Modeling Names

Xuân Baldauf[1]

[1] xuan--names--2009--tmra.de@academia.baldauf.org

Abstract. This paper argues that the hierarchy between *topic name item*s and *variant item*s of the TMDM resembles a hierarchy between *name*s and particular *rendering*s of names in the real world, but for this resemblance to be a better match, the requirement for *topic name item*s to always have a *value* property should be dropped.

Keywords: TMDM, Topic Name, Variant

1 Introduction

Names are the most widely used type of links between a sequence of characters (or a sequence of sounds) on the one hand and a concept on the other hand, at least for humans. Names serve to identify particular concepts, that is:

1. When inputting a name, existing names are used as a pattern to be matched against the input to decide which concept(s) are meant by the input.
2. When outputting a name, an existing name is used as data to be produced to represent a particular concept.

Typically, names are not particularly good at identifying without a context. For example, the word “bank” may refer to a type of financial institution as well as to a type of place of shallow water. This ambiguity is why the Topic Maps Data Model [ISO13250-2] provides means to identify concepts beyond human-centric names, namely subject identifiers, subject locators and item identifiers. For these identifiers, comparably strong assumptions can be made, such as: If two identifiers[1] are equal, they represent the same concept. If two identifiers are not equal, they represent a different concept unless indicated otherwise[2].

1 of the same class; As of [ISO13250-2] section 5.3.5, one class is the class of all subject locators and the other class is the class composed of all subject identifiers and all item identifiers.

2 e.g. by attaching both identifiers to the same topic

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

These beautiful properties, however, do not mean that it is viable to avoid human-centric names in most applications of Topic Maps, because most topic maps are used by humans, after all. The TMDM provides the *topic name item*s and *variant item*s exactly for this purpose: to model human-centric names as a link between

- a concept, represented by a *topic item* in the TMDM, and
- some short artifact of data (e.g. a character sequence, a sound sequence, an icon, ...), represented by the *value* of a *topic name item* or a *variant item*[3].

2 A hierarchy of name items

However, why do two different types of these links exist? The TMDM says:

> "A topic name is a name for a topic, consisting of the base form, known as the base name, and variants of that base form, known as variant names. [...] A base name is a name or label for a subject, expressed as a string. That is, it is something that identifies the subject (though not necessarily uniquely) and can be used as a label for the subject in user interfaces. The notion of a base name corresponds closely to the common sense notion of a name. [...] A variant name is an alternative form of a topic name that may be more suitable in a certain context than the corresponding base name."

If a variant name is just a topic name that may be "(more) suitable in a certain context" (than other topic names), then *variant item*s as a whole item class could be dropped and replaced by *topic name item*s with appropriate *scope*.[4] Indeed, among others[5], Robert Barta and Lars Heuer hold that *variant item*s should be removed [SC34-0705]. Going a step further, *name*s and *occurrence*s could be merged into *characteristic*s, the only remaining distinction between them being instance of either `tm:name` or `tm:occurrence` [ISO18048-WD-2008-07].

However, the current TMDM does not say that the variant name may be more suitable in a certain context than *any or some* base name. Rather, the current TMDM says that the variant name may be more suitable in a certain context than *exactly one* base name, the base name to which the variant name is attached to. This directed link between one type of name item and another type of name item would be lost when removing *variant item*s[6]. Reading the TMDM this way, the *value* property of a *topic name item* is merely a default value, subject to be overridden by the *value* property[7] of any *variant item* attached to that *topic name item*.

The main point of this paper is: There should not be any default values for *topic name item*s; or if there should be such default values, they should not be determined solely by the author of the particular topic map, but by some algorithm which may also take into account the cultural context of the current user of the topic map.

---

3 For brevity, the case where the *variant item*'s *datatype* is IRI is ignored.

4 Or they could be replaced by *occurrence item*s in case there the *variant item*'s *datatype* is anything other than *String*.

5 Benjamin Bock as of private conversation

6 Although this loss of information could be rectified by reifying both topic names (one being the successor for the removed variant) and linking the reifying topics with an association.

7 or a resource, in case the *datatype* is IRI.

3 Natural language considerations

When, actually, do names differ? For example, consider the city which is located around 41°N 29°E on the author's home planet. One of its names is “Constantinople” (in English). Another one of its names is “Konstantinopolis” (in Turkish). Are these names different to each other? Sure, the strings begin with either “C” or “K” and they end in either “le” or “olis”. But both strings are still remarkably similar.

Consider two other of this city's names, “Konstantiniyye” as well as “قسطنطينيه”. Are these names different? Well, they do not share any single Unicode character. However, they are the written representation of the same sounds in the Ottoman Turkish language, denoted in two different scripts (the Latin-based Turkish alphabet and the Perso-Arabic-based Ottoman Turkish alphabet). Considering the sound-sequence for this name, there is no distinction between “Konstantiniyye” and “قسطنطينيه”. Considering the character sequence, there is a considerable difference, as there is no 1-to-1 matching between the letters of “Konstantiniyye” and “قسطنطينيه”, as the latter string lacks representation for some sounds[8], as usual when writing in an Arabic script.

The examples so far may be more or less related to each other, in writing, in pronunciation, in etymological heritage or any combination of these. However, now consider the names “Istanbul” (in English) as well as “İstanbul” (in Turkish). These names are apparently closely related to each other. But these names are only very weakly[9] related to “Constantinople” or “Konstantinopolis”.

Apparently, we can build two groups of names for the same city: “Istanbul” as well as “İstanbul” are put in one particular group; “Constantinople”, “Konstantinopolis”, “Konstantiniyye” as well as “قسطنطينيه” are put in another particular group. In each group, there is only one string per (written) language, while across groups, there may be multiple strings per (written) language. Members within each group are closely related, while any two members across groups are either not at all or only very weakly related to each other.

This grouping seems to fit nicely to the concepts of *topic name item*s and *variant item*s. Each particular string is mapped to its own *variant item*, each particular group is mapped to its own *topic name item*. Apparently, such group (of related strings identifying a particular concept) is a *name* itself. Any member (any such string) is a particular *rendering* of the *name*. A similar distinction exists for URLs: A URL denotes the resource itself, while the data received when accessing the URL is a particular rendering of that resource.

8 a direct transliteration of “قسطنطينيه” would be “kstntinie”

9 The “bul” of “Istanbul” and the “pl” of “Constantinople” denote “polis”, which means “city”.

Following this idea, some *name*s and their *rendering*s of the city can be summarized as follows:

name	rendering	scope of rendering
(Name #1)	“Istanbul”	English
	“İstanbul”	Turkish
	“ഇസ്താംബൂൾ”	Malayalam (Malayalam script)
	“Ստամբուլ”	Armenian
	“ఇస్తాంబుల్”	Telugu
(Name #2)	“Constantinople”	English
	“Konstantinopolis”	Turkish
	“Konstantiniyye”	Ottoman Turkish (Latin-based script)
	“قسطنطينيه”	Ottoman Turkish (Arabic-based script)
	“Κωνσταντινούπολις”	Greek
	“Konstantinoúpolis”	Greek (latinized)
	“კონსტანტინუპოლი”	Georgian (Mkhedruli script)
	“Կոստանդնուպոլիս”	Armenian
(Name #3)	“Byzantium”	English
	“Βυζάντιον”	Greek
	“BYZANTIVM”	Latin
(Name #4)	“Tsargrad”	English
	“Цѣсарьградъ	Old Bugarian
	“Царьградъ”	Church Slavonic
	“Царьградъ”	Russian
	“Царгородъ”	Ukrainian
	“Цариградъ”	Serbian
	“Ţarigrad”	Romanian

Table 1: Names for the city located around 41°N 29°E. (Compiled from Wikipedia pages about this city in different languages.)

Provided suitable topics (or sets of topics) can be found for each scope, this table is a blueprint for how to define *topic name item*s and *variant item*s for the city. For each *variant item*, the properties *value*, *datatype*, *scope*, *parent* are determined properly by this table (and the other properties have sensible defaults). For each *topic name item*, however, the property *value* is not determined properly[10]. We could employ strings like “(Name #1)”, but this is highly ambiguous with respect to the concept to be identified. We could arbitrarily choose any *value* from any *variant item*, such as “ഇസ്താംബൂൾ” or “კონსტანტინუპოლი”, and while these *value*s are highly identifying, the procedure to choose them is both arbitrary and biased to a particular culture.

10 And if the *value* property is not determined properly, the *type* property may or may not be determinable properly.

It turns out that an appropriate *value* to use (and ultimately to display to the user) is dependent on the user and the cultural background of the user. Currently, there is no single culture, language or writing system in which all literate people are literate. Thus, for many *name*s (namely those which may have more than one *rendering*; and almost all *name*s will have many *rendering*s upon becoming popular across cultures), it is inappropriate to choose any particular *value*.

Thus, (if the notion of a hierarchy between *topic name item*s and *variant item*s remains) it should not be mandatory for the topic map author to define a *value* for each *topic name*.

4 Adjusting the Topic Maps Data Model

It is clear that the TMDM will not be changed any time soon. However, once this state changes, there is a multitude of possibilities to accommodate the requirement "*Topic name item*s should not have *value* properties set mandatorily.":

4.1 Making the *topic name item*'s *value* property optional

This is the plain implementation of the requirement. This solution opens up the possibility to keep a topic map culture-neutral. Additionally, it softly requires all applications which display topic maps to users to employ more appropriate value selection algorithms (e.g. dependent on the user's language or culture). However, it allows for bad Topic Maps design (that is, choosing a culture-biased default value as a *topic name item*'s *value* and failure to denote the culture-biasedness of a particular *value*). However, precisely because this solution allows for bad Topic Maps design, it is perfectly compatible with existing topic maps. This solution may also effectively be a no-solution at all, because the weak force onto applications to employ appropriate value selection algorithms may be too weak, hence it may happen that no application supports culture-neutral topic maps, which in turn may inhibit the proliferation of those topic maps.

4.2 Removing the *topic name item*'s *value* property

With this solution, all human-centric names have to belong to *variant item*s, each *old topic name item* would be converted into a *new topic name item* (without a *value* property) + a *new variant item* (whose *value* property is set to the *old topic name item*'s *value* property). In this case, the requirement for *variant item*s to have a non-universal scope (e.g. to have at least one topic as theme) needs to be dropped as well, at least for compatibility, as many *old topic name item*s have not any topic as theme. The advantage of this solution is that sub-optimal handling of value selection for user interfaces at least becomes apparent, as the software applications which display name item's *value*s as labels for topics cannot put the responsibility of proper (user's-culture-dependent) value selection on the topic map author, as the topic map author has no possibility anymore to make an arbitrary and culture-biased choice for the

value of a *topic name item*. As a result, it is expected that this problem would get a higher probability of getting tackled by the authors of applications which display topic maps.

4.3 Removing the *topic name item*

It is also possible to remove the *topic name item*, representing a particular *name* (and thus a group of related *rendering*s of the same *name*), completely and replace it with a new *name rendering item*. Each *name rendering item* has the same properties as each old *variant item* (together with a *type* property, inherited from old *topic name item*s), but unlike *variant item*s, the *parent* of a *name rendering item* does not point to a *topic name item*, but to a *topic item* directly. The *name rendering item* would also completely replace the *variant item*. The relationship between each name *rendering* and its *name* (e.g. between "ഇസ്ലാംമൃഗ" and Name #1 in the example above) would be expressed as an association between the reifying topic of the appropriate *name rendering item* and a special *name rendering group topic* (or simply *name topic*) which, while representing many name *rendering*s, itself has no canonical name.

Advantages of this solution are the simplicity of the resulting model (e.g. there are no *variant item*s and no *topic name item*s anymore) and perfect data-structure-level-compatibility between *name rendering item*s and *occurrence item*s (which in turn allows to generalize both into *characterstic* items, as suggested by ROBERT BARTA's TMQL draft [ISO18048-WD-2008-07]). This solution is also compatible to the current TMDM, as there are straightforward conversion rules from the current TMDM devisable[11].

Once disadvantage is the complex way to retrieve an appropriate subset of name *rendering* values, as for each *name rendering group*, only one *rendering* should be retrieved. Another disadvantage is the complex way to represent *name rendering group*s in Topic Maps serialization formats (such as XTM or CTM) unless special syntax for these constructs is provided.[12]

---

11 To convert from a *topic name item* or a *variant item* to a *name rendering item*, the *parent* property is set to the appropriate *topic item*; in case of a *topic name item*, the *variants* property is removed; all other properties are preserved as is; and the missing properties (e.g. *type*, *datatype*) are set with default values. If there have been *variant item*s, the *name rendering item* representing the *variant item* and the *name rendering item* representing the *topic name item* each get a reifying topic. Each such reifying topic is linked to a newly created *name topic*, using an "is name rendering for" association. The topic representing the *name rendering item* representing an old *topic name item* also gets a special link between itself and the *name topic*. That is, the special link will be a separate "is default name rendering for" association, or the existing "is name rendering for" association will be scoped specially.

12 Special syntax for common Topic Maps structures is not unusual. The "instanceOf" elements of XTM 2.0 [ISO13250-3] or the "isa" and "ako" keywords of the current CTM language [ISO13250-6-WD-2008-05] are some examples.

5 Future Work

There are two generic problems with respect to labeling topics: “Given a set of possibly redundant and culture-specific labels for a topic, select a non-redundant list of these labels tailored to the user.” (For our example city, a solution for a user with English as first language would be the list “Istanbul”, “Constantinople”, “Byzantium”, “Tsargrad”.) and “Given a set of possibly redundant and culture-specific labels for a topic, select exactly one of these labels tailored to the user.” (For our example city, a solution for a contemporary user with English as first language would be “Istanbul”.). These two related problems are not solved by this paper, they are merely identified to be not-solvable in a culture-neutral form at the level of Topic Maps authors. Thus, these problems are pushed to the authors of user-interface-bearing topic maps applications. As these authors often cannot limit the set of topics they are dealing with, automated algorithms for solving these problems have to be found.

The awareness for sub-optimal topic name modeling needs to be raised. It may be possible to develop some “bad modeling checker” software which deterministically[13] or statistically[14] checks for certain modeling patterns considered as improvable, such as giving a *topic item* two separate *topic name items* while the *topic item* should only have one *topic name item* but also *variant item*s.

Names are not only used for synthesis (e.g. for computer output to be matched by a human) but also for analysis (e.g. for computer input to be matched by a computer). For imperfect human-centric names (as opposed to TMDM identifiers), the process of analysis may require different data (e.g. a pattern) than the process of synthesis (e.g. just a static string). For example, voice recognition of a particular word may not work well if there is only one example of speaking of that word; it may work better if there is a multitude of examples of speaking that word. It is not quite clear or standardized how to model these different types of name-data (e.g. name data for analysis, name data for synthesis) within the TMDM, although the tools (e.g. scope, name types, datatypes) seem to be available.

Similarly, *topic name item*s and *variant item*s may have other purposes beyond synthesis and analysis. Sorting is one example. While sort names are long-established in the Topic Maps standards (see [ISO13250-2] section 7.4), their usage as suggested by the TMDM may actually conflict with the view held by this paper: Sort names annotate not a particular *name*, but a particular *rendering* of a name in a particular language and script. As such, they should be attached to *variant item*s, not *topic name item*s, which is not possible with the current TMDM recommendations on how to use sort names. Thus, if sort names are used as suggested, each *topic name item* would represent a particular *rendering* of a name, the *name rendering group* itself would become unmodelable with the TMDM alone.

13 That is, depending on a particular pattern.

14 For example, depending on a linguistic corpus which matches the same *name*s in different *rendering*s (e.g. different scripts or different languages).

Acknowledgements

Thanks go to Motomu Naito who raised the issue that there is a multitude of *rendering*s of the very same Japanese *name*.[15] This paper is based on and inspired by a mailing list thread which was effectively started by Lars Heuer suggesting that nobody would need *variant item*s[16]. Thanks also go to anonymous reviewers of this paper for useful comments.

References

[ISO13250-2]:
International Organization for Standardization/International Electrotechnical Commission — Joint Technical Committee 1 — Subcommittee 34 — Working Group 3: “ISO/IEC IS 13250-2:2006: Information Technology — Document Description and Processing Languages — Topic Maps — Data Model”
International Organization for Standardization, Geneva, Switzerland (August 2006)
http://www.isotopicmaps.org/sam/sam-model/

[SC34-0705]:
Robert Barta, Lars Heuer: “Proposal to remove variants in 13250-2 Topic Maps - Data Model” (January 2006)
http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0705.htm

[ISO18048-WD-2008-07]:
International Organization for Standardization/International Electrotechnical Commission — Joint Technical Committee 1 — Subcommittee 34 — Working Group 3: “ISO/IEC WD 18048: Information Technology — Document Description and Processing Languages — Topic Maps — Query Language”
International Organization for Standardization, Geneva, Switzerland (July 2008)
http://www.itscj.ipsj.or.jp/sc34/open/1054.pdf

[ISO13250-3]:
International Organization for Standardization/International Electrotechnical Commission — Joint Technical Committee 1 — Subcommittee 34 — Working Group 3: “ISO/IEC IS 13250-3:2007: Information Technology — Document Description and Processing Languages — Topic Maps — XML Syntax”
International Organization for Standardization, Geneva, Switzerland (August 2006)
http://www.isotopicmaps.org/sam/sam-xtm/

[ISO13250-6-WD-2008-05]:
International Organization for Standardization/International Electrotechnical Commission — Joint Technical Committee 1 — Subcommittee 34 — Working Group 3: “ISO/IEC FCD 13250-6: Information Technology — Document Description and Processing Languages — Topic Maps — Compact Syntax”
International Organization for Standardization, Geneva, Switzerland (May 2008)
http://www.itscj.ipsj.or.jp/sc34/open/1044.htm

15 See http://www.infoloom.com/pipermail/topicmapmail/2009q2/007486.html
16 See http://www.infoloom.com/pipermail/topicmapmail/2009q2/007482.html