Review of dissertation thesis of Jakub Klímek "XML Formats Evolution and Integration"

The submitted dissertation thesis focuses on the important and topical problem of managing large XML schemas in the heterogeneous and evolving real world. It contains a large amount of textual material, which is straightforwardly derived from the candidate's refereed publications. 13 main content chapters correspond to 2 journal articles and 11 conference papers; J. Klímek is the first author of 8 of them and co-author of the remaining 5. The chapters are clustered into three topic areas: schema integration, schema evolution and 'additional contributions'; first-authorship publications appear in all three areas. Inclusion of material primarily authored by colleagues follows from the fact that the research activities of J. Klímek had been an integral part of the research program of the XRG research group and directly followed up on the PhD (and subsequent) research of his supervisor M. Nečaský on conceptual modeling for XML.

The chosen approach to thesis *composition*, in which largely overlapping individual papers are directly transferred to thesis chapters, actually poses a challenge for the reader. The same notions (most notably, those of PIM and PSM, but also formulae for similarity, precision and the like) are introduced in several places, often in a different form and even with different content. In particular, the reviewer identified a number of technical flaws in the description of the PSM-PIM mapping methods in Chapter 3 (see the detailed comments below), while in the slightly extended treatment of the same topic in Chapter 4 many of these flaws are already eliminated. The thesis is thus a historical walk-through over the candidate's paper-writing carrier rather than either a truly consolidated, monograph-like thesis, or at least a selection of mature, self-contained and topic-wise distinct (preferably, journal) articles, which one would expect. The delta in many individual chapters is smaller than would warrant a whole thesis chapter. This *lack of content consolidation* is definitely the most significant flaw of the thesis.

The amount of original *scientific contribution* of the thesis, if picked bit by bit along the individual chapters, can be viewed as acceptable, although not ground-breaking.
Most results revolve around the problem of mapping a platform-specific model (PSM), or a concrete XML schema, for which the PSM is a 'conceptual proxy', to a pre-existing or newly constructed platform-independent model (PIM), which is basically a UML class diagram. Even Chapter 9, ranged under 'schema evolution', merely sets up integrity constraints for interpretation of a PSM wrt. a PIM, i.e. rather bridges between the two topic areas. However, parts of the thesis go beyond this central problem: in particular, the research on schema evolution by propagation of changes via the conceptual model (presumably, primarily designed by Nečaský and Mlýnková) is extended, by the thesis author, by explicit treatment of inheritance hierarchies during the evolution.

The *methods* used are in general adequate to the types of problems addressed. The formal apparatus introduced has definitional purpose (theorems with proofs only appear in Chapter 8, which has been primarily authored by the supervisor). The methods have been *validated* via a mixture of experiments on data and more comprehensive and qualitatively described case studies. The degree of validation is a bit uneven. Chapters 3, 4 and 7 contain lightweight quantitative evaluation, Chapter 8 additionally features a detailed case study, and Chapter 11 is itself a case study description; Chapters 5, 6, 7, 9, 10, 12, 13 and 14 contain either no validation at all or just a brief verbal mention. (Chapters 1 and 2 stem from a 'PhD intent' paper and a survey paper, hence no validation could be expected.)

The thesis is likely to obtain a certain degree of international *impact* in the specific community focusing on conceptual modelling of XML, both from the scientific and practical point of view (the latter is corroborated by the case studies included).

Aside the mentioned problem of content consolidation, the *readability* of the text is relatively good. At verbal level, the reader is neatly guided by chapter forewords, i.e. the thesis is well integrated at the 'surface' of individual chapters. Also the quality of English as well as adherence to typographic conventions is solid, although some chapters suffer from more frequent typos than others. I also missed numbering of formulae.

As far as the associated *publication activity* is concerned, I see it as surpassing the degree normally expected for a PhD thesis in the local conditions. Most of the publications have been presented at solid international venues. There are even two articles with highly respected journals (although there the candidate was not the first author). Among the first-authorship publications, at least [54] has already been *cited* (according to Google Scholar) beyond the candidate's own group.

Detailed comments to Chapter 3 (declared as 'core contribution' in the schema integration part, which is only 'further abstracted' in Chapter 4…):

- What is lacking in particular, from the understandability viewpoint, is a summary comparison of the structure of PIM vs. PSM (e.g., association in PSM is of a single type, the taxonomic one, in contrast to PIM with different – albeit sometimes named and sometimes unnamed – associations), and PSM vs. an XML schema (e.g., PSM attributes can model either an element or an attribute in an XML schema). The reader has to reconstruct these important distinctions by herself; otherwise the resemblance might be misleading.
- It is unnecessarily cryptic to only introduce the L_a and L_e sets of string labels, without saying that the former (presumably) contains attribute names and the latter contains element names. BTW, isn't it a bit restrictive dichotomy?
- Just for formal purity: where speaking about C_1 and C_2 (not sharing attributes or the like), they should be declared as different from each other.
- Minor comment: it's unusual to introduce symbols like 'xml' or 'content' with the 'prime' sign while there are no such symbols without this sign, in the formalism. 'Prime' here just indicates pertinence to the PSM rather than PIM – this could be done by other means.
- It is not semantically ideal to say that a class is 'child OF an association'. It is child of the parent class, and maybe 'child IN (the taxonomic) association'.
- Def. 3.3 uses the *C*' symbol first in general terms and then specifically for the root class. The root class would deserve a dedicated symbol.
- Def. 3.4: interpretation is a total function... over what? Union of *C*', *A*', *R*'? Furthermore: in the example below, the value of I for (Employer,Country) is a path (i.e. element of D^R and not of *R*) rather than a single association!
- It is unclear if S^{str} is normalized to (0,1) or if it is really the absolute length of the common substring. The latter case would be awkward as the measure is linearly combined with a binary feature (type identity).
- The notion of 'pre-order' is not completely clear, at least for someone outside the XML community (is it the order of depth-first left-to-right tree traversal?).
- In 3.4.2 I would expect a proof of the polynomial complexity of the algorithm.
- The notion of 'similarity adjustment' is not very instructive. Essentially, it is a graph-based similarity, although applied in an incremental manner, thus perhaps with an 'adjusting' aspect.

- Imprecise verbalization: "$S^{adj\text{-}class}$ is the average…" without mentioning the final addition of 1. Actually, it is a distance rather than similarity measure, thus it shouldn't use the *S* symbol. (This is fixed in Chapter 4.)
- Why is Fig. 3.3 structurally different form Fig. 3.2? BTW, it is not referenced in the text.
- "…if the PSM class *BusinessSector* is not present in the PSM…" – perhaps you meant "…in the PIM"?
- 'reversed value of an adjustment' – better, 'reciprocal value'? Actually, this shape of the function is not ideal. The impact of class proximity first decreases rapidly and then more slowly – it should better be the other way around or at least linearly (since very distant classes should not be considered at all). This is probably fixed in Chapter 4, if I understood it well.
- The heading 'Evolution of PIM' is not very adequate, as the paragraph rather deals with creation of a PIM, which is originally empty.
- Diagram in Fig. 3.5 lacks any label for the X-axis.
- Multiplication by 100 in the 'precision' formulae is useless.
- "If there are more PIM classes with the same similarity to *C'*, *order(C)* is the order of the last one." This doesn't make sense: there must always be *some* order when the classes are presented to the expert, i.e. the order of *C* is given and cannot be constructed at evaluation time. For ordering of classes with equal similarity, either some other heuristic is used, or the order is random. Random ordering of equal-similarity classes should however naturally lead to measuring the 'precision' wrt. the *median* class in the equal-similarity sublist, not wrt. the last one.
- The precision formula is not intuitive in the sense that the worst possible value is 1/*n* rather than 0. (Fixed in Chapter 4.)
- For 'local precision': "When *C* is the first class there can be other PIM classes before *C*…" – no class can then be *before C*! (Fixed in Chapter 4.)
- I would say that rather than 'local precision', some other term, such as 'distinctiveness', would be more adequate.

Detailed comments to Chapter 5

- The notion of 'structural representant' is not sufficiently explained (aside not looking like proper English). Furthermore, if it means that the structure is identical in several branches of the tree, rather than treating these replications in the described way, wouldn't it be simpler to transform the tree to a general graph with single occurrence of each such structure?

Detailed comments to Chapter 9

- Does the notion of 'structural inheritance' really deserve the name 'inheritance'? Isn't it simply 'reuse'?
- Def. 9: (function 'final' – BTW, it should be explicitly stated that these functions are Boolean) "…whether this class can be inherited from, respectively." (something missing in this sentence)

Detailed comments to Chapter 13

- Surprisingly, the Related Work only covers work specific to semantic web services and not general UML-OWL mapping, for which there has been much more extensive research published. The authors' approach is not specific to web services in any way.

I conclude that the submitted thesis, in my opinion, despite the mentioned weaknesses, satisfies the common criteria for dissertation thesis as original and substantial scholarly work, **proving the ability of the candidate to autonomously undertake scientific research**.

Questions to be possibly answered during the defence:

- Chapter 3, as well as some other chapters, explicitly refer to 'web service interfaces'. However, the method doesn't seem to address any specifics of WS interfaces. Are WS interfaces good representatives of XML documents in general, or do they have any peculiarities, which should be taken into account?
- When conceptually modelling XML using a UML-style PIM, as in Fig. 3.1, you think in more 'real-world' terms than in an XML schema, which is merely the structural description of a 'data container'. However, even there can still be different degrees of 'real-worldness': while a *company* and a *country* are indeed real-world 'objects', an *address* is still a data structure (or, at most, abstract information object) that allows to connect them; it does not have a meaning without the objects it connects. Could then an even 'higher-level' ontological model fit into the whole conceptual modelling framework, which would make this distinction explicit? (This way I am also trying to promote my own recent work on ontological background models, published at OWLED'13 and K-CAP'13, which I will provide you offline ☺)
- This is a fine-grained (and perhaps outsider's) one. The function 'rcard' assigns a cardinality to a pair 'class-association'. What if you have an association with equal ends, e.g., the 'parent' association anchored to class 'Person' on both ends? Can you specify that each end has a different cardinality constraint (without recourse to explicit introduction of 'Parent' class as subclass of 'Person')?
- Another fine-grained one. Does the similarity adjustment principle, which proceeds bottom-up (class similarity is adjusted according to its children or leaf nodes) entail that PIM classes corresponding to classes higher in the PSM hierarchy are presented with better accuracy than PIM classes corresponding to classes lower in the PSM hierarchy (as for the latter there is not enough 'adjustment evidence' yet in the time of the presentation)? Is there any empirical evidence in this respect? If yes – is it then practical from the point of view of result quality and human interaction overhead?

Prague, 31 July, 2013

Doc. Ing. Vojtěch Svátek, Dr.
Department of Information and Knowledge Engineering
University of Economics, Prague