/> (in Dutch), (Accessed 13.04.2011).

[17] Dijkema, G. P.J. and Mayer, I.S. Corporate strategy and Product-Oriented environmental policy. 6th World Congress of Chemical Engineering Melbourne, Australia. CD-ROM paper 642. 2001.

[18] Kolata, G. Rare sharing of data led to results on Alzheimer's *The New York Times.* August 12, 2010, pg. A1.

[19] Fung, A. and O'Rourke, D. Reinventing environmental regulation from the grassroots up: Explaining and expanding the success of the toxics release inventory. *Environmental Management,* Vol. 25, No. 2, 2000, pp. 115–127.