Neuroinform (2012) 10:1–3
DOI 10.1007/s12021-011-9136-8

EDITORIAL

The Promise and Shortcomings of XML as an Interchange Format for Computational Models of Biology

Ivan Raikov · Erik De Schutter

Published online: 7 December 2011
© Springer Science+Business Media, LLC 2011

A number of XML-based (Extensible Markup Language[1]) model description languages have been developed for representing models of dynamic biological processes independently of simulator platforms. NineML[2] is an effort that aims to provide an unambiguous description of large-scale neuronal network models. NeuroML[3] is a language for defining and exchanging descriptions of neuronal cell and network models. The Systems Biology Markup Language[4] (SBML) can represent cell signaling pathways, regulatory networks, and other kinds of biochemical network models. CellML[5] aims for a wider scope of model description and is not specific to any one field of biology.

XML, a data exchange language closely related to HTML, is a widely accepted standard for describing structured textual data. Its often advertised advantage is that XML documents with different structure can be read by the same generic reusable parser. However, XML by itself does not enable information interchange. Human readers of XML may be able to guess the meaning of a statement such as

<neuron> g = 0.1 </neuron>

but to a computer program, <neuron> and <axon> and <giraffe> are all equally meaningless. What a program should do with XML data is undefined by the XML standard.

Frequently, in practice, the semantics (or meaning) of XML-based model description languages are described in specification documents written in natural human language. Software programmers must read the specification documents and convert the requirements into programs (in Python, Java, etc.) for reading and writing model descriptions. But as new modeling approaches emerge, and new simulation code is written, the semantics of model description languages must be reimplemented. A specification written in a human language often hides ambiguities, and as the complexity of the languages and the number of supported software platforms increases, it is difficult or impossible to ensure that every language construct is implemented consistently in every software package.

The XML community has developed several schema languages that can specify rules for structuring and validating XML documents, but they offer weak support for data types, procedures, or complex dependencies between elements. Therefore, the expressive power of XML is greatly affected by its interpretation.

The use of XML for the syntactic structure of a modeling language does indeed eliminate the need for specialized

[1] World Wide Web Consortium. Extensible Markup Language (XML). http://www.w3.org/XML/
[2] INCF Multiscale Modeling Task Force (2011). NineML: declarative, mathematically-explicit descriptions of spiking neuronal networks (2011). Neuroinformatics 2011.
[3] Gleeson P, Crook S, Cannon R, Hines M, Billings G, et al. (2010) NeuroML: A language for describing data driven models of neurons and networks with a high degree of biological detail. PLoS Comput Biol 6(6): e1000815.
[4] Hucka M, Bergmann F, Hoops S, Keating S, Sahle S, et al. (2010). Systems biology markup language (SBML) Level 3 Core. http://sbml.org/Documents/Specifications/SBML Level 3
[5] Cuellar A, Nielsen P, Hallstead M, Bullivant D, Nickerson D (2002). CellML 1.1 specification. http://www.cellml.org/specifications/cellml_1.1

I. Raikov (✉) · E. De Schutter
Theoretical Neurobiology, University of Antwerp,
2610 Antwerpen, Belgium
e-mail: raikov@oist.jp

I. Raikov · E. De Schutter
Computational Neuroscience Unit,
Okinawa Institute of Science and Technology,
Okinawa 904-0495, Japan

parsers, but does not eliminate the need for a system of semantic rules that specify how a model must be validated and processed. Every XML-based modeling language must be accompanied by a set of semantic rules to give meaning to models written in this language. A common set of formal semantic rules ensures that a model will be read and processed meaningfully and correctly by different software applications because all implementations mean the same thing. 'Formal' refers to particular kinds of mathematically-based techniques for the specification of computer languages, as opposed to informal human language specifications.

The meaning of computer languages has been defined formally in a number of ways. In programming language research, the classical approach to semantics has been to use a subset of first-order logic to define a valuation function (or a relation) that defines the meaning of program by establishing formal relationships between its inputs and outputs. Valuation functions ("denotational semantics") were pioneered by mathematicians Dana Scott and Christopher Strachey[6,7]). Meaning relations ("axiomatic semantics") were pioneered by computer scientist Tony Hoare[8]. Scott and Hoare each received the Turing award for their work. They have shaped virtually all research in compiler technologies and programming languages for the past four decades.

A lightweight variant of the approach of Scott and Strachey has been used to define a model description language for conductance-based models of neuronal ionic currents. The language consists of distinct semantic layers to represent the various domain-specific biological modeling concepts and the underlying mathematical formalisms. The relationships between the layers are formally defined by semantic transformation functions, which if faithfully implemented ensure consistent interpretation of the language. Adding new features to the language (e.g. experimental protocols) is simply a matter of defining additional layers and semantic transformation functions[9].

An alternative approach taken by the editors of the SBML specification is to supplant the specification text with Unified Modeling Language[10] (UML) diagrams. UML is a widely-used software industry standard for specifying and visualizing the structure of object-oriented software systems. However, the symbols in the visual notation used by UML are informally defined, and the UML specification itself has been cited as being ambiguous and inconsistent.

The development of the Semantic Web[11] has led to solutions to build sharable ontologies, or common sets of interrelated definitions. These definitions, together with semantic annotations, serve as a basis for the unambiguous machine interpretation of documents and model descriptions. The annotation of models with semantic metadata is considered a promising approach to reach semantic interoperability in the domain of computational biology. Efforts such as the MIRIAM guidelines[12] for the annotation of computational models use descriptive logic (another subset of first-order logic) for formal knowledge representation. This allows a model to be integrated with biomedical ontologies and connections to be established between models and data sets through automated reasoning software.

Applications such as SBML Harvester[13] utilize the structure of SBML models and MIRIAM annotations to create complex ontology-based representation of SBML models. Such approaches make SBML's semantics explicit and enable verification and integration with different modeling frameworks.

In the field of neuroimaging, the web-based Query Manager[14] allows the semantic integration of experimental datasets and neuroanatomical ontologies. In neuroinformatics, the NeuroLex[15] project has as an aim the establishment of a dynamic neuroscience lexicon, and the International Neuroinformatics Coordinating Facility is developing a Neuron Registry[16] based on a proposal for machine-readable ontological descriptions of neuron types[17]. In principle, such efforts could be used to annotate computational neuroscience

[6] Scott D, Strachey C (1971) Toward a mathematical semantics for computer languages. In: Proceedings of the Symposium on Computers and Automata. volume XXI, pp. 19–46.

[7] Schmidt D (2000) Induction, domains, calculi: Strachey's contributions to programming-language engineering. Higher-Order and Symbolic Computation 13: 89–101.

[8] Hoare CAR (1969) An axiomatic basis for computer programming. Communications of the ACM Vol. 12 (10).

[9] Raikov I, De Schutter E (2011). The layer-oriented approach to declarative languages for biological modeling. (submitted).

[10] Object Management Group (2005). Unified Modeling Language Version 2.0. http://www.omg.org/spec/UML/2.0/

[11] Berners-Lee T, Hendler J, Lassila, O (2001). The Semantic Web. Scientific American Magazine, May 2001.

[12] Le Novère N, Finney A, Hucka M, et al. (2005). Minimum information requested in the annotation of biochemical models (MIRIAM). Nat Biotechnol 23 (12): 1509–15.

[13] Hoehndorf R, Dumontier M, Gennary JH, Wimalaratne S, et al (2001). Integrating systems biology models and biomedical ontologies. BMC Systems Biology, 5:124.

[14] Turner JA, Mejino JLV, Brinkley JF, Detwiler LT, et al (2010). Application of neuroanatomical ontologies for neuroimaging data annotation. Frontiers in Neuroinformatics 4 (0).

[15] Bug, WJ, Ascoli, GA, Grethe, JS, et al. (2008). The NIFSTD and BIRNLex vocabularies: building comprehensive ontologies for neuroscience. Neuroinformatics, 6(3), 175–194.

[16] INCF Neuron Registry Task Force (NRTF) http://www.incf.org/core/programs/pons/projects/nrtf

[17] Hamilton DJ, Bergamino M, DeFelipe J, Le Novère N, et al (2011). Machine-Readable Description of Neuron Types and Properties. Neuroinformatics 2011.

models with semantic metadata and integrate them with ontological databases.

The increasing use of formal methods in computational modeling of biology parallels the development of computer science and engineering, where the use of mathematical analysis techniques has contributed to the refinement and correctness of system design. The mere use of XML as a format for interoperable model description languages is not enough. An unambiguous semantic specification is a prerequisite for interoperable software tools based on a common model description language.