TATA CONSULTANCY SERVICES
Experience certainty.

TCS Estimation and Quality Framework

Charlie Halpern-Hamu

Senior Solutions Architect,
Tata Consultancy Services
charlie.hamu@tcs.com

International Symposium on Quality Assurance and Quality Control in XML,
Monday, 2012 August 6, Hotel Europa, Montreal

Context, team and goals

- Conversion of live legal data
- 2 languages
- Source: 273 SGML DTDs
- Target: 1 XML DTD

- Unified project management, and content architecture
- Analyst specifiers in multiple cities
- Programming teams on multiple continents

- Need good estimations
- Need to be able to work in parallel for speed
- Errors must be caught before switchover

Tools and techniques

- Data analysis and estimation
  - Counting function points in source to estimate effort
  - Counting function points in target to estimate slope
  - Autogeneration of tight schemas to discover variation
- Quality assurance
  - Listing parent-child pairs to guide specification
  - Always program for context
  - Always program for all content
- Quality control
  - Source-to-Target comparison for lost or duplicated content
  - Autogeneration of word-wheels to highlight anomalous data
  - Use of XQuery-capable XML database to quickly review data

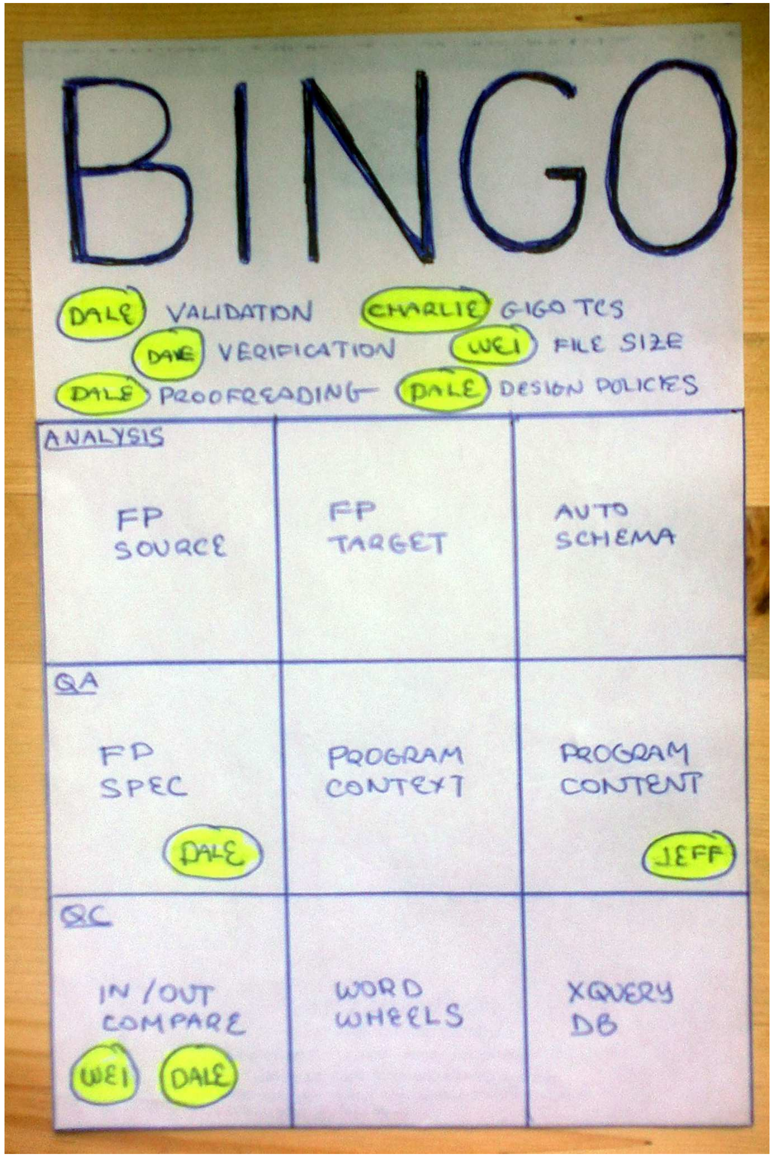
All about discovering *new* problems

I played during the other presentations

- Three sets of
  three techniques
  makes for a nice grid

- Not wanting to repeat
  what other people said,
  I kept track of the
  techniques they discussed

- I noticed that most of their
  techniques were focused
  on finding “known” errors,
  rather than exploring,
  hoping to discover trouble



BINGO

DALE VALIDATION
DAVE VERIFICATION
DALE PROOFREADING
CHARLIE GIGO TCS
WEI FILE SIZE
DALE DESIGN POLICIES

ANALYSIS FP SOURCE	FP TARGET	AUTO SCHEMA
QA FP SPEC DALE	PROGRAM CONTEXT	PROGRAM CONTENT JEFF
QC IN /OUT COMPARE WEI DALE	WORD WHEELS	XQUERY DB

Tools and techniques bingo sheet

B
I
N
G
O

Analysis Charlie
- *GIGO TCS*
- *Feedback* Murray
- *Fidelity* Jeff

QA Dale
- *Design policies*
- *Auto cleanup* Keith

QC Dale
- *Validation*
- *Verification* Keith
- *Proofreading* Wei
- *File size*
- *Link testing* Jeff
- *Regression*

FP source	*FP target*	*Auto-schema*
FP spec Dale	*Handle context*	*Handle content* Lynn Jeff
In/out compare Dale Wei	*Word wheels*	*XQuery DB*

Estimation: Function points in source

- Metric: conversion function points
  - Each parent/child pair counts for one
  - Each element/attribute pair counts for one
  - Text, processing instructions, and comments are ignored
- Function-point count estimates work
  - Specification, programming and QC come out to about an hour per function point
  - Already-specified function points can usually be deducted from estimate

Objective, transparent, repeatable

```
/chapter
colspec/@align
colspec/@colname
colspec/@colwidth
document_info/@document_name
document_info/@document_path
entry/@align
entry/@nameend
entry/@namest
entry/@valign
entry/bold
entry/break
entry/leader
figure/break
history/link
note/paragraph
note/section
paragraph/@font_size
paragraph/@keep-next
paragraph/@keep-previous
paragraph/@leading
paragraph/@no-keeps
paragraph/@prespace
paragraph/@type
paragraph/bold
paragraph/break
paragraph/link
```

Estimation: Function points in target

- When available, allows estimation of slope
  - Prospective input = 72
  - Prospective code = 73
  - Prospective annot = 53
  - Prospective output = 101
  - 101 / 72 = 40% bulk up
- Slope is more commonly estimated by number of text-pattern-to-element rows in spec

Recommendation: keep slope modest

```
dc:coverage/location:stateProvince
dc:identifier/@identifierScheme
dc:metadata/dc:coverage
dc:metadata/dc:identifier
default:colspec/@align
default:colspec/@colname
default:colspec/@colwidth
default:designator/@value
default:emphasis/@style
default:entry/@align
default:entry/@nameend
default:entry/@namest
default:heading/default:designator
default:heading/default:title
default:p/default:emphasis
default:p/default:table
default:row/default:entry
default:table/default:tgroup
default:tbody/default:row
default:tgroup/@cols
default:tgroup/default:colspec
default:tgroup/default:tbody
doc:metadata/@documentContentCountry
doc:metadata/dc:metadata
location:stateProvince/@stateProvinceCode
location:stateProvince/@stateProvinceCodeScheme
```

Estimation: Autogeneration of schemas

- Any number of XML files can be used to create a schema
- Schema created from 24*.xml 25*.xml
- Generated schema applied to 26*.xml
- Resulting validation errors indicate variations

```
…\Data\Input\260.xml:3:9115: error:
         element "hd" not allowed here; expected the element end-tag
…\Data\Input\260.xml:4:55383: error:
         element "br" not allowed here; expected the element end-tag,
         text or element "b", "doc_ref", "leader" or "statute_name"
…\Data\Input\260.xml:4:64593: error:
         element "br" not allowed here; expected the element end-tag,
         text or element "b", "doc_ref", "leader" or "statute_name"
```

- Similarly, can be used to find new patterns in the output

Highlights variation in the data

QA: List parent-child pairs

- Create framework for specification

Input	Context	Output	Notes
@align	colspec		
@align	entry		
b	entry		
b	paragraph		
b	sr		

- Create skeleton for programming

```
<template match="colspec/@align | entry/@align">
        <call-template name="to-do"/>
</template>

<template match="entry/b | paragraph/b | sr/b">
        <call-template name="to-do"/>
</template>
```

QA: Always program for context

- Mapping without context is deceptively fast

```
<template match="b">
        <call-template name="map-to-bold"/>
</template>
```

- But new contexts often require new consideration

```
<template match="entry/b">
        <call-template name="map-to-heading-cell"/>
</template>

<template match="paragraph/b">
        <call-template name="map-to-bold"/>
</template>
```

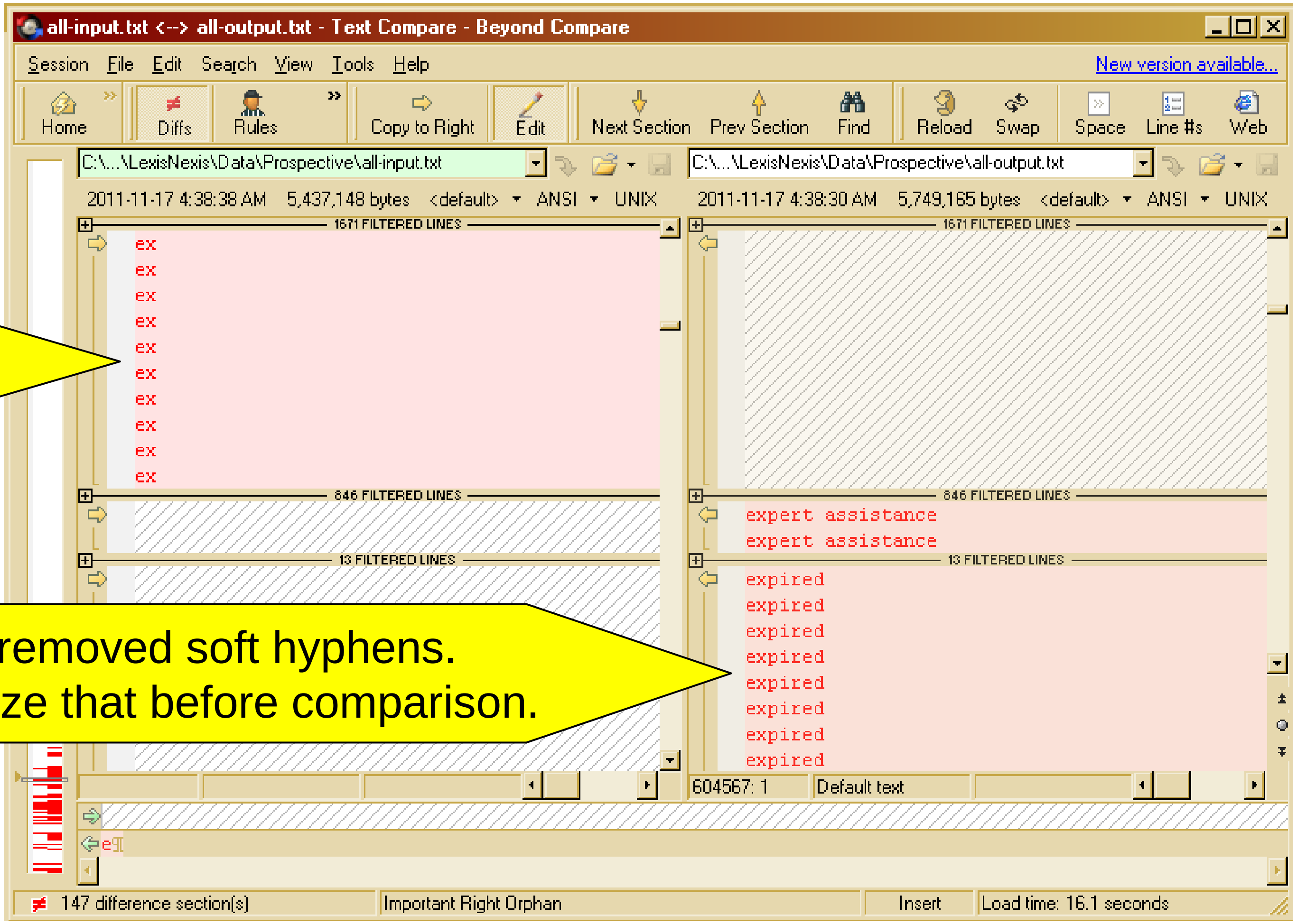
Defensive programming

QA: Always program for all content

- May script a paragraph by scripting:
  - paragraph/@font_size
  - paragraph/@keep-next
  - paragraph/@keep-previous
  - paragraph/@leading
  - paragraph/@no-keeps
  - paragraph/@type
- But will lose
  - paragraph/@prespace
- Unless we script for “and all other attributes”

Avoid data loss

QC: Source-to-Target comparison



all-input.txt <--> all-output.txt - Text Compare - Beyond Compare

C:\...\LexisNexis\Data\Prospective\all-input.txt

2011-11-17 4:38:38 AM 5,437,148 bytes <default> ANSI UNIX

C:\...\LexisNexis\Data\Prospective\all-output.txt

2011-11-17 4:38:30 AM 5,749,165 bytes <default> ANSI UNIX

```
ex
ex
ex
ex
ex
ex
ex
ex
ex
ex
```

```
expert assistance
expert assistance
```

```
expired
expired
expired
expired
expired
expired
expired
expired
```

604567: 1 Default text

e¶

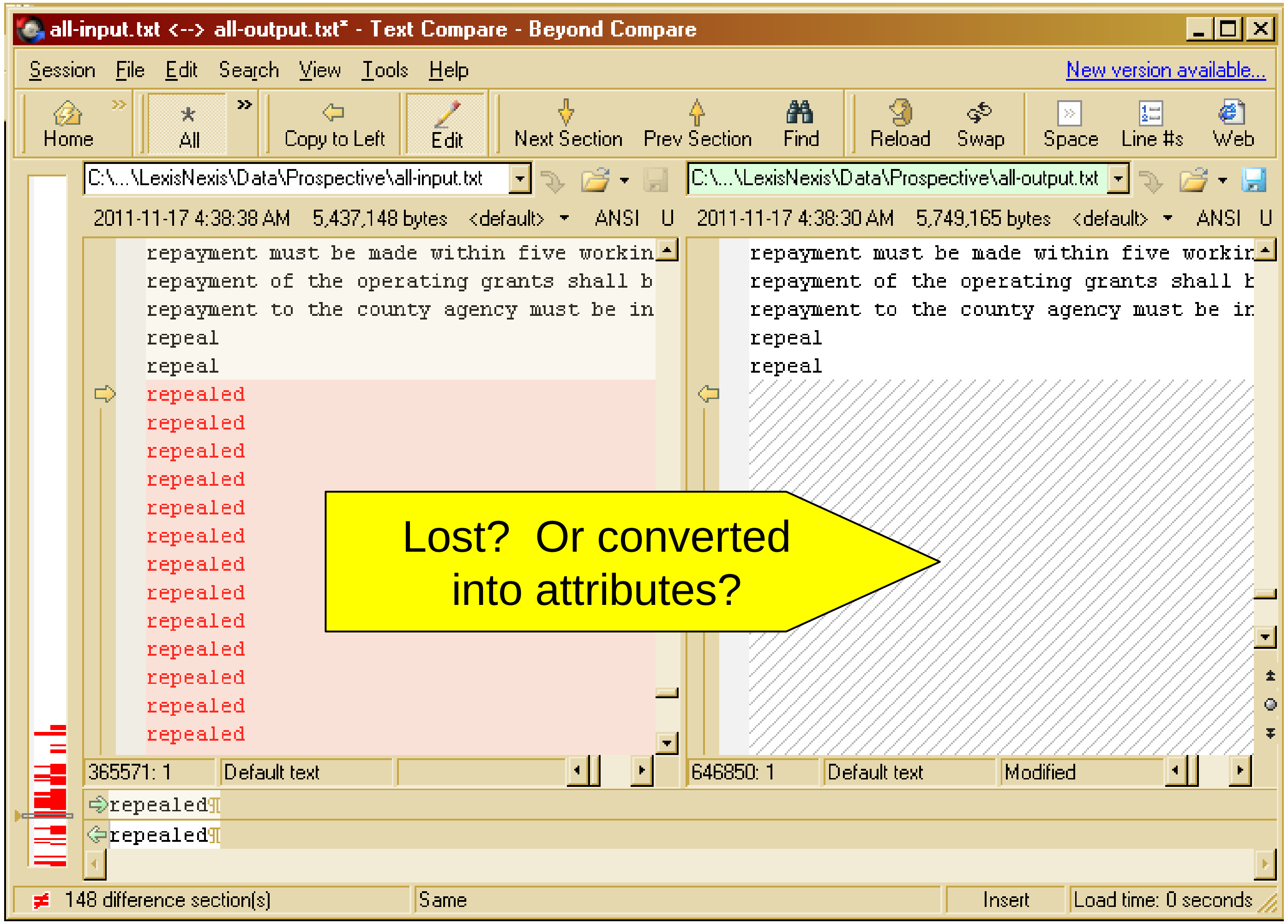
147 difference section(s) Important Right Orphan Insert Load time: 16.1 seconds

Where did these go?

We purposely removed soft hyphens.
We should normalize that before comparison.

Catch lost and duplicated content

QC: Source-to-Target comparison



```
C:\...\LexisNexis\Data\Prospective\all-input.txt
2011-11-17 4:38:38 AM   5,437,148 bytes   <default>   ANSI   U
repayment must be made within five workin
repayment of the operating grants shall b
repayment to the county agency must be in
repeal
repeal
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
repealed
365571: 1

C:\...\LexisNexis\Data\Prospective\all-output.txt
2011-11-17 4:38:30 AM   5,749,165 bytes   <default>   ANSI   U
repayment must be made within five workin
repayment of the operating grants shall b
repayment to the county agency must be in
repeal
repeal
646850: 1

repealed¶
repealed¶

148 difference section(s)   Same   Insert   Load time: 0 seconds
```

Lost? Or converted into attributes?

Catch lost and duplicated content

QC: Autogeneration of word-wheels

- Highlights anomalous data
- Very quick for a human to scan
- Out-of-place data jumps out

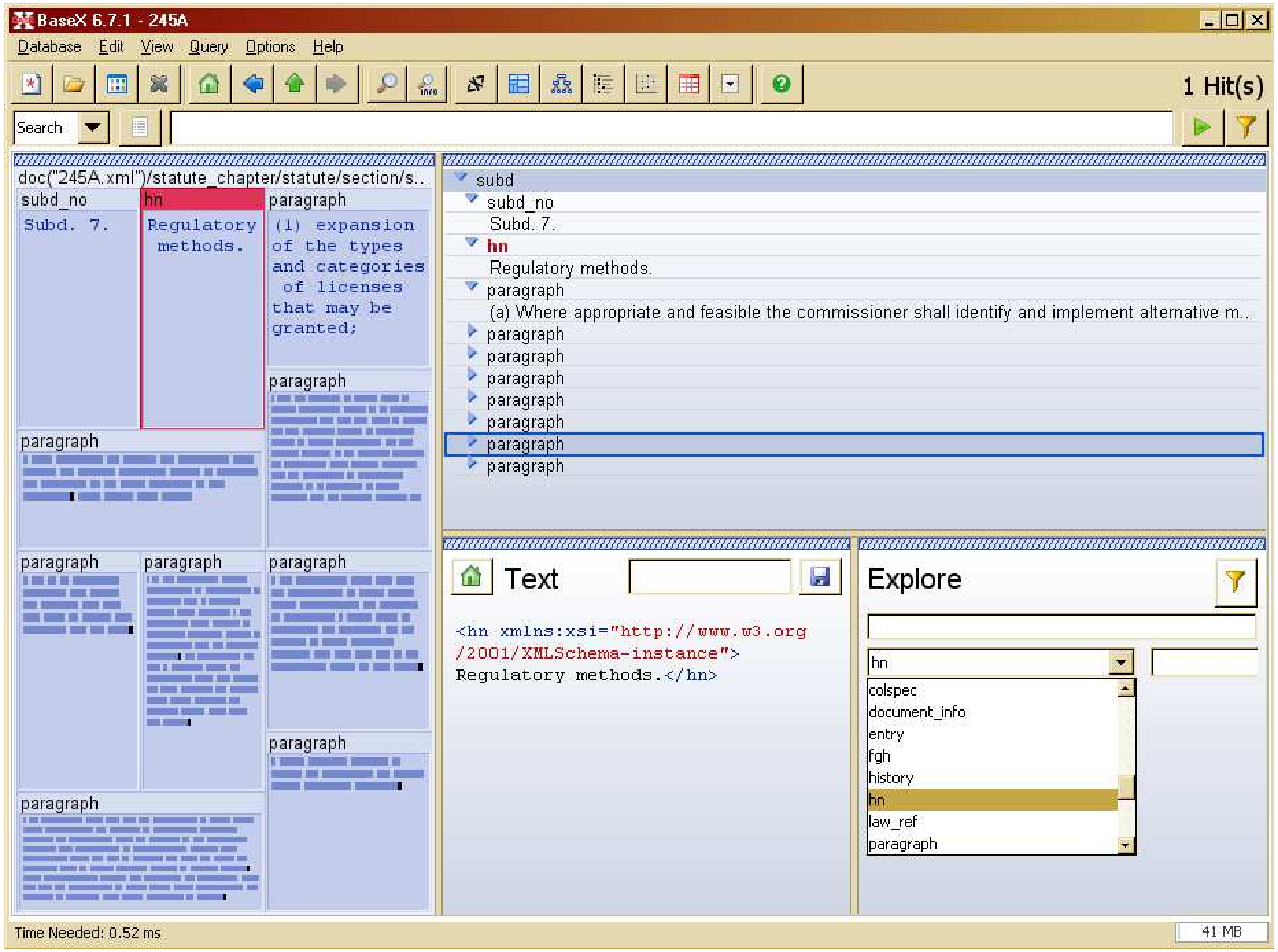
Makes human QC efficient

Why is this one so long?
Oh! It's two run together.

```
1995 c 263 s 8
1995 c 263 s 9
1995 c 264 art 6 s 4,5
1995 c 61 s 1-7
1995 c 82 s 1
1995 c 82 s 10,16
1995 c 82 s 11
1995 c 82 s 2-9
1995 c 86 s 1
1996 c 277 s 1
1996 c 281 s 1
1996 c 305 art 1 s 57
1996 c 305 art 2 s 44
1996 c 305 art 2 s 47
1996 c 305 art 2 s 48
1996 c 305 art 2 s 49,50
1996 c 305 art 2 s 51
1996 c 305 art 2 s 52
1996 c 305 art 2 s 53
1996 c 305 art 2 s 54
1996 c 312 s 1
1996 c 392 s 11996 c 392 s 2
1996 c 392 s 3
1996 c 392 s 4
1996 c 392 s 5
1996 c 392 s 6
1996 c 395 s 11
1996 c 395 s 12
1996 c 398 s 58
1996 c 408 art 10 s 5
1996 c 408 art 10 s 6
1996 c 414 art 1 s 35
1996 c 416 s 1
1996 c 416 s 10
1996 c 416 s 11,12
1996 c 416 s 13
1996 c 416 s 9
1996 c 421 s 1
```

QC: Use XML database with XQuery

- Quickly finds examples of any pattern



doc("245A.xml")/statute_chapter/statute/section/s..

subd_no	hn	paragraph
Subd. 7.	Regulatory methods.	(1) expansion of the types and categories of licenses that may be granted;

```
subd
  subd_no
   Subd. 7.
  hn
   Regulatory methods.
  paragraph
   (a) Where appropriate and feasible the commissioner shall identify and implement alternative m..
  paragraph
  paragraph
  paragraph
  paragraph
  paragraph
  paragraph
  paragraph
```

```
<hn xmlns:xsi="http://www.w3.org
/2001/XMLSchema-instance">
Regulatory methods.</hn>
```

hn

colspec
document_info
entry
fgh
history
hn
law_ref
paragraph

Time Needed: 0.52 ms

Tangible Benefits

- **Objective, repeatable and reliable estimation**
  from our conversion function point framework

- **High quality results**
  from programming best practices

- **High productivity**
  from sophisticated tools and techniques

- **Scalability**
  from systematic development approach