Applying intertextual semantics to Cyberjustice

Many reality checks for the price of one

Yves Marcoux
Université de Montréal, Canada

Overview of the talk

1. IS
2. Cyberjustice
3. IS + Cyberjustice
4. Conclusion

1. IS

IS – the elevator talk (1/2)

- Ever noticed that for some XML documents, reading aloud the start-tags and text content, makes a lot of sense?
- *What do you mean?*
- Well, if an XML document reads: "**`book title Autumn author John price currency USD 30`**", that sequence of words pretty much conveys the *meaning* of the document…

Visual aid
(not present in the elevator)

```
<book>
  <title>Autumn</title>
  <author>John</author>
  <price currency="USD">30</price>
</book>
```

IS – the elevator talk (2/2)

- *Right, that's because the markup tokens were carefully chosen*
- *But of course, it would be stupid not to choose meaningful tokens…*
- Exactly… I wonder how far we can push the idea; like, using very very long names
- *Sounds like a fun idea to play with… why don't you drop by my desk later today?*

Later in the day… (1/8)

- *Listen to that one:* "`this_book is_entitled Autumn is_authored_by John and_is_priced in_currency USD 30`"
- Ah! Same document as in the elevator, but you changed the markup tokens! Clever!
- *But I don't like long names; what if we could replace start-tags with arbitrary chunks of text? (Oh, and let's ignore attributes for now and assume US dollars…)*

Later in the day… (2/8)

- OK! For example, if I define this mapping:

```
<book> ⇒ "This book "
<title> ⇒ "is entitled "
<author> ⇒ ", it is authored by "
<price> ⇒ " and it costs (in USD) "
```

I get something slightly nicer:

```
"This book is entitled Autumn, it is
authored by John and it costs (in USD)
30"
```

Later in the day… (3/8)

- *We'd get more leeway if we could also replace some end-tags by text too…*
- *Try this:*

```
<book> ⇒ "This book "
<title> ⇒ "is entitled "
<author> ⇒ ", it is authored by "
<price> ⇒ " and it costs "
</price> ⇒ " US dollars."
```

- **"This book is entitled Autumn, it is authored by John and it costs 30 US dollars."**

Later in the day… (4/8)

- *Cool… But wait, that only works because all elements are required!*
  - Well, maybe the schema says they are…
- *And, then, you have to know the genIDs to set up the mapping, of course.*
  - Indeed, you can't do much of that if genIDs can be anything!
  - So, I guess all of this only makes sense for documents *valid* against a *known tagset*.

Later in the day… (5/8)

- *(pause) But I have a bigger problem… I don't think our mapping is right; I don't think the document actually means what our mapping says it means!*
- Wo, let me digest that sentence…
- *While you do that, let me rephrase what I think the document* does *mean…*

Later in the day… (6/8)

- *Our current mapping gives:*

  "This book is entitled (bla-bla)…"

  How can we talk about "this book"? Which book is that?

  I think what the document really says is rather only this:

  "There is that book, somewhere, which is entitled (bla-bla)…"

Later in the day… (7/8)

- You got a point… But, how can we decide which mapping is better?
- I think the only person entitled to determine whether your mapping for `book` is better than mine is the person who *invented* the element (and that would be the tagset designer, I guess).
- *Maybe the tagset designer didn't have any mapping in mind, only the genIDs…?*

Later in the day… (8/8)

- Well, in that case *any* mapping whatsoever (except the identity mapping `<book>` ⇒ `<book>`) would be an unjustified augmentation of the meaning intended by the tagset designer…
- *…*
- *I am tired and my head is slightly aching. I think I need to go rest a bit. Good bye.*

The next day…

- (Knock, knock…)
- I got that crazy idea: why don't we ask the tagset designer to provide her own mapping? That mapping would be considered as *defining* the semantics of each element in the tagset!
- *Your insight amazes me… It is indeed the craziest idea I have ever heard of. I hope we never meet again in an elevator!*

Introducing…

Intertextual Semantics

(EML2006)

IS: a few details (1/8)

- The name:
  - Semantics
    - It is a way to assign *meaning* to populated data structures (most importantly, XML documents)
  - Intertextual
    - The meaning of a document is a *network of interrelated texts*:
      - the concatenated segments
      - possibly other resources linked via hyperlinks

IS: a few details (2/8)

- Peritexts: text-before, text-after
  - They form an *IS specification* (ISS) for the model
- Can contain hyperlinks (URIs)
- Defined for each element type *in given* (possibly all) *ancestral contexts*

IS: a few details (3/8)

- Attribute handling:
  - Peritexts can contain "guarded" passages of the form

    `@attName[… @ …]`

  - Inserted only if attribute `attName` is present
  - `@` gets replaced by actual attribute value

Example

```
<story author="Bram Stocker">
  <para><person>Dracula</person> went to <place>France</place>.</para>
</story>
```

```
<iss xmlns="http://grds.ebsi.umontreal.ca/ns/ISS/">
  <rule paths="story" text-before="This document tells a tiny
    story.@author[ The author of this story is @.]"
            text-after="End of the tiny story."/>
  <rule paths="para" text-before="A bit of the story: " text-after=""/>
  <rule paths="person" text-before="THE PERSON NAMED " text-after=""/>
  <rule paths="place" text-before="THE PLACE NAMED " text-after=""/>
</iss>
```

```
This document tells a tiny story. The author of this story is Bram Stocker.

   A bit of the story:

       THE PERSON NAMED Dracula went to THE PLACE NAMED France .

End of the tiny story.
```

IS: a few details (4/8)

- More complex than simply mapping start- and end-tags, but still exceedingly simple
  - Simplicity intentional and important
  - Aim: documents that are understood likewise by modelers, authors, readers, developers…
  - Yet, it easily (but trivially) accounts for any existing way of assigning meaning to documents
  - Non-trivial IS rather hard for existing models

IS: a few details (5/8)

- Main ideas:
  - Bring the explanations about the data as close as possible to the data and in its context (i.e., in the peritexts), rather than provide them in physically separate locations (standoff documentation)
  - Make modelers aware of the interpretation paths necessary to understand the documents (and provide them in peritexts)

IS: a few details (6/8)

- IS suggest the following modeling method:
  - Start by thinking of peritexts; make them straightforward and simple
  - Determine content models
  - Markup tokens "abbreviate" peritexts
- "Literate modeling"
- Modular writing
- Actually fun…!

IS: a few details (7/8)

- Can be applied to other kinds of populated data structures:
  - Relational databases
  - Controlled vocabularies
  - etc.
- To interfaces and interface elements
- To interactions (in theory; not yet explored)

IS: a few details (8/8)

- The IS platform (Marcoux 2009):
  - Takes an XML document + an ISS, returns the IS "meaning" of the document (in HTML)
  - Used micro-formats in peritexts for attribute handling and hyperlinks
  - Textual "purism": no control over rendering
  - Heuristics for paragraph breaks
  - Automatic indentation

2. Cyberjustice

Cyberjustice

- Cyberjustice lab (M$CAN)
  - Faculty of Law, UdeM
  - Professor Karim Benyekhlef
- *Towards Cyberjustice*
  - SSHRC grant (M$CAN)
  - 25-30 researchers internationally
  - Leverage technologies for increasing access to justice and improving its various processes

3. IS + Cyberjustice

IS + Cyberjustice

- Justice: lots of documents
- Cyberjustice: lots of electronic documents
- Common goals:
  - Intelligibility by users (citizens)
- Project: modeling a document type (EDI)
  - Mid-2012 – late 2014
  - Can an IS model help in application dev?
  - Improve IS framework & platform

Modeling for a client (1/2)

- Need to communicate model hypotheses
  - More control over indentation & linebreaks
  - Emphasis & other styling
  - Orientation landmarks
  - Better printing
- Modeling now an iterative process
- Protocol for interviewing domain experts
  - Preparing example instances

Modeling for a client (2/2)

- Iterative process: model changes
  - Rationale Management Notes
  - Five types:
    - Modeling questions still pending
    - Modeling decisions
    - Tips for developers
    - Information for end-users
    - Notes for the modelers themselves
  - Appear in peritexts as icons: 
  - Content in tool-tip (or side-box if printing)

Examples

oXygen & web

Integrating in an IT project (1/2)

- There had been 2 earlier attempts at PDFying the EDI
- Our obsession with meaning made clear, however, that the nature of the target community for an authoring application was crucial. But it had never been decided on, nor their needs established!
  - Agenda modified accordingly

Integrating in an IT project (2/2)

- The "document approach" hard to understand / accept for IT people
- Search for meaning of elements often misunderstood as attempt to eliminate ambiguity / fuzziness
  - Easy to handle in XML
  - Mixed + choice content models
  - Define appropriate semantics

4. Conclusion

Conclusion

- Project very successful at prompting improvements to IS
- Not so successful at assessing usefulness of IS in application dev process
  - There was no application dev

Lessons learned

- Document approach difficult to sell
  - Problem within IT in general?
- Value of IS may lie more in:
  - asking good questions early on
  - addressing head-on terminology issues
  - quality (simplicity, reusability) of document model

  than helping directly the application dev itself

Thank you !

Questions ?

<ymarcoux@gmail.com>