1 Introduction

The research which led to this paper was motivated by the desire to have a very expressive language for talking about XML trees with a very simple syntax. For expressivity, monadic second order logic (MSO) seems a natural upper limit, so that was our original goal. For the syntax, we choose (the core of) XPath 1.0 as the guideline. There are several reasons for desiring such a language: the most abstract being that it can serve as a unifying framework in which different XML related technologies can easily be embedded. More concretely, we wanted a language in which we can validate documents, select sets of nodes and define paths in a document. Minimally both DTD’s and navigational (Core) XPath should be expressible in the language. Another abstract reason was to discover what expressive power is “really” needed for XML related languages. For that reason we only wanted to accept a language with a natural characterization in terms of either logic or automata. There are strong theoretical reasons, especially coming from the theory of tree automata and regular tree languages, for choosing MSO as the “golden standard”. On the other hand, an intuitive syntax driven approach might lead to a different standard with its own merits.

The main contribution of this paper is in the definition of such a language and a characterization of it in terms of tree walk automata with a simple pebbling regime. This language is the expansion of the caterpillar expressions of Brüggenmann-Klein and Woods with an operator $(\cdot)^{\text{loop}}$. The meaning is simple: a path from node $n$ to node $n'$ is a $R^{\text{loop}}$ path if and only if the path is an $R$ path and $n = n'$. Our most surprising result is that –on finite ordered unbounded trees– every first order definable path is definable in this language.