The XML World View
– a personal vision with challenges

Kristoffer H. Rose
krisrose@us.ibm.com
Dreaming
An easy way to access & query all information in the world*

* or Enterprise or ...
So, what's “the world”?

- **Information** in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- **Organizations** of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
  - Live feeds (“mylife”)
  - ...

- Organizations of the information:
  - Global addressing.
  - Interdocument relationships.
  - Embeddings.
  - ...

- Information in units:
  - Documents.
  - Databases.
  - Spreadsheets.
  - Mail and Notes.
Why is XML interesting?

- Everything with an **XML data model**\* can make use of the XML standards:
  - access with XPath,
  - manipulation and construction with XSLT (and XQuery),
  - full text search (in preparation),
  - standardized distribution (web services),
  - standardized encryption,
  - ...and more!

* Specifically the “Xpath/XQuery Data Model” (that maps to the XML Infoset).
So how do we get XML data models of everything?

_Virtualize,_

_Let no one else's work evade your eyes,_

_Remember why the good Lord made your eyes,_

_So don't shade your eyes,_

_But virtualize, virtualize, virtualize..._

_Only be sure always to call it please ... search._

With apologies to Tom Lehrer
It's the **Syntactic Web**!

- “Live” virtual instances of XML Data Models for
  - addressing, and
  - information.

- But “Just syntax”...
  - Expose *addressing directly in XML hierarchy*.
  - Expose *structure of information directly in XML*.

- Make sure we can “grow” the syntax as we gradually understand what we need from it.
  - *Make it easy to build derived virtual XML views.*
Putting it all together!

- **Virtual XML federation**
- **Actual organization**
- **Virtualization XML adapters**
- **Heterogenous information units**
Waking up
The world* is a big place.
Organization (I): file systems

- /root/dir[@name='etc']/file[@name='passwd']/@text
  - Directories are represented as elements.
  - The root is the root directory.
  - The children of a directory are the sub-directories, files, etc. (links).
  - Actual file contents text is available but in general the child of a file should be the root of it's XML representation.
Organization (II): the web

- /www/link[@href = "http://ibm.com/developerworks"]
- /page/link[contains(@href, "watson")]
- /file[@type = "application/pdf"]

- Root is virtual list of all possible links.
- Children of a link is the referenced web page and/or it's representation as a file.
- Children of web pages are their links (recursive).
- Don't build this for real...
Information units (I): the structured case

- Easy for native XML and structured documents (SGML).
- Relational data maps easily into XML.
Information units (II): customized

- .../mail/message[@from = “krisrose@us.ibm.com”]
  /attachment[@type = “application/pdf”]
- .../passwd/record[@uid = “krisrose”]/@full-name

→ In each case the “surface structure” is mapped into XML (DFDL).
Derived views

- .../file/transform[@by = "my.xsl"]/...

  ➔ Live transformed data!
Monday morning...
The world*
always
has more complexity
than we think...
Persistence

• The world changes.
  ➔ *We do not yet have a nice XML-level standard for updates.*

• Can we define persistent subsets (“profiles”) of the data model?
  ➔ *For example: order cannot be observed, children only added, etc.*

• How is it ensured that updates are well defined on virtual XML data models?
  ➔ *Constrains the “cleverness” of the virtual models.*
Evolution

- Data evolves.
  - XML Schema “evolution” is still a research topic.
  - Other formats each have their own notion of evolution e.g., (e.g., version control).

- Can virtual XML specifications be robust wrt. evolution?
  - Seing “evolved” data as “derived” could help.
  - Constrains the “cleverness” of the virtual XML models.
Challenges...

- Can we map all our data into useful virtual XML?
  - *This is happening already.*

- Can we build virtual XML data models of the various ways the world* is organized?
  - *Is execution of XPath (etc.) over such virtual data feasible?*

- Does it scale properly?
  - *Even over multiple derivations/evolutions and mutations?*
  - *Can multiple organizational principles coexist?*

- What is needed to seed the growth of this?
Thank you
It's Here: HTTP + Google™

- Just text (and URL) search.
- Results returned in their native form or as text.
- No combination of searches (join/filtering).
- No customization of result format.
Why Not Semantic Web?

- Requires full specification at data source.
- Hard to retrofit onto legacy data.
- Value is in shared “ontology space”...is the world ready to share?