XML Processing by Streaming

XML Prague 2007

Bienvenue

2007/05/15 version

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 @ W3C

W3C (World Wide Web Consortium):

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

Greetings – 1/4

Hello

Dobrý den

Bonjour

Greetings – 2/4

Mohamed ZERGAOUI

- INNOVIMAX (small French company)
- W3C Member (XSL, XProc, XQuery, ...)
- ISO DSDL invited expert
- 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 ?

Why harassing you for more than one hour ?

- Nice place (many nice thing around)
- Drink, foods
- Nice blue shirt around (missing angle brackets ☹)
- Look forward for tricky questions
-

Why am I wearing so serious clothing ?

- Because I'm from Paris
- To look more serious
- Because that's Norm's Birthday (**<NaN />** years)
- To be ready for disco tonight
- ... So don't be scared, if tomorrow, I'm wearing the same clothes....

Plan – 1/3

Plan

First part

- Definition of Streaming
- State of the art
- Products
- API
- Languages
- Specifications
- Evolutions
- History

Second part

- Fields of research related to Streaming
- Interest from WG
- Questions ?

FIRST PART

FIRST PART

Definitions – 1/13

Definition

Definition of Streaming

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

Definition of Streaming

- Related to memory use of the process
  - Bounded ?
  - Grow slower than linear : o(InputFileSize)
  - Isn't memory use related to Complexity Theory ?

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 size of the input
  - Infinite input (Quotes, logs, etc.)
  - Is process time a good candidate ?
    Process time belongs to Complexity Theory, too
  - Incident question:
    Is streaming out of reach of NP-complete programs ?
    (not so naïve answer : no)

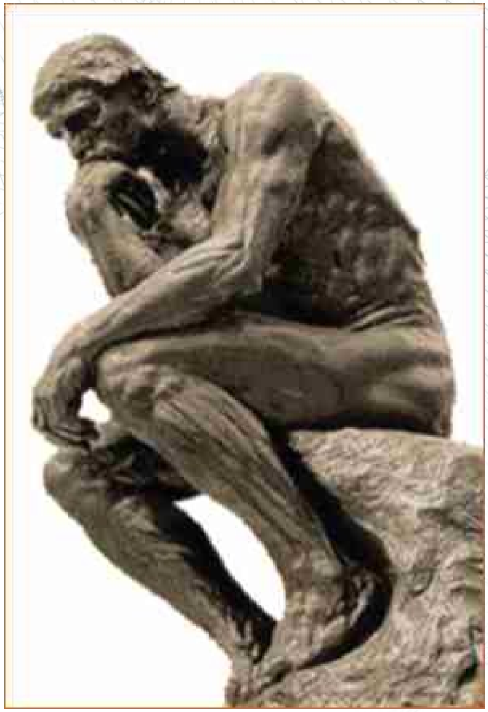
Definition of Streaming

- 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 know if you read a state (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.)

Definitions – 11/13

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

Definitions – 12/13

Parsing and Lexical analysis

- Decoding and lexical analisys is a fully streamable process;
  XML has been designed for that : no look ahead, no complex model
- Grammar (of XML) can be streamed (SAX, StAX)
- A tree is a tree, but tree like representation can be streamed too (take care of forward axis):
  XDM Streamed (not fully equivalent to SAX and StAX)

Validation

- Parsing is good
- But validating could be better
- Is DTD stremable ? Yes definitely !
- Is XML Schema streamable ?
  - MSM says yes
  - Some other says ...not really
- Is Relax NG streamble ?
  - Of course, that's Tree Automata Theory !!

State of the Art – 1/8

State of the Art

State of the Art

- XSLT 1.0 / XPath 1.0 (Clark, DeRose, 1999: W3C Rec)
  - No streaming facilities
  - Worse the spec 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 subset of XPath 2.0 (intersect with ancestor)
  - Higher level than other proposal
  - XSLT Fans : not functionnal, Yet Another XSLT-like
  - W3C folks look at it, DSDL folks look at it

State of the Art

- XProc 1.0/XPath 1.0 (Walsh and al., 2007 : W3C WD)
  - Even more high level (combining steps of transform)
  - Designed to keep maximum streaming facilities (hard)
  - More details (Norm's presentation)
  - DSDL folks look at it (for Validation Management)
  - Isn't everyone waiting for it ?

State of the Art

- 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

State of the Art

- 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)

State of the Art

- 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 ---> propriatary extension in Saxon for example

Products – 1/2

Products

Products

- STX : Joost (Java, Sourceforge, May 29 2007), XML:STX (Perl, CPAN, v0.43 Dec 22 2004)
- Cocoon (Java, Apache, v2.1.10 Dec 21 2006)
- XSLT 2 : Saxon (Java and .Net, Sourceforge, v8.9 Feb 12 2007), Gestalt (Eiffel, Sourceforge, vBeta 1, Apr 22 2006), Altova (?
- ServingXML (Java, Sourceforge, v0.7.0 Jun 13 2007)
- The philosophy here : Just let the product do ~~optimisation~~ streaming when it can !

API – 1/2

API

API

- Event model
  - SAX (Push) : (Java, C, Perel, etc.)
  - StAX (Pull) : JDK 6, JAXP 1.4
    - Intermeditate : 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 – 1/4

Languages

Languages

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

Languages – 3/4

Languages

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

Languages

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

Specifications – 1/2

Specifications

Specifications

- STX
- XML Processing (XProc)
-
- Another approach :
- XQuery Update (Is this XML Processing : almost idempotent transformation could be written easily

Evolutions – 1/2

Evolutions

Evolutions

- XSL WG looking at streaming
- DSDL looking at STX
- DSDL looking at Xproc
- Intel looking at streaming
- etc...

History – 1/2

History

History

- How did we do that before ?
- SGML time ?
- SGML vision of processing (Balise and Omnimark)
- cursor idea that can be find in Arbortext OID in ACL (not fully streamable) ---> see StAX in XML (remove reverse parsing)

SECOND PART

SECOND PART

Research Fields – 1/10

Research fields

Research Fields – 2/10

Active research fields :

- See above
  - XPath subset
  - Static analysis
  - Model aware
- TBD
  - Efficient representation of Streams for buffer

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 !

Reseach 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?

Reseach fields

- Constraints
  - Schematron : today implementation is 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 have taken a very very small subset of Xpathfor the current draft. But Mike Kay has gone rescue them...

Reseach fields

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

Reseach fields

- Streamable path
  - Old gimick
  - 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

Reseach fields

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

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 – 10/10

Research fieldsl

- 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 !

WG interest – 1/5

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
- 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

- Start of 2007, 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

Optimisation – 1/2

Optimising

Optimisation – 2/2

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

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

Questions ? – 1/1

15/06/2007

innovimax learning

innovimax