Sanitizing using Metadata in MetaXQuery

Hao Jin and Curtis Dyreson
School of E.E. and Computer Science
Washington State University
USA

Outline
• XQuery
• Metadata
• Sanitize
• Experiments
• Summary

Book Data in XML

✦ <book>
✦ <authors>
✦ <author>Millicent Marigold</author>
✦ <author>Montana Marigold</author>
✦ </authors>
✦ <content>
✦ <chapter>Usability testing is…</chapter>
✦ …
✦ </content>
✦ </book>

Parsed into DOM

✦ book
✦ authors
✦ content
✦ author
✦ author
✦ Millicent Marigold
✦ Montana Marigold

Querying in XQuery

✦ FLWR expression
  ■ For some nodes
    • Use XPath locate nodes
  ■ Let – grab other nodes
  ■ Where some condition holds
  ■ Return some value

✦ Example: List book authors
  for each $author in //book/authors/author
  return $author

Locating Nodes

✦ /book/authors/author
  [*<author>Millicent Marigold</author>*,
   *<author>Montana Marigold</author>*]
Outline

- XQuery
- Metadata
- Sanitize
- Experiments
- Summary

Metadata

- Metadata is *data about data*

  - Descriptive
    - Language
    - Subject (Dublin Core)
    - Metadata author

  - Proscriptive
    - Security
    - Privacy
    - Time (when in the data store)

Embedding Metadata in XML

```xml
<book>
  <meta:metadata>
    <meta:subjects>
      <meta:subject>Usability testing</meta:subject>
    </meta:subjects>
    ...
  </meta:metadata>
  <content>
    <chapter>Usability testing is...</chapter>
    ...
  </content>
</book>
```

Using RDF

```xml
<RDF xmlns="rdf-syntax-ns#">
  <Description about="document(books.xml)/book[1]">
    <subject>Usability testing</subject>
  </Description>
  ...
</RDF>
```

MetaDOM

- Support arbitrary levels of metadata
- Reuses DOM at each level
- Separates metadata and data into different scopes
  - //book/authors/author – locates only data nodes
  - //book*/user – ascend into the security metadata (MetaXPath)
- Upwards compatible with DOM/XPath/XQuery/XSLT etc.
Outline

- XQuery
- Metadata
- Sanitize
- Experiments
- Summary

Sanitize

- Ensures that the data conforms to the metadata perspective
- Semantics differs for each kind of metadata
  - Language – equality
  - Time – overlaps
  - Implementation specific
- Recursively evaluate each level of metadata

Querying with MetaXQuery

- Data-only (implicit metadata)
  - Implicit perspective
    - Data as of Time 3
    - Security of Joe
    - Language is English
  - List book authors in MetaXQuery
    
      for each $author in
      metaxq:sanitizeByPerspective(//book/authors/author, P)
      return $author

- Metadata-only (implicit meta-metadata)
  - List the users who can access a node

Sanitizing //book/authors/author

Certify

- Metadata/data pulled from various sources
  - May conflict
- Is MetaDOM node reachable?
  - Transaction time semantics – a child can exist only if a parent exists, certify that child time is subset of parent time
- metaxq:certify(//book/authors/author)
Outline

- XQuery
- Metadata
- Sanitize
- Experiments
- Summary

Experiments – Testing Certify/Sanitize

- Overhead is linear is amount of metadata

Conclusions

- Extend XQuery to support metadata
- Extend querying
  - Implicit sanitizing using metadata
- Performance results suggest overhead is linear in amount of metadata

Thanks!