State of the art of streaming

Why W3C XProc, W3C XSLT WGs and ISO SC34 WG 1 are looking (so) closely at streaming?

Balisage Montréal 2008

Présentation Innovimax

De nombreuses technologies émergent chaque jour et toute société a besoin de s'approprier et d'intégrer ces atouts pour leurs développements. A travers la jungle des sigles, XML, Java, .Net, SOA, XSLT, AJAX, XUL, vous cherchez à comprendre et à utiliser la bonne technologie. La société *Innovimax* a été créée dans cette optique. *Innovimax* vous accompagne dans toutes les phases de votre projet en vous fournissant le conseil, le suivi, les prestations et la formation nécessaire à sa bonne réalisation.

Basée à Paris (France), *Innovimax* est une société privée spécialisée en technologies émergentes et en innovations. *Innovimax* propose donc ses services regroupés autour de quatre pôles : *Média*, *Software*, *Consulting* et *Learning.*

Contactez-nous / Contact us

Innovimax
9, impasse des Orteaux - 75020 Paris

Tél: +33 8 72 47 57 87
Fax: +33 1 43 56 17 46

contactus@innovimax.fr
http://www.innovimax.fr

SARL au capital de 10.000 €
RCS Paris 488.018.631

Innovimax Learning

Le pôle **Innovimax Learning** est le second pôle important de la société. Point clefs de la réussite de toutes évolutions technologiques, la formation se doit d'être clair, accessible et adaptée. Les technologies émergentes sont légions et il vous semble difficile de faire le tri parmi les sigles : HTML, XML, XSLT, CSS, AJAX. Pour ce faire, le département Learning d'Innovimax vous propose des formations pour vous y retrouver dans ce dialecte et savoir quels sont les technologies dont vous avez besoin.

A destination des décideurs, les formations *Manager* vous propose des formations concrètes expliquant les tenants et les aboutissant de chaque technologie, les gains attendus et les success stories.

A destination des utilisateurs/collaborateurs, les formations *Client* vous propose des formations essentiellement axées sur les technologies en place dans leur environnement de travail et rétablisse les réflexes à prendre avec les nouvelles technologies (sauvegarde, sécurité, spam, etc.)

A destination des acteurs technologiques, les formations *Designer* vous propose des formations ciblées sur votre domaine (Web, graphique, applicatif) afin de vous enseigner les bases approfondies de chacune des technologies et d'être en capacité de mettre rapidement en application ces technologies.

Innovimax & Standards

W3C, ISO & AFNOR

Innovimax is a member of W3C at XSL, XML Processing, XQuery, CSS et MathML WG and apply those standards to its customers.

We are also member of AFNOR (French national Body) and editor of DSDL Part 10

Greetings – 1/2

Hello

Bonjour

Kia ora tatou

Mohamed ZERGAOUI

- INNOVIMAX (small French company)
- W3C Member (XSL, XProc, XQuery, ...)
- ISO DSDL invited expert (editor of Part 10)
- AFNOR (French national body) ; French Official Publication Office (DJO) ; OECD Publication
- Studies : SGML and XML ecosystem
- Hobbies : SGML and XML ecosystem
- Work : Make a guess ?

Definition

Definition of Streaming

- Difficult to define
- May be we should start by what is not streaming
- Proposition (necessary condition) :
  - A transformation that do not mandates a blocking preprocess of the document
    - blocking means that you should wait until the end (DOM loading, Full Validation, Indexing, etc.)

Definition of Streaming

- Multiple ways to handle that meaning
  - Related to memory use of the process
  - Related to size of the input
  - Related to latency time of the process

Definition of Streaming

- Related to memory use of the process
  - Bounded (independent of the size)
    - small device
  - Grow slower than linear (grow slowly than the document size)
    - $\varepsilon * \text{InputFileSize}$ with $0 < \varepsilon < 1$
    - $\varepsilon \geq 1$ may mean everything is streamable?

Definition of Streaming

- Related to memory use of the process
  - Questions:
    - Which memory use are we talking about?
      - ...and isn't memory use related to Complexity Theory?

Definition of Streaming

- Related to size of the input
  - Infinite input (Quotes, logs, etc.)
    - What does it mean (Infoset, XDM, Validation) ?
  - Is process time a good candidate ?
    - Process time belongs to Complexity Theory, too
  - Incident question:
    - Hence, is streaming out of reach of NP-complete algorithms ?

Definition of Streaming

- Related to latency time of the process
  - First input event/First output event
  - Last input event/Last output event (non infinite)
  - Mean

    Need to have some hints on relations between input and output;

    Difficult in general case;

    Not so difficult in almost-copy, decorator or wrapper design pattern

Definition of Streaming

- Related to latency time of the process
  - First input event/First output event (FIFO)
    - *The lower this one is, the better we can queue processing*
  - Last input event/Last output event (LILO) (non infinite)
    - *This one might augment in a streaming context (tradeoff is complex)*

Definition of Streaming

- Simple XPath examples
  - **//*, //A :** node()* (Streamable => No memory, processing time ≈ 0, FIFO = 0, LILO=0, Process= ALL)
  - **count(//A) :** number() (Streamable => small memory, processing time ≈ 0, FIFO = Huge, LILO=0, Process=ALL)
  - **count(//A) > 3 :** boolean() (Streamable => small memory, processing time ≈ 0, FIFO = Huge or small, LILO=N/A or 0, Process=4 elements A at max)

Definition of Streaming

- Less simple XPath examples
  - **//A[last()] :** node()? (Streamable => memory=max size of A, processing time ≈ 0, FIFO = Huge, LILO=0, Process=ALL)
  - **//A[(2+count(.//node()))*(2+count(.//@*)) mod 7 = 1] :** node()* (Streamable => memory=max size of A, processing time important, FIFO = Huge, LILO=0, Process=ALL)
  - **/*[*[last()]/@*] :** node()? (Streamable => memory=**document size**, processing time ≈ 0, FIFO = Huge, LILO=0, Process=ALL)

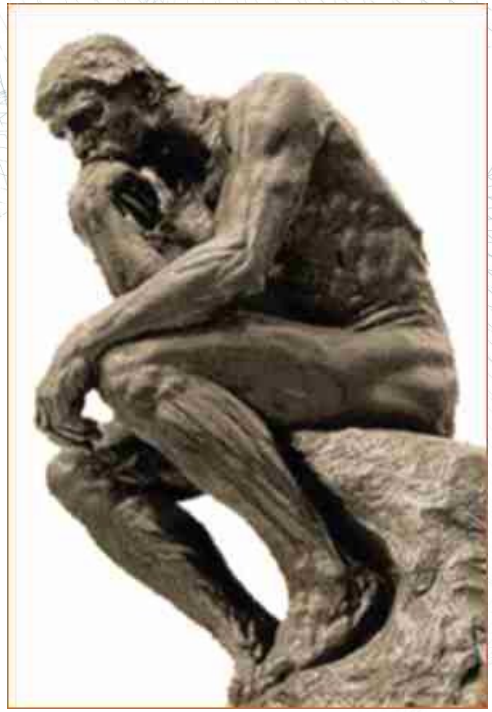
Definition of Streaming

- Pragmatic definitions
  - « Don't hold the input tree in memory » (Comity of the XML Forest Defense)
  - « Just use the minimum » (Comity of IT Communists)
  - « Just use the resource you find » (Yet another Greenpeace Comity)
  - « Don't hold anything » (Comity of XML Streaming Extremists)

Definition of Streaming

- Pragmatic consequences
  - Need other form of memory (buffering, state automaton)
  - Swap or reread (multipass or random access)
    - Multipass :
      - fixed number of pass
      - Unknown : need to have a way to test the endpoint (fixed point, in case of sorting)
    - Random access ? How that ?

Definition of Streaming



- Isn't just streaming a philosphy ?
- To stream or not streaming ?
- ...or another name for optimisation ? (as the trade off Memory vs. Time)

Processing ?

- Need to define processing ?
  - Not really
- Processing:

  Action to generate a ***result*** from zero or one ***main input source***, and zero or more auxillary input sources, with respects to zero or more ***parameters.***
- ***Use cases : Generate TOC, Generate HTML file, Generate FO from SVG, etc.***

IO ?

- Need to define inputs and outputs ?
  - Of course inputs are XML, but which form ?
- How to see an XML Instance is important
  - Byte Stream (very low level)
  - Character Stream (low level)
  - XML Event stream (mixed level)
  - Tree (XDM 1.0 or 2.0, DOM, etc.)

IO ?

- Need to define inputs and outputs ?
  - Of course inputs are XML, but which form ?
- How to see an XML Instance is important
  - Byte Stream (very low level) DECODING
  - Character Stream (low level) LEXICAL
  - XML Event stream (mixed level) GRAMMAR
  - Tree (XDM 1.0 or 2.0, DOM, etc.) STRUCTURE

Parsing and Lexical analysis

- Decoding and lexical analysis is a fully streamable process;
  XML has been designed for that: no look ahead, no complex model (well it's not SGML ☺)
- Grammar (of XML) can be streamed (SAX, StAX)
- A tree is a tree, but tree like representation can be streamed too : XDM Streamed (not fully equivalent to SAX and StAX – no CDATA, no Entities, etc.)

Validation

- Parsing is good
- But validating could be better
  - Is DTD streamable ? Yes definitely !
  - Is XML Schema streamable ?
    - MSMcQ says yes
    - Some other says ...not always (PSVI annotated)
  - Is Relax NG streamable ?
    - Of course, that's Tree Automata Theory !!

Validation

- Anyway, big troubles with the very notion of errors
  - Tree state logic (Valid, Invalid, Dunno)
  - Pipeline problem already known in chip industry (invalidation, atomicity)
  - Can a stream be valid to a certain point and not after ? What does it mean ?
  - Can we recover on error ? Wasn't the redundancy </end> done for that ?

State of the Art

State of the Art

- XSLT 1.0 / XPath 1.0 (Clark, DeRose, 1999: W3C Rec)
  - No streaming facilities
  - Worse the specs enforce « stability » --> two access to same info, need to answer same result
- SAX 1/2 (Megginson and al., 1998, 2001: de facto Rec)
  - Dedicated to streaming
  - No help for complex transformations

State of the Art

- XSLT 2/XPath 2 (Mike Kay and al., 2007: W3C Rec)
  - Even less room for streaming
  - More high level facilities
- XQuery 1/XPath 2 (Chamberlin, Robie and al., 2007: W3C Rec)
  - Designed for streaming
  - ...but also designed for database ☹
  - Not very handy for document transformations (see XTech 2005, Mike Kay's « Comparing XSLT and XQuery »)

State of the Art

- STX 1.0/STXPath (Cimprich, Becker and al. 2007 : WD)
  - Designed for streaming
  - Special (poor ?) subset of XPath 2.0
  - Higher level than other proposal
  - XSLT Fans : not functional, Yet Another XSLT-like
  - W3C folks look at it, DSDL folks look at it

And what about our subject ?

Why? – 2/9

Why W3C XProc, W3C XSLT WGs and ISO SC34 WG 1 are looking closely at streaming?

Why ?

- Streaming is ~~kewl~~ ~~cuil~~ cool
  - Less memory
  - Less time
  - probably more tricky

Why ?

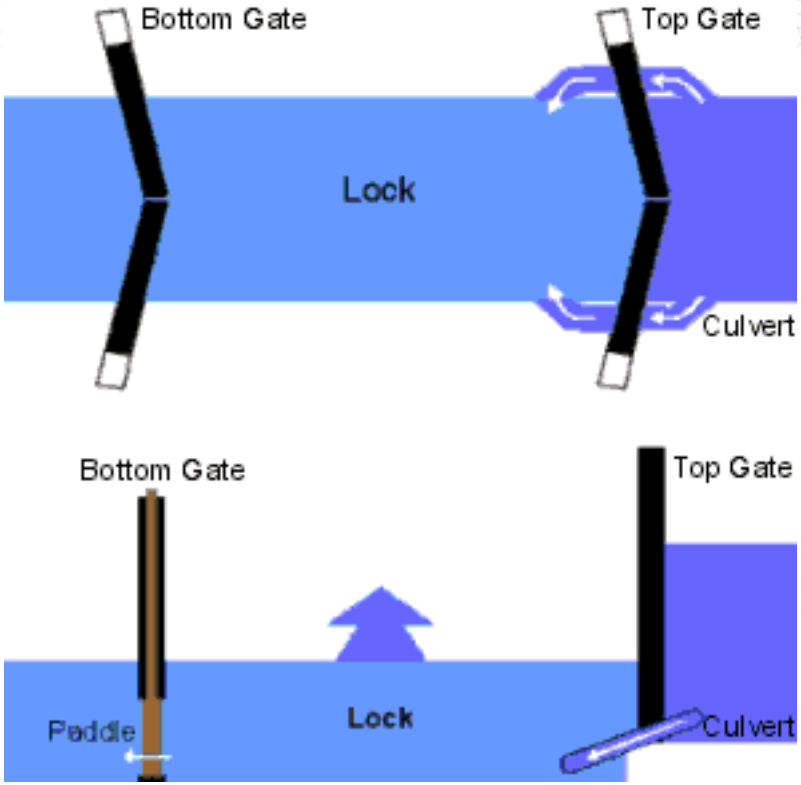
- Memory is now cheap ? so why bother ?
  - probably but time is not
  -

Why ?

- But Database (NXD and others) already do the job ?
  - Extra import (and indexing) time for one-shot transform

Why ?

- Well, this is XML processing today



Why ?

- You see what I mean

Why ?

- But we want this

Why ?

- ...or this...

Why ?

- ...and this...

The BANZAÏ PIPELINE !!!!

WG interest

Validation (DSDL)

- DSDL, a great WG (Clark, Murata, Tenisson, …)
  - Relax NG (XML and compact) : Grammar
  - Schematron : Rules
  - NVDL : Namespace aware validation and dispatch
  - DTLL (Datatype library)
  - DSRL (Renaming tools)
  - CRDL (Character repository)

Validation (DSDL)

- Part 6 : Streaming
- The current editor is also the editor of STX
- DSDL looks at STX for streaming Schematron

Validation (DSDL)

- Great Soloists, but need a conductor ...
- Part 10 : Validation Management
- DSDL looks at XProc
- XProc need streaming to be efficient

XSL WG

- Early 2007, there was plan to make a XG for Streaming (Thanks to Nokia, Art Barstow)
- Then XSL was interested in working on
- Most of the XG decided to join XSL
- Recently, Intel join XSL
    - Hardware implementors join the group : EXCITING

XSL WG

- Still fighting to know what is the definition of streaming
- We have a very early (internal) WD of requirements
- We will have an important Face to Face in Prague at the end of september

Next Step

- XSLT 2.0+
  - Introduction of instruction and function that are streaming friendly
    - xsl:iterate instruction (already available in Saxon 9.1, as saxon:iterate) : sequential for-each
    - stream() function (idem, as saxon:stream(), or read-once) : removing stability

XProc WG

- XProc 1.0/XPath 1.0 (Walsh and al., 2008 : W3C WD (second last call is now ready to be published))
  - Even more high level (combining steps of transform)
  - Designed to keep maximum streaming facilities (hard)
  - More details (See the spec http://w3.org/TR/xproc)
  - Will be the base for DSDL Part 10
  - Isn't everyone waiting for it ?

XProc WG

- Implementation being created
  - Norman Walsh's implementation (Java, StaX, Saxon)
  - Jim Fuller's in XQuery
  - Orbeon
  - All are in pipes for the moment...

Other work

- Other approach : mathematical and theoretical
- Mainly based on OCaml (functionnal language) :
  - CDuce (Frish) : highly typed, higher order
  - XTiSP (Nakano)
  - XStream (Frisch, Nakano) : Turing complete, term rewriting ⚐ Powerful need to take a look

Other work

- The Graal of streaming
  - For academic:
    The biggest XPath subset fully streamable : lots of research with no obvious solution
  - New way : tree automata theory, visibility pushdown automata (WWW2007)

Other work

- Pragmatic approach:
  - Keep all XPath and just ~~optimize~~stream when it is possible --> New academic field : schema aware static analysis
  - Possible enhancement : help the processor with some hint on what to drop from memory ---> proprietary extension in Saxon for example

Products

Products

- STX: Joost (Java, Sourceforge, May 28 2008), XML:STX (Perl, CPAN, v0.43 Dec 22 2004)
- Cocoon (Java, Apache, v2.2.0 May 15 2008)
- XSLT 2: Saxon (Java and .Net, Sourceforge, v9.1.0.1 Jul 2 2008), Gestalt (Eiffel, Sourceforge, 1.0, Dec 16 2007), Altova
- ServingXML (Java, Sourceforge, v0.9.2c Aug 13 2008)
- The philosophy here: Just let the product do ~~optimisation~~ streaming when it can!

API

API

- Event model
  - SAX (Push) : (Java, C, Perl, etc.)
  - StAX (Pull) : JDK 6, JAXP 1.4
    - Intermediate : XOM (Java) v1.1 (2005)
- Another approach
  - Based on Binary XML (Efficient XML Interchange (W3C WG)
    - VTD-XML : Java, C, C#, Sourceforge, v2.1 Jun 14 2007 --> XPath not enough complete

Languages

Languages

- CDuce (OCaml extension)
- XDuce (OCaml extension)
- XStream (make a guess ?)

Languages

- XML as first class citizen
  - E4X (Javascript)
  - XLinq (C#)
  - XJ (Java extension, Nov 22 2006)
  - xmlsh (Shell)

Languages

- Omnimark (Own language / Proprietary)
- Balise (Has anyone else heard about ?)
- And many more ad hoc garage version

Specifications

Specifications

- STX
- XML Processing (XProc)

- Another approach :
- XQuery Update

History

History

- How did we do that before ?
- SGML time ?
  - SGML vision of processing (Balise and Omnimark)
  - Optimized but not streaming
    - Because SGML needed always a full preprocess

Research fields

Research Fields – 2/10

- Active research fields :
  - XPath subset
  - Static analysis
  - Model aware
- Still to come
  - Efficient representation of Streams for buffer
  - Error management

Research Fields – 3/10

Active research fields :

- Annoying VERY USEFUL things : sorting, grouping
  - Removed from STX (Sorting)
  - Difficult to stream (obvious tortuous use case)
- Let's get a List !

Research fields

- Constraints
- Normalizing
- Streamable path
- Multilayer transformation
- Constraints aware streamable path
- Static analysis of XSLT and XQuery to detect streamable instances
- What are the needed evolutions of the cursor model?

Research fields

- Constraints
  - Schematron: today's implementation is in XSLT 2.0 for last ISO Schematron and 1.5. But DSDL is interested in using STX to implement ISO Schematron (it's already allowed but less expressive: the aim is to keep expressivity)
  - XML Schema 1.1 is trying to implement Constraints that could be streamable. They were on a path to take a very very small subset of Xpath, but they're now back to full XPath 2.0 (with some trick, like copy...)

Research fields

- Normalizing
  - Normalizing documents (Canonical XML)
  - Normalizing « frozen stream » (a.k.a Stream buffers, VTD-XML)

Research fields

- Streamable path
  - Old gimmick
  - Subset
    - Academic result : cannot be used seriously in implementation (XPath without predicate, keep predicate but remove all but 2 axes, etc.)
  - XPath rewriting
    - Interesting but it would be better if done dynamically

Research fields

- Multilayer transformation
  - Definition this explicitly as a Design Pattern
  - Multipass or Multilayer ?
    - Layer could be different / Pass the same process many time
    - Need streamable comparison

Reseach fields

- Constraints aware streamable path
- That's the most interesting field at the moment
- XSLT 2.0 has defined a Schema Aware version
- Lots of work on XQuery (Colazzo 2006), XPath (Geneves 2006),

Research fields

- What are the needed evolutions of the cursor model?
  - Cursor has to be bidirectionnal (forward and backward move on the input document)
  - XQuery Update Facility ---> different approach : modify the document not transform it !

Optimization – 1/2

Optimising

Isn't streaming just a high level approach to optimization ?

- Hints on the answer
  - Fuzzyness of the definition
  - Difficult
  - Complexity Theory in the hood...
  - XML is 10 years now

Questions ? – 1/1

innovimax