How to Model Names in Topic Maps – Pitfalls and Guidelines

Hendrik Thomas¹, Tobias Redmann², and Bernd Markscheffel³

¹ KDEG, School of Computer Science and Statistics, Trinity College Dublin, Ireland ² University of Applied Sciences Berlin, Germany

 $^{3}\,$ Technical University of Ilmenau, Ilmenau, Germany

{hendriktho,tobias.redmann}@gmail.com, bernd.markscheffel@tu-ilmenau.de

Abstract. Names can be a tricky thing to represent due to their natural ambiguity, plurality of meanings and context dependency. In this paper we will discuss 10 common pitfalls in dealing with names and based on that present three modelling templates for complex name structures in Topic Maps.

1 Introduction

Ontologies and in particular Topic Maps [1] are an important component for the implementation of the Semantic Web vision. Despite the efforts in establishing unique published subject identifier (PSI) [1, 2], from a pragmatic point of view, human readable names are still one of the best and most common ways to identify a topic subject. However, names can be a tricky thing to model because of their natural ambiguity, plurality of meanings and context dependency [4]. Due to the natural diversity in the modelling process there is also always more than one valid way to express names [1]. As a result pragmatic guidelines, modelling templates and best practice recommendations are highly needed. However, publications focusing on issues related to modelling of names in Topic Maps are rare [1, 2] and primarily focused on technical aspects. Also only sparse information has been published on the modelling of meta-data for names, e.g. provenance, valid context.

In this paper we address the question: how to model complex name structures in Topic Maps? In particular, we ask which general pitfalls related to the identification and expression of names needed to be considered and which modelling approaches are suitable.

2 Pitfalls for Modelling Names

To get a better understanding of the problems involved in modelling names we want to share our experience gathered in the "Digital Mechanism and Gear Library" (www.DMG-Lib.org) project [3]. In this project we extracted names automatically from a domain specific dictionary (IFTOMM) and manually from

Maicher, L.; Garshol, L. M. (Eds.): Linked Topic Maps. Fifth International Conference on Topic Maps Research and Applications, TMRA 2009 Leipzig, Germany, November 12–13, 2009 Revised Selected Papers. Leipziger Beiträge zur Informatik. ISBN 978-3-941608-06-1

50 H. Thomas, T. Redmann, and B. Markscheffel

three selected text-books. Overall approx. 4.000 names in four languages were modelled. During this project we identified the following 10 pitfalls for modelling names in ontologies:

1. Ambiguity and plurality of meanings: A subject of interests can have multiple names to describe [4]. In particular, problematic are synonyms where different names are used for the same subject [4]. Common are full synonyms (strict synonyms) which always refer to the same subject in terms of exact same meaning [4], e.g. car and automobile. However, in specialized domains we can often find *partial synonyms* which refer to the same subject only in a specific context. In terms of that meaning is similar but not always identical, e.g. long and extended are synonyms but a long arm is not the same as an extended arm. Even more problematic are homonyms where a group of words share the same spelling but have different meanings, e.g. the German name Läufer has over 30 different meanings. In addition, it is common to use specific symbols rather then words to identify subjects, e.g. J stands for Joule but is also used on car registration plates for the town Jena. Another example is the name Topic Maps - written with capital letters it refers to the technology and with small letters to a collection of specific topics and associations[2].

2. Multilingual names: Naturally in different languages different names are used to identify a subject [2, 4]. If a topic map should be shared internationally then it is necessary to include translations and their valid contexts. However, often a simple one-to-one translation is not enough because the meaning of a name and in particular the knowledge structure can be different in two languages. For example commonly the German name "Mechanismus" is translated into the English term "mechanism" and "gear" into "Getriebe" [3]. This can be represented by assigning the German and the English names to the each topic. However, this is not valid for an engineering expert because for him "Mechanismus" is the topic type and "Getriebe" is the instance. In contrast in English "gear" is the topic type and "mechanism" is the instance. Thus the content relation between the names is exact the other way around then the dictionary translation. Thus the interpretation of such semantic relations in each language needs to be considered carefully, e.g. dictionaries can translate names different depending on the addressed community. Another tricky question is, what to do if no one-to-one (x = y) translation is available, e. g. a direct translation of the term "river" does not exists in French, only the two more specific terms "riviere" and "fleuve". Also problematic is the situation if no translations is available at all, e.g. the Eskimos have many more names for snow.

3. Context specific names: A name can be valid in a clear or fuzzy defined context, e.g. a valid time frame, a region or popularity in research communities or school of thoughts. For example, in publication in West Germany the name "Koppelgetriebe" is used for a specific gear. In East Germany the same gear is labelled with the name "Kurbelgetriebe". Both communities claim that their name is the standard. Modelling multiple names is easy but the challenge for Topic

Maps lies in the expression of the valid context and thus to select the appropriate display name. However, the real challenge is the identification of such valid contexts based for example on the analysis of cross-domain communications, publications or experts.

4. Default names: A rule of thumb is, that every topic should have at least one human-readable name [1] to make it easier to identify the subject. However, the choice of a default name can be difficult. Who decides what the default name is? [4]. Instead of aiming for a *detente* in which everyone is feeling equally miserable it may be better to define a default name for each relevant context. But what if the current context is unknown or not considered in the model?

5. Evolution of names: As time goes by new names emerge, others are no longer used or their meaning has changed. For example the name AIDS for the "acquired immune deficiency syndrome" was first introduced in the late 70's. Previously no specific name was available, only a plurality of fuzzy descriptions in medical reports. Due to the research efforts in the 80's the understanding of AIDS evolved and thus the associated interpretation. For most people such historical or time differences may not be important but it is highly relevant for reuse of knowledge in a wider scope.

6. Paraphrase descriptions: Another issue, in particular relevant for the automated extraction of names, is that for some subjects no specific and established name exists [4]. Many subjects can only be described in one or multiple phrases. For example aqua planning refers to the fact that a wet street is slippery for cars. But how do you name it, if a surface is slippery due spilled milk? Multiple descriptions exist but which one should be modelled?

7. Names of persons: Other pitfalls are names of persons. One can assume that names are relative static and unique. However, this is not true, e.g. titles are added during life like Dr. or PhD. If a woman get married the last name change. In addition different names are uses for a person like art name or nickname.

8. Construction rules of names: Names and in particular formal names common in research or in library science often contain implicit knowledge, e.g. in the IFToMM dictionary [3] alternative names were encode using brackets, "couple (floating) link" is resolved in couple link and floating link. If names from dictionaries, taxonomies or classification systems are reused, it is important to know how the name is constructed because in a Topic Maps such knowledge needs to be modelled explicitly.

9. Different spellings: On a syntactical level another pitfall are different spellings of names. For example, are terms modelled in singular or plural or both? How to handle names containing more than one term, e.g. use-case or use case? How to deal with spelling errors in documents - model only correct names or misspelled once, too? From a retrieval perspective all spelling variants are useful to increase the recall but could make the Topic Maps quite complex and messy.

10. Different alphabets and transcription rules: Another pitfall for multilingual ontologies is the heterogeneity of writing principles and alphabets in the world, e.g. the Latin alphabet is common but not in Russia or China.

52 H. Thomas, T. Redmann, and B. Markscheffel

One option is to model the original names but in libraries and dictionaries it is common to transform such terms into Latin terms based on standardized transcription rules, e.g. Hanyu Pinyin is a system for Mandarin. However, these rules are heterogeneous and not always specific and comprehensive. Thus depending on the expert the resulting Latin names can be quite heterogeneous.

These 10 pitfalls are generic problems and not limited to Topic Maps but relevant for all semantic languages. Ontologies imply reuse and sharing of formal expressed knowledge [1, 2]. Thus the above described pitfall may not be relevant during the creation time of an ontology but might become relevant if the knowledge is reuse in a different context.

3 Modelling Approaches for Names in Topic Maps

The key question for an ontology engineer is two folded. One issue is the identification of all relevant names but this is clearly a domain analysis problem. The second is a rather technical one: how to model complex naming information in Topic Maps? In the following section we will present three templates to express names and associated information in Topic Maps.

3.1 Modelling Names With Topic Names

In the DMG-Lib [3] we applied the following approach which can be processed by any current TMAPI implementation. Each subject is represented by one topic. For each topic one topic name is modelled without any scopes. This is interpreted as the default name and should be the most common name used in the specific domain. In scientific domains we suggest to uses English names. The selection of a default name is a task for the domain experts and NOT for the ontology engineer. Each translation is modelled as a variant names of a given topic name. The corresponding language is modelled as a scope attribute for each variant name. Please note, that it is necessary to model the default name as a variant name, too. For example: basename "Gear" and the variant names "Getriebe" (scope German) as well as "Gear" (scope English). This redundant approach ensures that for each supported language a suitable name can be found as a variant name including the language information. This is the ugly part of the model but it is easy to process. We recommend to use common PSIs for the language topics, e.g. http://topicmaps.org/xtm/1.0/#de. Abbreviations are modelled as variant names by adding the addition scope for abbreviation (PSI http://www.tmedit.org/psi#abbreviation. Synonyms represent different names used for the same subject. They can be modelled as additional topic names within the scope synonym (PSI http://tmedit.org/psi#synonym). Homonyms refer to a common name for different subjects. Simply model each subject as a topic and assign the same name to each of them. An explicit annotation is not necessary because application can compare all names and thus identify all homonyms.



Fig. 1. Modeling Template: Express Name with Topic Names



Fig. 2. Modeling Template: Express Name with Topic Names and Reification

Furthermore, contextual meta-data are helpful to choose a name and they can be modelled using additional scopes. In particular relevant for interdisciplinary projects is the provenance of a name. A list of well-known publications which contain the name be used as references and expressed as scopes of the topic type "source". Figure 2 shows a sample of this modelling approach. Please note that GTM^{alpha} as a graphical notation for Topic Maps is used for all following graphs [2].

3.2 Modelling Names With Topic Names and Reification

The discussed pitfalls demonstrated that topic names are complex and thus many addition information are needed (e.g. context, language, time) in order to label a topic appropriately. Unfortunately the TMDM [2] provides only basic options for modelling information related to topic names, more specifically scopes and topic types. In the previous modelling template we encode all relevant aspects in scopes, e.g. provenance of the name etc. This is correct because the name is

54 H. Thomas, T. Redmann, and B. Markscheffel



Fig. 3. Modeling Template: Express Name with Topics

a valid label in the context of the specified publication BUT it is not limited to it. In other words we simply want to express that a particular name can be found in a dictionary and not really that the publication is the only valid context for the name. As a result we propose a second template in which each subject is represented as a topic and all corresponding names are assigned as simple topic name without any scopes. Every topic names is then reified as an individual new topic and suitable meta-data can be expressed using associations. Figure 3 shows a sample for the topic name Bank.

The disadvantage of this approach is that we model statements about the wrong things. Reification enables us to express statement about other statements. But in this scenario we want to make statements about the name (string) and not about a particular topic names element which is bound to a specific topic. E.g. for a topic X the name "Bank" is modelled and reified as the Topic Bank-Name-1 in the context "finances" and for a second topic Y the name "Bank" is modelled and reified as the Topic Bank-Name-2 in the context "park and garden". Thus two different reified topics and both represent different topic names of different topics thus different subjects. This is problematic because the interpretation perspective is wrong, because the original intention was to express a statement about the name "Bank" independent from meanings or relations, e.g. a different spelling version would need to be assigned to both reified topics which is not correct because the spelling version is bound to the label not to the meaning (at least in this particular case.)

3.3 Model Subject and Names as Individual Topics

According to the standard a topic can represent ANY subject. In your context we are interested in subjects as well as in the available names to address the subjects. Therefore we prose a third rather drastic approach: Every subject is modelled as

a topic with NO topic names at all. Every available name is modelled as a separate topic, because we want to make statements about the name itself (about the term or string) not the subject it refers. Each name is then linked via an association to the subject, e.g. PSI of association type http://www.tmedit.org/psi#name-of-a-subject. Relevant information about the name can be modelled with additional topics and associations. By modelling every name as a topic we can explicit express further statements about it. To say it more forcefully the building blocks of Topic Maps are topics, associations and occurrences and by using this approach we do not need the topic name element. However, in contrast to other the templates this one is semantically more correct and expressive but also more complex and not supported by most Topic Maps applications. They can process it but common functions like getAllTopicNames() will not work. Figure 4 shows a sample for this template.

4 Summary and Outlook

To sum up modelling of names is a complex task. In this paper we presented 10 selected pitfalls to highlight common problems involved in dealing with name. In addition, we discussed three modelling templates which offer either expressiveness or a simple to process model. The key problem is that currently no one knows in which context which modelling approach is appropriate. Therefore further research is needed to develop a comprehensive guidebook to make it easier to model a domain and finally to support the flexible communication of a common understanding.

5 Acknowledgements

This work is partially funded by the Science Foundation Ireland FAME project (award No. 08/SRC/I1408).

References

- Garshol, L. M: Towards a Methodology for Developing Topic Maps Ontologies, in Maicher, L, Siegel, A, Garshol, L. M. (eds.): Leveraging the Semantics of Topic Maps – 2nd International Conference on Topic Map Research and Applications, TMRA 2006, Leipzig, Ger-many, October 11–12, 2006, Berlin Heidelberg New York, Springer, 2007, pp. 20–31
- Thomas, H. Redmann, T., Markscheffel, B. GTM^{alpha} Towards a Graphical Notation for Topic Maps. in: Fourth International Conference on Topic Maps Research and Applications Subject Centric Computing, Leipzig, October 16, 2008, pp 310–316
- 3. Brix, T., Döring, U., Trott, S., et. al.: The Digital Mechanism and Gear Library: a Modern Knowledge Space, in: Knowledge Media Technologies, in: Jantke K.-P., Kreuzberger, G. (eds.): Diskussionsbeiträge des IfMK, Ilmenau, 2006
- 4. Fugmann, R.: The Five-Axiom Theory of Indexing and Information Supply. Journal of the American Society for Information Science Vol 36 (2), 1985, pp. 116–129