

Document Variant Relations for Identifying and Supporting Scientific Communities of Practise

Achim Mahnke¹

DFKI GmbH, Universität Bremen

achim.mahnke@dfki.de

Abstract. The variant relation approach facilitates to capture the specific differences between variants of documents or their parts. We delimit the idea of document variants from other document management issues like sharing and reuse in general and introduce the variant relation approach. Based on advanced features, this approach is offering in contrast to existing work in the area of variant documents, we present ideas on how these features can be explored for supporting scientific communities of practice.

1 Introduction

There are different scenarios for creating variants of documents. The most common approach addresses the requirements of document authors who want to adapt their documents to different groups of readers. Here, in the document creation phase, a methodology for building variants is already chosen and appropriate management functionalities are implemented in the authoring environment. This situation suggests a document model in which all variants of a document are occurrences of the same logical (or abstract) document and therefore share a common root – often, all variants of one document are in fact created by the same author.

This scenario is suitable for closed environments, e.g. institutions or publishers where fixed policies and systems for document processing are in effect. For open environments, where a huge amount of documents are created by authors mostly unrelated to each other and where different document processing techniques and systems are used, the 'single-root' approach is not appropriate. Furthermore, the lack of a common methodology for managing document variants in the past has led to the current state of documents on the web: Many documents contain the same information with slight variations, but these relationships cannot be explored because they are not explicated. This not only holds for the big scope of the world wide web but also for more 'closed' collections of documents like databases for scientific articles or eLearning repositories because variant management for documents has not been commonly used in the past due to several reasons.

Our approach of specifying document variants addresses the latter szenario, where variant relationships between existing documents are discovered and explored for document management processes. In this article, we describe how scientific communities of practise can benefit from this approach.

2 Related Work

The main question that has driven the development of methods and techniques for managing document variants has been the problem of representing content dependend information in a coherent way. Stavarakas and Gergatsoulis have presentend an approach for *multidimensional semistructured data* [SG02] in which they enhance an existing model for semistructured data (OEM) by introducing a new multidimensional node which serves as a placeholder for context-dependent facets of a piece of information. The author attaches context specifiers consisting of a set of pairs of a dimension and a respective value to the facets in order to specify, under which worlds (contexts) the facet is valid and will fill the place of its multidimensional parent node. Consistency rules and a special query language makes sure that the multidimensional structure can always be reduced to a normal graph structure which corresponds to one document which is valid for a context in question.

Stavarakas and Gergatsoulis build upon the ideas of *Intensional HTML* ([Wad00]), where authors of HTML files can annotate fragments of their web pages with context information based on dimensions like natural language etc.

Norrie et.al. ([NP03]) have developed a similar approach for the contextualization of web applications. They have enhanced their web content management system OMS with a technique for annotating the objects of a web application with context dependency information and a flexible content selection mechanism.

All these approaches try to solve the problem of context dependent content in the single root scenario described in Sect. 1, where one logical document is tailored to different contexts like varying audiences, media etc. In our work so far, we too have concentrated on this perspective and developed a repository for eLearning material which addressed the management of lecture variants as well as the semantic interrelation of documents ([KBHL⁺03,KBLL⁺04,MTea04]).

Beginning with the integration of our work into OMDoc ([Koh06,KMM07]), the relation of two documents or parts of it came into focus and the variant relation approach aims at enhancing the existing systems by taking the semantics into account, a variant relation between two documents can offer.

3 Variant Relations

3.1 Sharing, Reuse and Variants

One of the main issues in the area of document management has been the copy and paste method, many authors are (still) using in order to reuse existing material for creating new documents. This kind of reuse bears the problem that the original object and the copied one have no connection any more and that changes on the original, e.g. error corrections, will not be propagated to the copies.

Document markup languages and management systems therefore facilitate the sharing of objects through inclusion by reference. The original object is not copied but referenced in the new document; updates on the original are automatically reflected in the new documents, e.g. object 13 in document D in Figure 1(a) is reusing object 5 of document C.

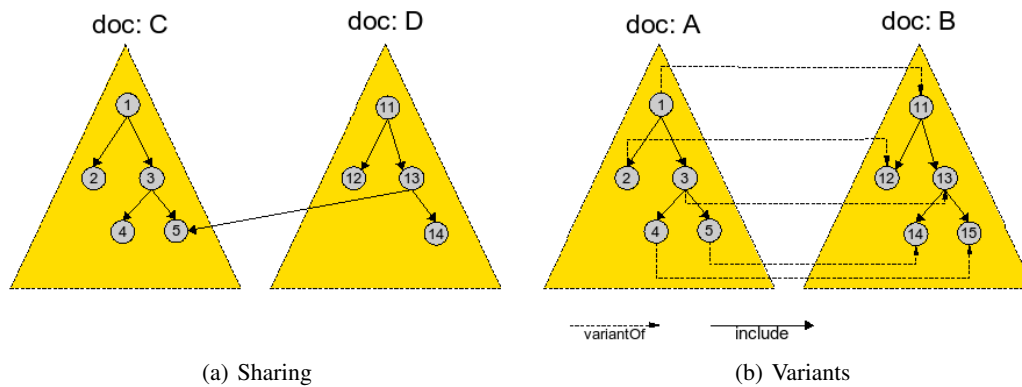


Fig. 1. Examples for Variant Relations

Documents sharing a considerable amount of objects might be considered as variants of each other, e.g. if somebody is preparing a short and a long version of an article. In the long version parts of the short one are reused and new content is added. This situation is sometimes described as if these versions are variants. In our point of view, however, this is a reuse scenario, because existing parts are augmented by new ones, but there are rarely objects which exist in two alternative (re)presentations.

In contrast to that, we speak of variants as being different representations of the same abstract knowledge. Variant objects can *replace* but should not complement each other. Figure 1(b) depicts this situation: each object of document A and B are unique but are variants of each other.

3.2 Semantic of Variant Relations

Variant relations provide the means for expressing the fact that two documents – or parts (objects) of documents – basically conveys the same knowledge in different ways. The important underlying assumption is that, when someone is stating a variant relationship between two document objects, he or she is correctly judging that both objects capture the same knowledge and therefore are – in a quite general sense – equal. The second important information, a variant relationship carries is *in what respect* the objects differ in presenting or representing this knowledge: Natural language, wording, format, character coding, conceptualization etc.

A variant relation type is a named binary relation with no restriction on the type of objects being related – besides the requirement that they uniquely identify documents (or parts thereof).

The name of a variant relation is reflecting the aspect, in which the related objects differ from each other, e.g. in the natural language, their text parts has been written in. If we define a relation with the name **naturalLanguage**, the fact that there is a difference in this aspect can be expressed, but there are no means to denote the concrete

languages involved. Therefore, each end of a variant relation can be accompanied with an attribute which specifies the characteristic of the respective object. This could be realized by defining an attribute for the dimension of natural language with values like 'en' for english etc. (see Figure 2(b)).

If we consider the management of change of documents, variant relations carry a basic meaning of dependency: as they express semantic equality on the content aspect of the two objects related to each other, a change of the content semantics on one object breaks the relationship if the other object is not altered accordingly. But we can also express dependency on the representation aspect: if a document is translated from, eg. english to german, then the german object is dependant not only on the semantic content but also on the concrete wording of the original english variant. In Figure 2(b) we denote this by one-directional arrows.

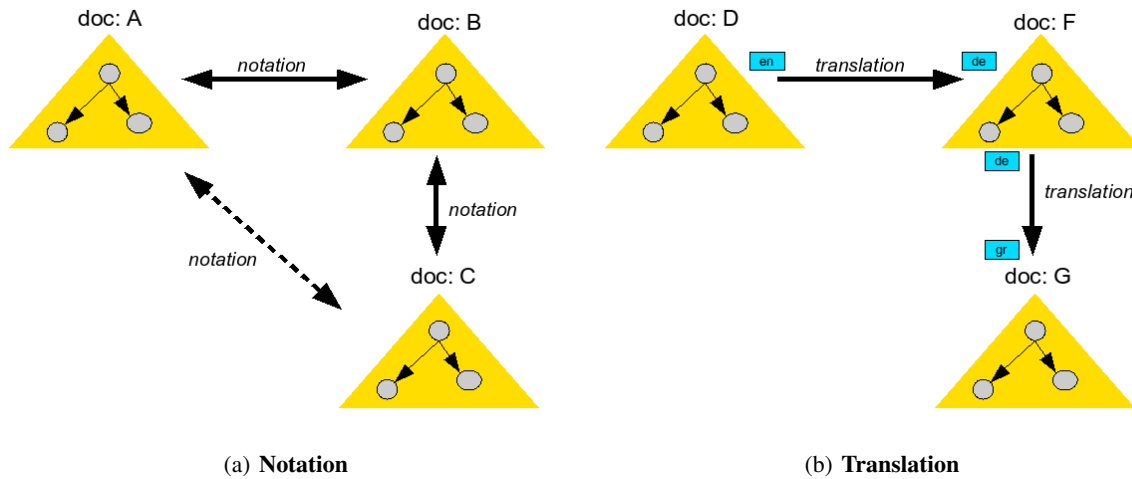


Fig. 2. Sharing vs. Variants

3.3 Example

We assume that a repository for lecture material for mathematics encompasses documents from many different authors and that several introductory courses on mathematics for different faculties (e.g. computer scientists, electrical engineers) has been inserted into the repository over time. It is likely that the same topic is addressed in several courses but each time presented slightly different.

In mathematics, a *complex number* is a *number* which can be formally defined as an *ordered pair of real numbers* (a, b) , often written: $a + bi$ where $i^2 = -1$.

...

In mathematics, particularly in combinatorics, a binomial coefficient is a coefficient of any of the terms in the expansion of the binomial power $(1+x)^n$. It is given by the following formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

In mathematics, a *complex number* is a *number* which can be formally defined as an *ordered pair of real numbers* (a, b) , often written: $a + bj$ where $j^2 = -1$.

...

In mathematics, particularly in combinatorics, a binomial coefficient is a coefficient of any of the terms in the expansion of the binomial power $(1+x)^n$. It is given by the following formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

In mathematics, a *complex number* is a *number* which can be formally defined as an *ordered pair of real numbers* (a, b) , often written: $a + bi$ where $i^2 = -1$.

...

In mathematics, particularly in combinatorics, a binomial coefficient is a coefficient of any of the terms in the expansion of the binomial power $(1+x)^n$. It is given by the following formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Fig. 3. Example lectures A, B and C

Figure 3 shows an example where the concepts of complex numbers and the binomial coefficient are explained ([Mül08]). Because electrical engineers are used to denote the imaginary unit of complex numbers with j instead of i as in mathematics, the learning object B differs from A in the *notation* of this mathematical construct. Object C differs from B in using a different notation convention for the binomial coefficient.

We therefore define a variant relation **notation** which conveys the information that the related document objects are differing with respect to the notation of some or all mathematical constructs encompassed by the document objects. Given that the above relationships have been identified and marked up by a user, we have the following situation: $(A, B) \in \mathbf{notation} \wedge (B, C) \in \mathbf{notation}$.

Essentially this means that all three document objects contains the same information although presenting it slightly different in respect of mathematical notations. We can assume that all of them are explaining the mentioned mathematical concepts and that all notation differences are equitable. We therefore define **notation** as symmetric and transitive, so that we can derive additional facts about variant relations between A, B and C: $(B, A) \in \mathbf{notation} \wedge (C, B) \in \mathbf{notation} \wedge (A, C) \in \mathbf{notation} \wedge (C, A) \in \mathbf{notation}$. Figure 2(a) visualizes that.

4 Variant Relations for Communities of Practice

For scientific communities of practise, we assume that practices are inscribed into artifacts like documents, emails etc. (see [MK08]) We propose that variant relations between documents can be used to describe or even identify communities of practice. Both main statements, a variant relation makes about the interrelated documents can be explored: the equivalence of the semantic of their content and the difference in some aspect of their presentation or representation.

4.1 Identifying CoPs Based on Different Practices

In Figure 4 we broaden the view on the situation depicted in Figure 2(a) where the same learning material is presented with different notation in documents A,B and C. Repository users are aware of these variants and have the choice which one they prefer and bookmark for further reference.

Although, all readers of documents A, B or C are interested in the same topics, because the semantic content is equivalent, they differ in the preference for a certain notation practice. The notation variant relations reflect this and give rise to the assumption, that the readers of these documents can be divided into three CoPs corresponding to these variation in notation practice.

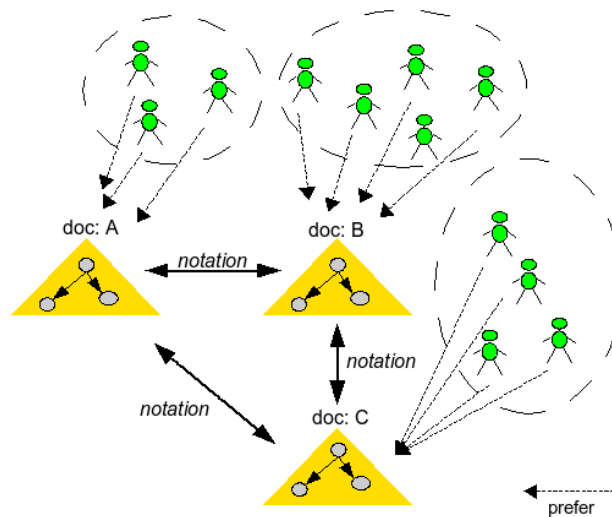


Fig. 4. Different Practices

It would be beneficial to know something about the nature of the notation styles and in which context they are used, but the sole information that the documents (only) differ with respect to notation is an indicator that readers could be grouped along this line.

4.2 Identifying CoPs Based on Equivalent Content

We assume that there are CoPs where certain documents play an important role in defining the subjects or topics, the CoP is interested in. For example: if there is a CoP *CopAReaders* for which the document A is important and someone is discovering a variant relationship to document B which plays a role in CoP *CopBReaders* this indicates that the members of both CoPs are interested in the same topics because the second important information incorporated in variant relations is, that the variant documents are basically conveying the same information. New personal contacts could be initiated and possibilities for collaboration could be found on this insight. Perhaps, the creation of a new CoP combining *CopAReaders* and *CopBReaders* is beneficial.

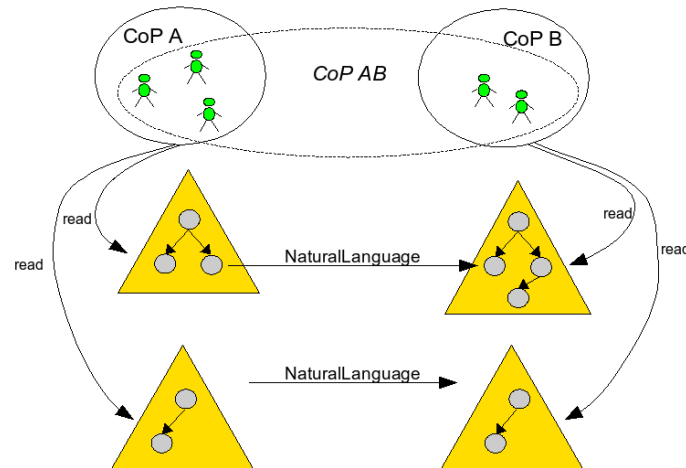


Fig. 5. Equivalent Content

5 Conclusion and Outlook

We have presented work in progress on the *variant relation* approach for managing document variants which builds upon the idea of capturing the kind of relation between to variant objects. This approach is more suitable for large amounts of documents created independently than existing approaches in this area.

We illustrated ideas on how variant relations can be explored for supporting communities of practice. Equivalences and differences expressed through variant relations can hint the CoP building process by revealing common interest for the same topics or significant differences in practice incribed into documents.

The next step would be to test these ideas within a suitable tool for communities of practice, e.g. *panta rhei* [pan08] by implementing facilities to create and explore variant relations. For severe testing of the approach, a large corpus of documents containing variants is needed – for this, the language variants of Wikipedia [wik08] is under investigation at the moment.

Acknowledgment

We would like to thank the members of the KWARC group at Jacobs University and the DFKI Bremen group for the fruitful discussions and continuous feedback on our work.

References

- KBHL⁺03. B. Krieg-Brückner, D. Hutter, A. Lindow, C. Lüth, A. Mahnke, E. Melis, P. Meier, A. Poetzsch-Heffter, M. Roggenbach, G. Russell, J.-G. Smaus, and M. Wirsing. Multimedia instruction in safe and secure systems. In M. Wirsing, D. Pattinson, and R. Hennicker, editors, *Recent Trends in Algebraic Development Techniques*, volume 2755 of *Lecture Notes in Computer Science*, pages 82–117. Springer-Verlag, D-69121 Heidelberg, Germany; <http://www.springer.de>, 2003.

- KBLL⁺04. Bernd Krieg-Brückner, Arne Lindow, Christoph Lüth, Achim Mahnke, and George Russell. Semantic interrelation of documents via an ontology. In G. Engels and S. Seehusen, editors, *DeLFI 2004*, volume P-52 of *LNI*, pages 271–282. Springer-Verlag, 2004.
- KMM07. Michael Kohlhase, Achim Mahnke, and Christine Müller. Managing variants in document content and narrative structures. In Alexander Hinneburg, editor, *Wissens- und Erfahrungsmanagement LWA (Lernen, Wissensentdeckung und Adaptivität) conference proceedings*, pages 324–229, 2007.
- Koh06. Michael Kohlhase. OMDOC – *An open markup format for mathematical documents [Version 1.2]*. Number 4180 in *LNAI*. Springer Verlag, 2006.
- MK08. Christine Müller and Michael Kohlhase. Copit - towards a community of practice toolkit based on semantically marked up artifacts. 2008. in submission.
- MTea04. Robert Meersman, Zahir Tari, and Angelo Corsaro et al., editors. *On the Move to Meaningful Internet Systems 2004: OTM 2004 Workshops*, number 3292 in *LNCS*. Springer Verlag, 2004.
- Mül08. Christine Müller. A Survey on Mathematical Notations, 2008. <http://kwarc.info/publications/papers/kwlnotationSurvey.pdf>.
- NP03. Moira C. Norrie and Alexios Palinginis. Versions for context dependent information services. In Robert Meersman, Zahir Tari, and Douglas C. Schmidt, editors, *CoopIS/DOA/ODBASE*, volume 2888 of *Lecture Notes in Computer Science*, pages 503–515. Springer, 2003.
- pan08. The panta rhei Project. <http://kwarc.info/projects/panta-rhei/>, seen May 2008.
- SG02. Yannis Stavrakas and Manolis Gergatsoulis. Multidimensional semistructured data: Representing context-dependent information on the web. In Anne Banks Pidduck, John Mylopoulos, Carson C. Woo, and M. Tamer Özsu, editors, *CAiSE*, volume 2348 of *Lecture Notes in Computer Science*, pages 183–199. Springer, 2002.
- Wad00. William W. Wadge. Intensional markup language. In Peter G. Kropf, Gilbert Babin, John Plaice, and Herwig Unger, editors, *DCW*, volume 1830 of *Lecture Notes in Computer Science*, pages 82–89. Springer, 2000.
- wik08. Wikipedia project. <http://www.wikipedia.org/>, seen June 2008.