Institution: The University of Edinburgh

Unit of Assessment: B11 — Computer Science and Informatics

Title of case study: Shaping the XML technologies used to manage the world's data

1. Summary of the impact

The success of the eXtensible Markup Language (XML) has been due in large part to the technologies built around it for constraining, querying, styling and otherwise processing XML documents. Research carried out at Edinburgh has been instrumental in the creation and/or design of many of these core XML technologies, including XSLT, XML Schema, XInclude, XQuery and XProc. Edinburgh staff played key roles in bringing these technologies into widespread use in both the private and public sectors through participation in standards development work.

2. Underpinning research

[Unless otherwise explicitly noted, all the work discussed in this and the following section was carried out at the University of Edinburgh]

University of Edinburgh staff details
(Unless otherwise noted, Edinburgh employment began before 1993 and continues to date)

- C. Brew, Research Fellow, left 2000
- P. Buneman, Professor, since 2002
- J. Carletta, Senior Research Fellow
- W. Fan, Professor, since 2004
- C. Grover, Senior Research Fellow
- J. Cheney, Lecturer, since 2004
- L. Libkin, Professor, since 2006
- D. McKelvie, Research Fellow, left 2001
- A. Mikheev, Research Fellow, left 2002
- M. Moens, Senior Research Fellow, left 2004
- H. S. Thompson, Professor
- R. Tobin, Research Fellow
- P. Wadler, Professor, since 2003

2.1. Research overview

Research into the use of the Standard Generalised Markup Language (SGML) for more than just the encoding of language data grew into a major component of the work of the Language Technology Group of the Human Computer Research Centre (HCRC) at Edinburgh by 1993. Led by Thompson and Moens**,** in 1994 the group developed and in 1995 distributed a software toolkit (LT-NSL). This enabled the development of efficient modular pipelines of simple SGML-to-SGML processing steps for the implementation of complex natural language processing tasks. In 1997, this led to Thompson's participation in the standards group at the World Wide Web Consortium (henceforth W3C), which designed XML itself, and proposed languages such as XSL (see [1]). Thereafter the research focus shifted to XML, with grant support from Sun Microsystems, Microsoft and EPSRC. The use of XML pipelines for language processing was rapidly adopted in the NLP community following the release in 1998 of the second generation toolkit (LT-XML), largely the work of Tobin and Brew**,** and the success at the 7th Message Understanding Conference in 1998 of a Named Entity Recognition system built on top of it by Mikheev**,** Moens and Grover (see [6]).

A related research effort which began in 1996, involving Thompson, Tobin and McKelvie**,** focussed on developing a new architecture for multi-level annotations for language data, known as 'stand-off markup' (see [2]). This, together with the toolkit work, supported an extensive period of work on the use of first SGML and then XML to structure and publish large-scale multi-language research corpora. EU grants funded this effort from 1997 onwards. This strand of work continues to the present day, funded by EPSRC and the EU, under the leadership of Grover and Carletta. This work is primarily in the area of the interaction between markup architecture and workbench design for working with language resources, covering both written and spoken language and, in the latter case, dual- and multi-party interaction as well as single-speaker data.

Thompson, with support from Microsoft, initiated another new thread of work in 1997 aimed at providing a way to define the structure of XML documents *using* a type of XML document known as a schema (see [3]). Along with several other parallel efforts, this stimulated the creation of W3C XML Schema Working Group in 1998, where Thompson co-edited several Schema standards. Thompson and Tobin fed additional research on a novel and efficient approach to recognising languages constrained by regular expressions including occurrence indicators (exponents) into the on-going development of those standards (see [4]).

In the early 2000s, the arrival of Buneman, Cheney, Fan, Libkin and Wadler strengthened and broadened our XML research activities. Their work on the formal properties of XML, XML schemas and XML querying, informed by database antecedents and often targeted at the integration of XML and relational data, opened up a whole new field of research (see e.g. [5]), with support from EPSRC, Google and the EU. In conjunction with Thompson's work on XML transformations, this work was instrumental in further strands of standardisation through membership in the W3C's XSL (Thompson), XQuery (Wadler) and Provenance (Cheney) Working Groups.

Starting in 2000, Thompson and Tobin returned to the earlier work on XML pipelines and helped develop a new semantics for XML in terms of the 'information set'. They built on this with the idea of pipelines to initiate an understanding of XML processing as information flow, and the control of XML processing as something expressible in XML itself (see [6]). They carried this idea into practice both through a start-up company (Markup Technology, 2001) and through their membership of the XML and XML Processing Model Working Groups at the W3C, where they edited the resulting standards.

3. References to the research

1. S. Adler, H.S. Thompson, *et al.* (1997) *A Proposal for XSL*, *http://www.w3.org/TR/NOTE-XSL*
2. H.S. Thompson and D. McKelvie (1997) *Hyperlink semantics for standoff markup of read-only documents*. In SGML Europe '97, P. Gennusa, ed., Graphical Communications Association, Barcelona. *http://bit.ly/194PWJi*
3. C. Frankston and H.S. Thompson (1998) *XML-Data Reduced*. Technical Report, Microsoft, Redmond, WA. *http://bit.ly/194PPxv*
4. H.S. Thompson and R. Tobin (2003) *Using Finite State Automata to Implement W3C XML Schema Content Model Validation and Restriction Checking*. In XML Europe 2003, E. Dumbill, ed., IDEAlliance, London. *http://bit.ly/15pspPd*
5. M. Arenas and L. Libkin (2008) *XML data exchange: Consistency and query answering*, Journal of the ACM, 55(2), article no.7, May 2008. DOI *10.1145/1346330.1346332*
6. A. Mikheev, C. Grover and M. Moens (1998), *Description of the LTG system used for MUC-7*, Proceedings of 7th Message Understanding Conference (MUC-7), Fairfax, VA. *http://acl.ldc.upenn.edu/muc7/M98-0021.pdf*

References [1], [5] and [6] are the references which are most indicative of the quality of the underpinning research. These are peer-reviewed works in the most significant relevant venues.

4. Details of the impact

There is a common route to impact across the four specific areas of XML technology research reported above:

- one or more Edinburgh staff help to launch a standards effort;
- they join the group responsible for the new standard;
- they contribute a theoretically well-grounded perspective to the work of the group along with specific details from Edinburgh research; and
- they take on some of the work of writing the standards themselves.

Those standards, in turn, drive the development of both open-source (in several cases from Edinburgh) and commercial implementations that underpin wide adoption of the now-standardised technology. In all cases the standardisation work itself stretches from the time of the research described above through into the impact timeframe beginning on 2008-01-01. The most recent editions of the relevant standards are listed in the Sources section below.

It should be noted that although their roles are officially listed as 'editors' of these standards, in all cases the Edinburgh staff identified as such played a major role in design, development and detailed authoring of the standards they edited, for review and ratification by the Working Group concerned.

4.1. XML Pipelines

The Edinburgh research (see [6], which is of necessity only an indicative sample of our work in this area) established both

i. the theoretical framework, that is, the re-interpretation of a wide range of XML technologies as best understood as operations on documents not as sequences of characters, but rather as structured containers of information, and
ii. the practical evidence that this could form a sound basis for implementation.

Thompson and Tobin's membership in the W3C's XML Working Group was the initial conduit for this work, leading to Tobin's co-editing of the XML Information Set standard.

This work, as well as its exploitation via the creation and distribution of the LT-NSL and LT-XML toolsets, led to Thompson and Tobin taking a lead role in getting the W3C to launch an XML pipeline standardisation effort in the form of the XML Processing Model Working Group.

The resulting XML Processing Model standard (XProc, published May 2010), of which Thompson was a co-editor, incorporated key results from Edinburgh, including the standard's basic dataflow model and the way its stated semantics are carefully insulated from implementation details (e.g. threading, sequencing).

As of July 2013, the web page *http://xproc.org/implementations/* lists four available current XProc implementations (*Calabash*, *Calumet*, *QuiXProc*, and *Tubular*). Calumet is also distributed as the *EMC Documentum XProc Service* (*https://community.emc.com/docs/DOC-10477*). QuiXProc is available as a commercial service at *http://www.quixproc.com/quix/homeQ*. Calabash is incorporated in one of the major XML-orientated IDEs, *oXygen* (*http://www.oxygenxml.com*). oXygen has added explicit editing support for authoring and debugging XProc pipelines.

4.2. Corpora

The use of first SGML and then XML for encoding language resources, both the raw data and analyses and annotations thereof, raises many research questions at the boundaries of linguistics and computation. Edinburgh's involvement in the creation, publication and distribution of such resources, with support from both EPSRC and the EU, meant that our innovative approach to managing and recording complex multi-layered annotations became widely adopted. Alongside the corpora themselves, two separate standardisation efforts contributed to this impact:

i. the Corpus Encoding Standard (CES, and its XML version, XCES), developed via several EU projects, depends on the technique of remote or stand-off markup (see [2]), and
ii. the W3C XInclude standard, produced by the XML Working Group with substantial input from Tobin and Thompson, incorporates features to support precisely this kind of usage, based on Edinburgh's corpus development experience.

4.3. XSLT, XQuery and Databases

Thompson was responsible for introducing the idea of iconic templates to the style language for XML, now known as XSLT, which fed directly into the creation of the W3C XSL Working Group, where Tobin joined him. Building on XSLT and his XML work (some of which predated his arrival in Edinburgh), Wadler helped edit one of the families of specifications of the query-language successor to XSLT as a member of the W3C XML Query Working Group. Cheney's work fed directly into the creation of the W3C Provenance Working group, and he is co-editor of the resulting standard.

XSLT and XQuery have been the most successful of the second generation XML technologies:

- Google reports 2 million XSLT stylesheets visible on the web
- Much of eBay's websites, including auction details, is built using XML and XSLT

- The BBC deployed XSLT and XQuery extensively in their coverage of the 2012 Summer of Sport [*http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html*]
- XQuery is also the basis for a very successful XML-based company, MarkLogic [*http://www.marklogic.com*] , with annual revenue of over $50 million and customers including Warner Bros, Lockheed Martin, Boeing, J P Morgan Chase, and United Airlines.

4.4. XML Schema

In the late 1990s, Thompson (see [3]) and others explored various routes to bring the definition of XML document structure into the emerging XML-Infoset-based consensus (see XML Pipelines, above) by defining XML languages for use in defining XML languages, known as *schema* languages. As a direct result of these efforts, the W3C formed the XML Schema Working Group, and invited Thompson to edit the resulting standards.

Thompson and Tobin carried out implementation experiments throughout the development of the standards. A number of areas of the XML Schema design reflected these implementation experiments, including:

i. the provision of element equivalence classes; and
ii. the use of several varieties of inheritance in support of the object-oriented approach to schema definition that was adopted by the group.

Their theoretical work (see [4]) was crucial in providing a sound basis for implementations of schema validation. Both open-source and commercial schema validation software incorporate this theoretical work. The XML Schema technology itself is in widespread use throughout many sectors of government and industry.

- Google reports 1.7 million XML Schema documents visible on the Web.
- The UK government's *legislation.gov.uk* site uses XML Schema in the publication on the Web of all UK legislation since 1988.
- The Inland Revenue provide a wide range of online web services. The services to accept PAYE information from employers and tax returns from individuals are implemented in XML and validated using XML Schema (as well as other validation technologies).

5. Sources to corroborate the impact

A. Director of the W3C, for corroboration of contribution to XML development
B. Interaction Domain Leader at the W3C, for corroboration of XML schema influence
C. Deputy Director of the W3C, for corroboration of contribution to XML implementations
D. W3C Standards with Edinburgh authors and significantly influenced by Edinburgh work
   D.i *Constraints of the Provenance Data Model*, J. Cheney, P. Missier and L. Moreau, Sept 2012, *http://www.w3.org/TR/prov-constraints*
   D.ii *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures*, N. Mendelsohn, S. Gao, C. M. Sperberg-McQueen, D. Beech, M. Maloney and H. Thompson, Apr 2012, *http://www.w3.org/TR/xmlschema11-1*
   D.iii *W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes*, S. Gao, A. Malhotra, C. M. Sperberg-McQueen, D. Peterson, H. Thompson and P.V. Biron, Apr 2012, *http://www.w3.org/TR/xmlschema11-2*
   D.iv *XML Information Set (Second Edition)*, R. Tobin and J. Cowan, Feb 2004, *http://www.w3.org/TR/xml-infoset*
   D.v *XML Processor Profiles*, H. Thompson, N. Walsh and J. Fuller, Jan 2012, *http://www.w3.org/TR/xml-proc-profiles*
   D.vi *XProc: an XML Pipeline Language*, H. Thompson, A. Milowski and N. Walsh, May 2010, *http://www.w3.org/TR/xproc*
   D.vii *XQuery 1.0 and XPath 2.0 Formal Semantics (Second Edition)*, M. Fernández, M. Rys, K. Rose, P. Fankhauser, M. Dyck, J. Siméon, D. Draper, A. Malhotra and P. Wadler, Dec 2010, *http://www.w3.org/TR/xquery-semantics*