

Michael Sperberg-McQueen

Containment and dominance in Goddag structures

Goddag structures (generalized ordered-descendant directed acyclic graphs) are an alternative to trees as a data structure for thinking about and manipulating textual material; informally, they may be thought of as sets of tangled trees with shared sub-structures. This paper describes some possible revisions to the notion of dominance in Goddag structures.

The original description of Goddags [3] specifies that whenever one element A contains all the character data contained by another element D, then A dominates D either directly or indirectly (unless A and D contain exactly the same character data, in which case one must dominate the other). This leads to deterministic results for the structure of a Goddag over a given body of material, but the structures thus generated sometimes feel intuitively awkward, often because the parent-child relations seem inconsistent and unhelpful. When elements of two different types X and Y share content, sometimes the X is the ancestor of the Y, sometimes vice versa, and sometimes neither is ancestor of the other.

Related phenomena can be observed in concurrent SGML [2] and XML [5], where the constraints on the multi-tree structures forbid structure-sharing in some cases where it seems natural. The problem can also be illustrated using the technique of fragmenting virtual

elements in order to fit them into an XML structure. In the case of fragmentation, a simple mechanical association of containment and dominance results (for example) in pages sometimes containing paragraphs, paragraphs sometimes containing pages, and chapters containing a bewildering mixture of pages and paragraphs, some fragmentary and some whole, as siblings of each other. Colored XML [1] provides further examples, in which the relation between dominance and containment is seen from a very different perspective.

A tentative analysis of intuitively satisfactory and intuitively unsatisfactory cases leads to the conjecture that two distinct relations among nodes, 'containment' and 'dominance', should be postulated. Containment is a purely empirical superset-subset relation between the contents of elements, while dominance entails not only containment but also some further relationship between the two nodes, whose exact nature remains elusive.

Structures which are intuitively more satisfactory and easier to work with may be made possible by distinguishing containment and dominance, and by modifying the definition of Goddag structures to make the parent-child relation reflect dominance not containment. The introduction of 'colored Goddags' [4] may also help address this problem area.

[1] Jagadish, H. V., Laks V. S. Lakshmanan, Monica Scannapieco, Divesh Srivastava, and Nuwee Wiwatwattana. 2004. "Colorful XML: One hierarchy isn't enough". Proceedings of the 2004 ACM SIGMOD International conference on management of data, Paris, sponsored by the Association for Computing Machinery Special Interest Group on Management of Data. New York: ACM Press.

[2] Sperberg-McQueen, C. M., and Claus Huitfeldt. 1999. "Concurrent document hierarchies in MECS and SGML". *Literary & Linguistic Computing* 14.1: 29-42.

[3] Sperberg-McQueen, C. M., and Claus Huitfeldt. 2000. "GODDAG: A Data Structure for Overlapping Hierarchies". In DDEP-PODDP 2000, ed. P. King and E.V. Munson, *Lecture Notes in Computer Science 2023* (Berlin: Springer, 2004), pp. 139-160. Available on the Web at <http://www.springerlink.com/index/98J1VBU5NBY73UL3> and <http://www.w3.org/People/cmsmcq/2000/poddp2000.html>.

[4] Sperberg-McQueen, C. M. 2007. "Representation of overlapping structures". In *Proceedings of Extreme Markup Languages 2007*. Available on the Web at <http://www.idealliance.org/papers/extreme/proceedings/html/2007/SperbergMcQueen01/EML2007SperbergMcQueen01.html>

[5] Schonefeld, Oliver. 2007. "XCONCUR and XCONCUR-CL: A constraint-based approach for the validation of concurrent markup". In *Datenstrukturen für linguistische Ressourcen und ihre Anwendungen / Data structures for linguistic resources and applications: Proceedings of the Biennial GLDV Conference 2007*, ed. Georg Rehm, Andreas Witt, Lothar Lemnitzer. Tübingen: Gunter Narr Verlag. Pp. 347-356.