Between close and distant:

historical editing methods at intermediate scale

Douglas Knox, Newberry Library, Chicago, Illinois

What to do with...

Valued literary texts?
Mark them up intensively one by one

Millions of books?
Scan en masse, OCR, text mine

Tens of thousands of texts?

Transcribe, organize, edit
in domain-appropriate ways

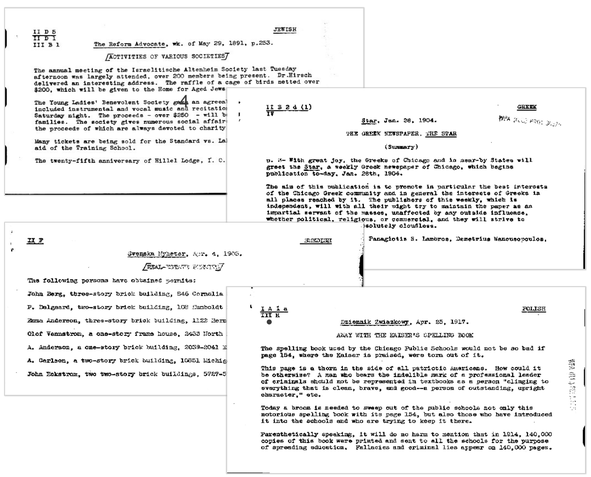
"Give us editors!" —Gregory Crane

Case Study:
The Chicago Foreign Language Press Survey

- A project of the Works Progress Administration, 1935-39
- 120,000 typescript sheets
- ca. 50,000 articles selected and translated from Chicago's foreign-language press, 1861-1938
- 22 ethnic groups

Group	Sheets
German	18,448
Polish	16,368
Jewish	16,298
Bohemian	15,811
Greek	10,706
Norwegian	7,654
Swedish	6,780
Russian	5,963
Lithuanian	5,950
Danish	3,847
Italian	2,950

Group	Sheets
Hungarian	2,688
Spanish	1,909
Croatian	1,321
Ukrainian	997
Dutch	795
Filipino	588
Slovak	509
Chinese	398
Slovene	197
Serbian	124
Albanian	91



Digital Humanities 2011 Stanford University poster session, updated July 2011

Validate against vendor schema

```
vendor/5420780_2/5420780_2_1808.xml

<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="flps0.1.rnc" type="compact"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader><fileDesc><titleStmt>
    <title/>
   </titleStmt>
   <publicationStmt>
    <date>Mar 05, 2010</date>
    <idno>5420780_2_1808</idno>
   </publicationStmt>
  </fileDesc>
 </teiHeader>
 <text>
  <front>
   <pb facs="5420780_2_1808.jpg" n="1"/>
   <div type="group">
    <list>
     <item>DANISH</item>
    </list>
   </div>
   <div type="codes">
    <list>
     <item>II D 10</item>
    </list>
   </div>
   <div type="citation">
    <bibl>
     <title>Revyen,</title>
     <date>Apr. 19, 1913.</date>
     <title level="a">BENEFIT CONCERT</title>
    </bibl>
   </div>
  </front>
  <body>
   <p>The Society Dania is giving a concert on Sunday, May
11, in Dania Hall, for the benefit of Danish victims of the
recent storm and flood in Omaha, Nebraska.</p>
  </body>
 </text>
</TEI>
```

Check that filename, file identifier, and image name match as expected.

XSLT: Copy and constrain Group and subject code designations. The new schema requires these to match a fixed set of values enumerated in a taxonomy.

XSLT: Copy article information to header, automatically normalizing case and regularizing dates where possible.

Validate against editing schema

```
flps/5420780_2/5420780_2_1808.xml

<?xml version="1.0" encoding="UTF-8"?>
<?oxygen RNGSchema="flps0.2.rnc" type="compact"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
 <teiHeader><fileDesc><titleStmt><title>
     <bibl><title>Revyen</title>, <date when="1913-04-19">Apr. 19,
1913</date>. <title level="a">Benefit Concert</title></bibl>
    </title>
   </titleStmt>
   <publicationStmt>
    <idno>5420780_2_1808</idno>
   </publicationStmt>
  </fileDesc>
  <encodingDesc>
   <xi:include xmlns:xi="http://www.w3.org/2001/XInclude"
href="flpstaxonomy.xml">
    </xi:include>
  </encodingDesc>
  <profileDesc>
   <textClass>
    <catRef target="#grp-danish"/>
    <catRef target="#grp-danish #code-II.D.10"/>
   </textClass>
  </profileDesc>
 </teiHeader>
 <text>
  <front>
   <pb facs="5420780_2_1808.jpg" n="1"/>
   <div type="group">
    <list><item>DANISH</item></list>
   </div>
   <div type="codes">
    <list><item>II D 10</item></list>
   </div>
   <div type="citation">
    <bibl><title>Revyen</title>,
<date when="1913-04-19">Apr. 19, 1913</date>.
<title level="a">BENEFIT CONCERT</title></bibl>
   </div>
  </front>
  <body>
   <p>The Society Dania is giving a concert on Sunday, May 11, in Dania
Hall, for the benefit of Danish victims of the recent storm and flood in
Omaha, Nebraska.</p>
  </body>
 </text>
</TEI>
```

Report XPath expressions for a batch of files, and compare against paths already seen (to notice if vendor starts using the schema in a valid but new way):

```
 7927 TEI/text/body/p
 5272 TEI/text/front/div/@type
 5272 TEI/text/front/div
 5244 TEI/text/front/div/list/item
 3516 TEI/text/front/div/list
 3112 TEI/text/front/div/bibl/title
 1765 TEI/text/front/pb/@n
 1765 TEI/text/front/pb/@facs
 1765 TEI/text/front/pb
 1765 TEI/text/front
 1765 TEI/teiHeader/fileDesc/titleStmt/title
 1765 TEI/teiHeader/fileDesc/titleStmt
```

Explore data values in bulk, batch by batch, field by field, looking for anomalies.

Examine group values with XML Starlet and simple Bash shell scripting:

```
for i in *.xml;
do xml sel -N tei=http://www.tei-
c.org/ns/1.0
-t -m "//tei:div[@type='group']"
-v "tei:list/tei:item" $i;
done  | sort | uniq -c | sort -nr
```

Examining date values:

```
 6 May 23, 1919.
 3 May 1, 1919
 2 September 26, 1879.
 2 Sept. 31, 1914.
 2 Sept. 25, 1893.
 2 October 8, 1913.
 2 Nov. 24, 1920.
 2 Mar. 11, 1914.
 2 1913-14. p. 243.
 1 wk. of Sept. 26, 1891.
 1 Wk. of March 7, 1903. p.100.
 1 Wk. of March 2, 1923. Vol. 4, p.4.
 1 Week of February 9, 1929, Volume
77, Page 35.
 1 Volume 5. Week of July 20, 1923,
Page 4.
 1 June 26, 1877, 4:1.
 1 8 - 3 - 19
 1 8-1-19
 1 7/21- 1916.
 1 1915-16. p. 357.
```

Examining title values:

```
436 Denni Hlasatel,
 84 Svornost,
 76 Denní Hlasatel,
 10 The Chicago Tribune,
  7 SVORNOST,
  4 Illinois Staats-Zeitung,
  3 The Chicago Daily Tribune,
  2 Svornost, Chicago,
  2 Chicago Tribune,
  1 The Illinois Staats-Zeitung.
  1 The Denni Hlasatel,
  1 Chicago Tribune, Vol. XLVI,
  1 Svornost, Vol. III, No. 224
  1 Svornost
  1 Illinois Staats - Zeitung,
  1 Dennií Hlasatel,
  1 Denni Hlasetel,
  1 Denni Hlasatel
  1 Denní Hlasatal,
  1 Denni Hlasatal,
  1 Chicago Daily Tribune,
```

Examining group values:

```
593 LITHUANIAN
  1 NORWEGIAN
  1 LUTHUANIAN
  1 LITHUMANIAN
  1 LITHUANLAN
  1 LITHUANIAN (1)
```

Seeing editing issues at this scale requires semi-automated tools guided by editor-directed questions about the material.

Test-driven editing

Apply automated tests with human exception handling, followed by automated correction and testing or hand-edited correction, as appropriate.

Validation:

Many formerly valid files became invalid after transformation to the editing schema, which restricted values for ethnic groups and codes. The vendor's task was to accurately document what appeared on the sheets. The editing task involved checking the work of the vendor, as well as making sense of and correcting the work of the 1930s editors.

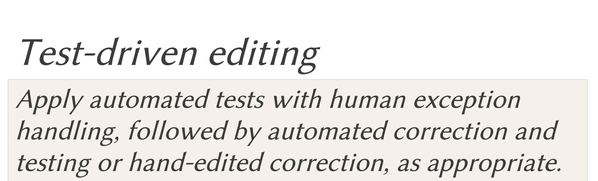
Preparation for testing:

- record <idno> element values
- record @facs attribute values (sheet images)

Tests:

- check that IDs have expected form
- check for missing image references
- check for duplicate references & IDs
- check that article ID matches first image ID
- check for continuity in sheet @n values

Mind the <gap>



5423967_2_1109.xml

```
   <p>The concert of the Russian pianist Mr. Joseph
Levin, was extraordinarily successful. The artist
performed many compositions of the great masters,
such as Schumman, Beethoven, Chopin, and many
others. A numerous and<gap unit="chars" extent="unk"
reason="illegible"/> rewarded every number of the
artist, with great ovations and enthusiasm. It was
noticed in many American papers, that the concert of
the artist Mr. Joseph Levin, was exceptionally
successful.</p>
```

The <gap> element above marks the limits of a bulk transcription phase of the project, and becomes a potential task list for more labor-intensive editing work.

Missing or uncertain information presents new challenges at intermediate scale. Matching the scale of labor to the expected benefits is an act of editorial judgment.

The 1930s editors were managing a database of tens of thousands of records with nothing more than typewriters, paper, and human attention. Digital tools would be new to them, but this scale of material would not. This project is working at a scale the Press Survey editors chose 70 years ago.

This poster represents work from a Newberry Library project funded by a Humanities Collections and Reference Resources grant from the National Endowment for the Humanities.
Any views, findings, conclusions, or recommendations do not necessarily represent those of the National Endowment for the Humanities.

Douglas Knox, Newberry Library, 60 West Walton Street, Chicago, Illinois 60626, knoxd@newberry.org