Language-acquisition inspired sustainability modelling for application profiles

Emma Tonkin
University of Bristol
United Kingdom
e.tonkin@bristol.ac.uk

Abstract

The ongoing accessibility of digital material is challenged by the constantly changing environment in which it exists. In particular, application profiles are threatened by a number of factors such as loss of context, social change and linguistic change. In this paper, we draw on observations taken from a number of application domains to build simple mathematical models for community growth and change, to explore the impact of community structure on the sustainability model required for application profiles over time. Finally, we discuss the use of similar models in evaluating application profile sustainability in general, and lessons to be drawn for DCMI.

Keywords: application profile; sustainability; user community; implementation

1. The application profile

The concept of the application profile is widely used in the world of Dublin Core, and expresses the idea that metadata, as it is experienced by its user communities, is situated within its context of use. To quote Heery and Patel (2000), 'implementors use standard metadata schemas in a pragmatic way'; those making day-to-day use of implemented systems are very likely to make use of the system to fulfil their task to the greatest extent possible. Ideals of semantic purity seldom survive exposure to the furnace of everyday pragmatism.

Application profiles reflect interdisciplinary boundaries and 'ways of seeing' (Berger, 1972) and may therefore be viewed as artefacts worthy of evaluation and exploration in their own right. Much as Olson (1998, 2001, 2002) makes use of library catalogues in the exploration of 'the cartography of marginalised domains' (Olson, 1998), so the creation and use of metadata application profiles provides a mirror through which practitioners may view institutional and individual practice.

Few of us explore the mirror images that application profile development makes available to us, with justification, given that these are functional artefacts intended to support the development of a computer-supported system that solves a problem. Invisibility could be said to be a design goal in application profile development: when the user finds themselves wondering about an application profile, it may plausibly imply that the profile has failed to achieve a stated goal. For practitioners, an application profile attracts little interest, beyond the question of whether it adequately reflects the needs of those working in the domain or with the system. Far less do practitioners find their gaze trapped, like a mythical Narcissus, in the reflection of their work. Indeed, it could be said that what Heery and Patel (2000) refer to as 'standards-makers' have a far greater propensity to the Narcissan fascination with reflection of self, being more often driven by the search for integrity, consistency and contemporary ideals of design and implementation.

1.1. Sustainability and the application profile

Application profiles represent a localisation of terms drawn from one or more relatively decontextualised concept spines (namespace schemas). Where parent resources may be viewed as subject to the pressures of social and cultural change (Kapitzke, 2001), the sustainability of the resource is called into question. Although metadata is one of the key pillars upon which data preservation efforts rest, it is the metadata that may cause greater concern than the preservation of data objects themselves; metadata is expensive to generate and its use can be expected to rely to some greater or lesser extent on the availability of standard components, such as metadata registries, or other components of the OAIS functional model (Day, 2002). Such components are reliant on a level of ongoing support and continuity, and (as shared resources in a broadly shared context) on a coherent multiorganisational or even multinational commitment to collaboration.

1.2 Evolution of an application profile

Application profiles themselves, representing a form of internationalisation or localisation, may be expected to suffer from the ongoing processes of change imposed by the drivers acting on that domain. Some result from changes within the organisation or community; some are the consequences of external change. Consider for example:

- external or internal political or strategic mandates
- staff turnover within an organisation
- organisational structure and project lifecycle
- changes in social attitudes

It may be gathered from this that the speed of change imposed on an application profile is not uniform. It is dependent on the characteristics of the community that the profile is designed to support. The maintenance requirements, and consequentially the sustainability of an application profile, can be expected to depend on situational and environmental factors. This broader set of contextual factors also includes the commercial, legal, regulatory and market context, which is referred to by Messerschmitt and Szyperski (2003) as the 'software ecosystem' in which any given system can be seen to operate.

Given that this short section covers a large number of factors, we cannot hope to explore all of these issues within a single paper; hence, we narrow our focus to a specific question: what is the effect of rapid change in user community on the rate of change imposed upon, and hence the sustainability of, an application profile?

1.3 Semantic evolution, shift, drift and change

In this paper, the mutability of various aspects of the system is considered. In particular, we explore the factor of *semantic evolution,* informally definable as a change in some part of a system, which typically results in a shift in the way in which a term or concept is understood. These concepts originate in linguistics, where they are primarily used in the fields of sociolinguistics or historical linguistics to describe variation in the use of spoken or written language over time or distance.

In simple terms, a semantic change is a change in the way in which terminology is used; when we begin to use the word 'cool' to mean 'I agree' or 'excellent' rather than to describe a temperature beneath that of 'hot', then we have implemented a semantic change. Semantic evolution is understood to be a destabilising factor in software ontologies (Cudré-Mauroux et al, 2006). The term 'semantic drift' is sometimes used, as with Gulla et al (2010), who define the term as 'the gradual change of a concept's semantic value as understood by the relevant community'. Gulla et al divide the term into two main areas: *intrinsic* and *extrinsic* draft, in which an intrinsic drift reflects change with respect to other concepts within the same frame of reference (such as an ontology or similar structure), and an extrinsic drift represents change with respect to the real-world referent.

Semantic change can take various forms and have been modelled by a number of researchers (Bloomfield, 1933). To a certain extent, models mirror the well-known thesaural relations of broadening (increasing the breadth of use of a term) and narrowing (reduction in the breadth of use of a term), although many other dimensions of semantic change are tracked by various models.

Baruzzo et al (2009) remark that 'preservation of [digital] information is about maintaining the *semantic* meaning of both the digital object and its content'; social change plays a significant role in patterns of change observed within the user community, and hence user requirements evolve over time. For Baruzzo et al, semantic evolution occurs within three *evolution dimensions*, including

- the informational domain (metadata and knowledge organisation)
- the technological domain (technological infrastructure, human-computer interaction issues and information transfer issues)
- the social domain (human and organisational factors, legal, social and procedural change)

We may hypothesise that semantic change is particularly likely to occur in situations in which items or systems are not often accessed or used. As Kanhabua (2013) states, items that are not in active use may require a form of 'recontextualisation' in order to retrieve the item as it would originally have been perceived. That is, in plainer terms, if we cannot remember what something was supposed to mean or how it was intended to be used or perceived, we will have to spend time and effort developing and testing a hypothesis and resolving any issues encountered along the way. Change that remains unnoticed is more likely to be disruptive, since it is unremarked and consequently uncompensated.

2. Methods: modelling for sustainability

In order to understand the likely development path of a domain, it is common to make use of a simulation-based modelling approach. Due to the problematically high complexity of software systems, models are generally designed with the intention of a simplified representation of some subset of the domain. The interdisciplinary nature of sustainability evaluation means that models are often interdisciplinary in focus, reach and usage; There are a large number of modelling approaches designed or applied to support sustainability evaluation. For example, Penzenstadler et al (2012) reviewed available literature for sustainability in software engineering, identifying a number of models proposed by authors over time.

Models proposed include, amongst others:

- conceptual and reference models designed towards specific areas of sustainability, such as the GREENSOFT model (Naumann et al, 2011), which are themselves typically used as inspirations for specific modelling instances rather than serving as operative models in their own right; the GREENSOFT model, for example, powers various subprocedure models applied through creation and manipulation of UML sequence diagrams, guidelines, checklists and so forth;
- agent-based models (Axelrod & Tesfatsion, 2006);
- evolutionary theory (Safarzyńska et al, 2012);
- probabilistic approaches making use of Bayesian networks (Calero et al, 2012);
- ontology-based ecosystem modelling (Franch et al, 2013);
- goal-oriented techniques for stakeholder modelling, using modelling languages such as *i**, essentially a graph-based modelling approach (Cabot et al, 2009);
- cognitive modelling and fuzzy inference (Rajaram & Das, 2010).

Selecting an appropriate model clearly depends on the model's purpose: in the words of Box (1987), 'Essentially, all models are wrong, but some are useful'. Prior to choosing a model, we must therefore define our purpose, which, in our case, is the development of a model that models the effect of factors identified in Section 1.2 of this paper on the evolution of the application profiles.

In this instance we explore the use of a model that to our knowledge has not previously been used for the purpose of sustainability modelling, but which has previously been used for the analogous purpose of computationally modelling the acquisition of language: a straightforward model of language acquisition. A discussion of computational modelling in language learning may be found in Kaplan et al (2008), although detailed evaluation of the model's original purpose exceeds the scope of this paper.

2.1 A simplified model of language acquisition

For the purposes of this paper, we apply a simple model based loosely on Niyogi (2006) and comparable to that discussed by Kaplan et al (2008). We make the following assertions: firstly, we accept that the linguistic knowledge and behaviour that underlies an application profile can be described as a formal system (Niyogi., p.37), and that human agents hold a range *H* of these systems. In order to successfully learn any given system *h* under this model, an individual must be exposed to events in which the term is used by a competent speaker of *h*. Secondly, we assert that *h* may be learned completely by an agent new to this system, by means of learning all terms used within the system. Finally, the process of learning a given term depends on two factors: exposure to at least one situation in which the term is correctly applied, which provides an opportunity to learn, and on the learnability *l* of the term. Learnability here refers to the probability that a given event in which an individual is exposed to a usage of the term will lead to a successful acquisition of the term. An individual who has been successfully exposed to all terms within *h* may be viewed as a competent user of *h*.

This model is unrealistic for several reasons: it discounts the possibility that a number of variants of any given system *h* may exist, whereas in practice variation within a formal system is likely to occur. Similarly, it presumes that a system must be completely learned in order for a user to be classified as competent. Additionally, it presumes that agents are entirely dependent on exposure to events in which terms are used to develop an understanding of how terms should be used. In practice, agents may also learn from documentation, although the learnability of examples given in documentation may diminish over time, as Kanhabua (2013) suggests, hence decreasing the accessibility of the material and reducing the efficacy of the documentation.

3. Qualitative case study: Continuous and discontinuous communities

In this section, we apply the model described in Section 2, above, to two sample cases. The first case describes a close-knit team with low staff turnover, which regularly makes use of an application profile. This case is similar to that found in many museum or archive contexts, in which continuity of practice is a significant factor. The second case describes a team which establishes an application profile, uses it for a certain period of time and then disbands; the data is then retrieved by another team, which attempts to make use of the application profile in question. This case resembles that often found in scientific research contexts, in which a project-driven team works for a certain period of time; the data and metadata created is preserved, and may well be retrieved at some later date for use in another context, such as a rapid innovation event or a later research project.

As a further simplification, we assume that the learnability of all terms in each case is total (i.e. l=1). Both case studies are dependent on the probability of the learner agent receiving evidence about terms in set *h*. Consequentially, this behaviour can be represented by a Markov chain.

We assume an application profile of ten terms, t_1-t_{10}. We assume an equal probability that any of these terms are used, although in practice, evidence of the active usage of application profiles shows us that some terms are used markedly more frequently than others (Dushay & Hillmann, 2003). Hence, a learner with moderate competence is more likely to be confident on commonly used terms.

3.1 Application profile acquisition in a highly connected team context

A learner has a high probability at any time of encountering a learning event. We model the transition matrix accordingly; a section of the full transition matrix is shown below, showing the initial state and the final absorbing state, in which a learner has correctly grasped all terms and has therefore fully completed the learning process.

$$P(x)=\begin{bmatrix} 0.1 & 0 & 0 & 0 \\ 0.9 & 0.2 & \cdots & 0 \\ 0 & 0.8 & \cdots & 0 \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

In accordance with the high probability of observing events from which they can learn (i.e. expert uses of the terminology), the learner rapidly begin to learn terms. Once the process has begun, they learn rapidly.

3.2 Application profile acquisition in a sparsely- or disconnected context

In a context in which no learning events take place, it is clearly impossible for a new learner to become fluent, since no term learning events can occur. There is no need to model this explicitly since it is trivially clear that the transition matrix is empty, and cannot lead the learner to a productive state.

Instead, we model a context in which learning events take place with relatively low frequency (a 1:10 ratio relative to the first community). Whilst this still permits learning, it reduces the probability that any given simulation timestep will be*productive* (that the learner will learn something new during that timestep). We therefore alter the transition matrix to take account of this assumption.

$$P(x)=\begin{bmatrix} 0.01 & 0 & 0 & 0 \\ 0.99 & 0.02 & \cdots & 0 \\ 0 & 0.98 & \cdots & 0 \\ 0 & \cdots & \cdots & 1 \end{bmatrix}$$

We expect this to reduce the learner's learning rate relative to the highly connected case.

4. Results and discussion

For each case in Section 3, we apply the transition matrix to a starting vector representing the initial state of our learner: [1 0 0 … 0]. The transition matrix is re-applied until equilibrium is reached, which in the case of this model concretely means that the learner has completely learned the terms in the application profile.

Graphically evaluating the results of cases 3.1 and 3.2 in figure 1, we find that the results comply with our expectations, showing that our learner picks up term usage rapidly in the connected state, and slowly in the sparse state. We have also observed that a learner without opportunity to learn will not acquire terms in this model, although in practice alternative learning strategies would undoubtedly be applied, such as learning from available documentation or available exemplars of use. Whilst an explicit model of this is beyond the scope of this paper, we remark that reduced learnability would have the result of slowing the process of learning further, stretching the S-shaped curve.

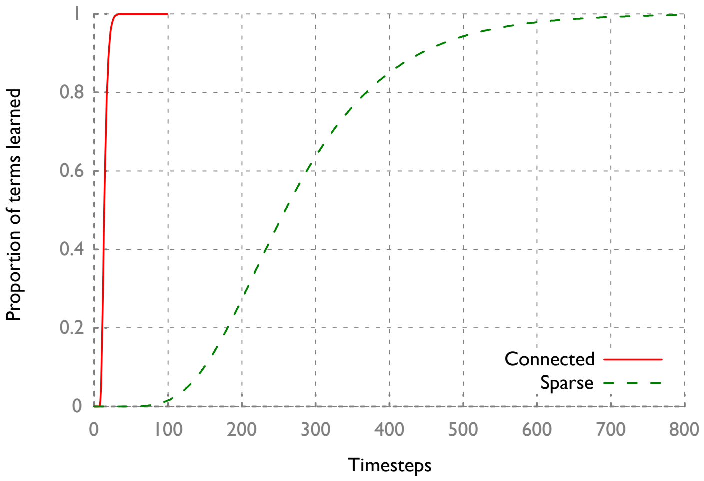

Figure 1: Model of term acquisition in connected and sparse community states

A further point of note is that, while the term acquisition rate varies significantly between the states, the curve itself does not. The *S*-shaped curve occurs in both states. Similar curves appear in many discussions of language change (see Niyogi 2006, pp. 29–30).

If insufficient exemplars or documentation were available, we would find an extremely low probability that a learner would successfully learn the usage of certain terms. This could have a number of possible effects. A learner might simply fail to learn any usage of the term, effectively truncating h by excluding the term entirely. Alternatively a learner may learn a differing interpretation of the term, resulting in the learner developing (and propagating) a variant form of the termset h, which we might refer to as h'. In the event that this occurs, this learner has experienced and will propagate a semantic change within the termset.

4.1 Discussion

We have shown that the availability of an active community has a significant effect upon the learner's ability to develop an understanding of terminology. We have also discussed that a would-be learner without an active community from which to learn must rely on available exemplars which act to demonstrate terminology in use, as well as upon formal documentation. Since the accessibility of such resources, following Kanhabua (2013), may be expected to diminish over time, we expect that the learnability of terms degrades as time passes. We also expect that the fidelity of the learner's understanding of the term may likewise be subject to change, resulting in an increased likelihood of change in the way that the learner chooses to apply terms. If the learner actively makes use of the terminology acquired, this has a relatively high probability of resulting in propagation of acquired semantic term shift to future learners.

The predictions made by this model appear to fit well with intuitions about these two cases, but it is important to stress that the model significantly simplifies events. In particular, we have made the assumption that a learner who is directly exposed to the use of a term by an expert user learns it with perfect fidelity, which we know is not the case. In practice, learning may be a partial or incomplete process, which raises the probability that a variant form of h will be created and come into use. In the event that a variant is created, a large and active community may prove to be *more* prone to propagating the variant, just as they would be more likely to rapidly learn any termset. This is especially true if it proves to be 'fitter' in an evolutionary sense than the original. For example, if a variant fits a group of users' needs better than the original, the variant will be more attractive and hence propagate more rapidly than the original, although this aspect of the model is out of scope for this paper.

4.2 Risk management

A useful outcome of a sustainability model is the ability to power decision-support applications on the individual and organisational level. This model uses observable features of a terminology set in use, notably a combination of community size and level of connectivity, to estimate, in the absence of detailed information, the 'learnability' (in terms of time cost) of a terminology set. The first of the proposed extensions to this model allows the effects of time to be modelled, drawing a distinction between a venerable application profile that is in frequent use and a similarly aged application profile that is in a state of abandonment. The second permits probable fidelity of duplication to be estimated; the practical use of such an approach is likely to depend on validation against real-life datasets. If validated experimentally, however, this model permits us not only to discuss the 'vitality' of a metadata artefact in terms of user count, but also to take into account the effects of periods of disuse and discontinuity in user community. Finally, it also permits us to take into account the likely effects of community structure and size on semantic shift and eventual evolution, where semantic evolution is here defined as propagation of opportunistic or accidental changes that prove to be beneficial to users.

It may with justice be remarked that the likelihood of popular metadata artefacts suffering from temporary abandonment or periods of disuse is low, and this is certainly the case. However, in many domains, especially in the experimental sciences, we find that temporary uptake and use of a metadata standard is a common phenomenon, and is often aligned to the vagaries of funding as well as to trends within the relevant research community. In such cases it is common to see temporarily active 'islands' of usage of specialist standards; understanding the likely outcome of this pattern is useful in understanding how artefacts resulting from such activity may best be understood, preserved and shared.

4.3 Metadata management best practices

Existing best practice in the domain of metadata management handles change (popularly termed evolution) of metadata schemas via an all-or-nothing approach: either a term is deprecated, or it is not; either a term is used, or it is not. Provenance has therefore become extremely significant in DCMI terms as the number of extant records continues to rise, as provenance metadata provides us with useful clues as to the characteristics of each record. Yet with attentive observation of an application domain, it is likely to become possible to actively and explicitly track change, information that can be used to guide further use of schemas and application profiles themselves and to guide our use of the information annotated: it is also a useful resource in mapping change within the application domain itself. For now, many questions remain: how do we gather and store such information? If it were available to us, how might we make use of it in our thinking and practice?

5. Conclusion and further work

In this paper, we have made use of a model inspired by theories of language acquisition to explore the effect of sparse and connected community groupings upon a learner hoping to develop an understanding of the usage of a specialised termset such as an application profile. This model suggests that scenarios involving discontinuity or high rates of change in community membership are more likely to suffer from issues with making use of that application profile. To increase the speed of term acquisition under these circumstances, users will be more likely to make use of lower fidelity learning strategies, including access to documentation, which unless updated becomes less accessible over time, and the use of undocumented exemplars from which to learn. We suggest that these are likely sources for semantic change. Finally, we remark that some occurrences of semantic change may have beneficial effects on the pragmatic usefulness of the termset, and are therefore likely to propagate within the relevant user community when they do occur; hence, while group discontinuity reduces the speed of adoption of termsets, we also expect it to increase the proportional likelihood that semantic evolution occurs; we expect to explore this possibility in future work.

References

Axelrod, R., & Tesfatsion, L. (2006). Appendix AA Guide for Newcomers to Agent-Based Modeling in the Social Sciences. Handbook of computational economics, 2, 1647-1659.

Baker, T., Dekkers, M., Heery, R., Patel, M., & Salokhe, G. (2001). What terms does your metadata use? Application profiles as machine-understandable narratives. *Journal of Digital information*, *2*(2).

Baruzzo, A., Casoto, P., Dattolo, A., & Tasso, C. (2009). Handling Evolution in Digital Libraries. *IRCDL*, *9*, 34-50.

Berger, J. (1972). Ways of seeing. London: BBC.

Bloomfield, Leonard (1933), Language, New York: Allen & Unwin

Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

Cabot, J., Easterbrook, S., Horkoff, J., Lessard, L., Liaskos, S., & Mazón, J. (2009, May). Integrating sustainability in decision-making processes: A modelling strategy. In *Software Engineering-Companion Volume, 2009. ICSE-Companion 2009. 31st International Conference on* (pp. 207-210). IEEE.

Calero, C., Moraga, M. Á., Bertoa, M. F., & Duboc, L. (2015). Quality in Use and Software Greenability.

Cudré-Mauroux, P., Aberer, K., Abdelmoty, A. I., Catarci, T., Damiani, E., Illaramendi, A., & De Tré, G. (2006). Viewpoints on emergent semantics. In Journal on Data Semantics VI (pp. 1-27). Springer Berlin Heidelberg.

Day, M. (2004). Preservation metadata initiatives: practicality, sustainability, and interoperability.

Dushay, N., & Hillmann, D. I. (2003). Analyzing metadata for effective use and re-use.

Franch, X., Susi, A., Annosi, M. C., Ayala, C. P., Glott, R., Gross, D., & Siena, A. (2013). Managing Risk in Open Source Software Adoption. In ICSOFT (pp. 258-264).

Gulla, J. A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., & Cerrato, O. (2010, April). Semantic Drift in Ontologies. In WEBIST (2) (pp. 13-20).

Heery, R., & Patel, M. (2000). Application profiles: mixing and matching metadata schemas. *Ariadne*, *25*(September).

Messerschmitt, D. G., & Szyperski, C. (2003). Software ecosystem. Understanding an Indispensable Technology and Industry. Massachusetts Institute of Technology, Cambridge, MA.

Kanhabua, N., Niederée, C., & Siberski, W. (2013). Towards concise preservation by managed forgetting: Research issues and case study. In Proceedings of the 10th International Conference on Preservation of Digital Objects, iPres (Vol. 2013).

Kapitzke, C. (2001). Information literacy: The changing library. *Journal of Adolescent and Adult Literacy*, *44*(5), 450-456.

Kaplan, F., Oudeyer, P. Y., & Bergen, B. (2008). Computational models in the debate over language learnability. Infant and Child Development, 17(1), 55-80.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge: MIT press.

Olson, H. A. (1998). Mapping beyond Dewey's boundaries: Constructing classificatory space for marginalized knowledge domains. *Library trends*, *47*(2), 233-254.

Olson, H. A. (2001). The power to name: Representation in library catalogs. *Signs*, 639-668.

Olson, H. A. (2002). *The power to name: locating the limits of subject representation in libraries*. Kluwer Academic Pub.

Penzenstadler, B., Bauer, V., Calero, C., & Franch, X. (2012). Sustainability in software engineering: A systematic literature review.

Rajaram, T., & Das, A. (2010). Modeling of interactions among sustainability components of an agro-ecosystem using local knowledge through cognitive mapping and fuzzy inference system. Expert Systems with Applications, 37(2), 1734-1744.

Safarzyńska, K., Frenken, K., & van den Bergh, J. C. (2012). Evolutionary theorizing and modeling of sustainability transitions. Research Policy, 41(6), 1011-1024.

Thibodeau, K. (2002). Overview of technological approaches to digital preservation and challenges in coming years. The state of digital preservation: an international perspective, 4-31.