FDLP, Cornell, and the CFR

2010-10-18

The CFR

- What it is: the Federal Register and the CFR
- What it is: the data

How we met

- The background: two data sets
  - Oldskool data: locator code
  - Newstyle: XML from FD/SYS
- The exchange: data for expertise

Why Cornell and the LII?

- LII history
- LII staffing and expertise
- LII relationships and communities

LII US Code experience

- 1994: first edition based on ASCII from the Office of the Law Revision Counsel
- 2000: first XML edition based on "locator code" -- the same format that FDLP wanted to make available for CFR
- 2010: US Code is the most popular LII collection
- CFR will be at least that popular

CFR: consistent with LII mission and research interests

- CFR as a target for open access
- Builds on LII work with:
    - administrative law data
    - ABA e-rulemaking committee
- Resonates with allied work on notice and comment rulemaking (CeRI)
- Holds strong technical interest

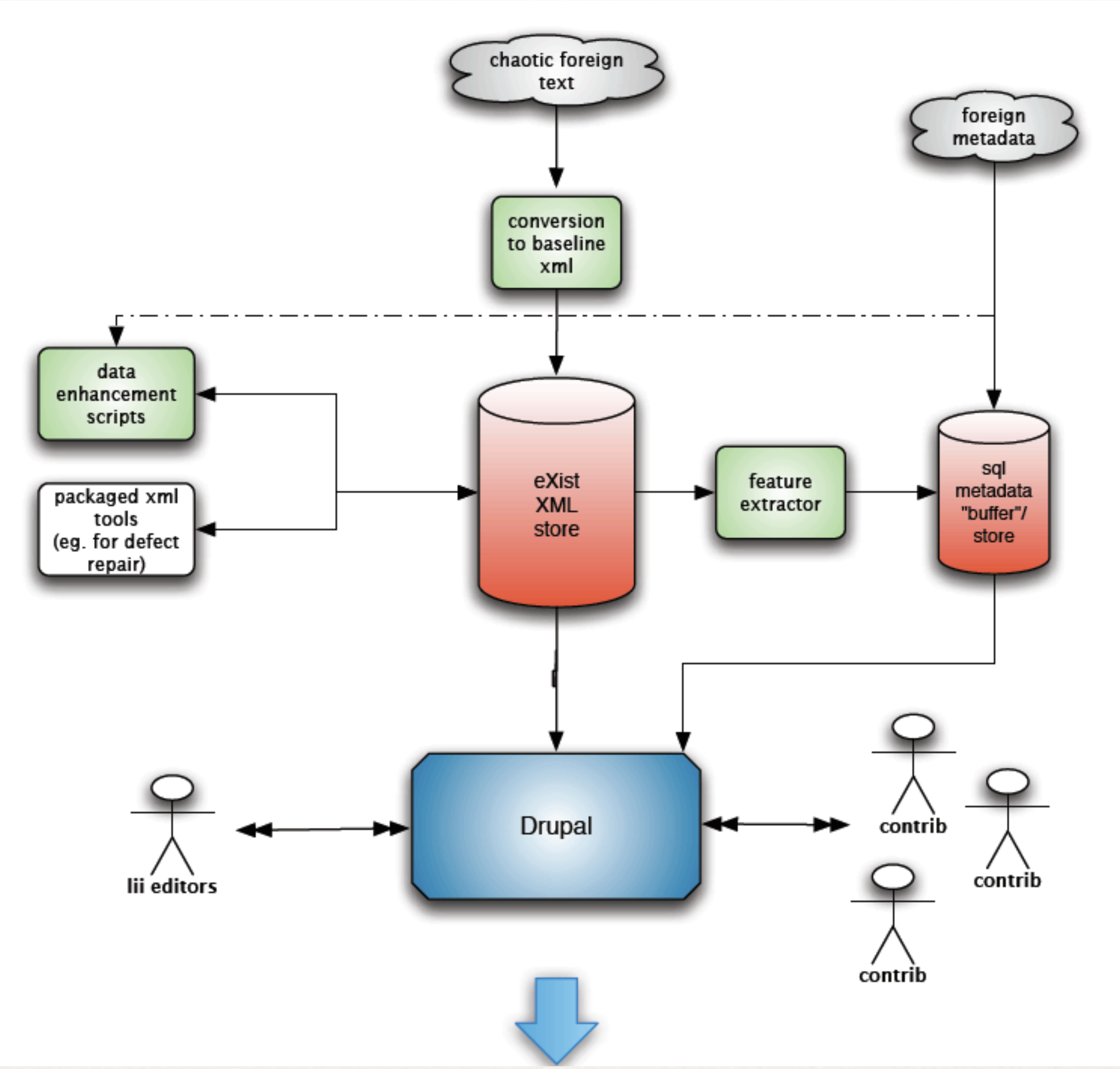
Sometimes it's good not to be the government

- Web 2.0 poses problems for providers of official data.
- Need for authoritative information is at odds with techniques like crowdsourcing.
- Outsiders better able to low-cost experiments with high failure rates.
- In short, we can do a bad job and then improve it.

The CFR as dataset : challenges

- CFR is very, very big
- CFR comes from many, many sources
- CFR has long-tailed problems
  - Inconsistent labels, textual structures, approaches
  - Problems of verification and certainty
- Many eyes might be the cure

LII delivery architecture



Current features of the “alpha” version

- Supersection structure
- Subsection structure
- Cross-references
- Parallel Table of Authorities

Development roadmap: next release

- Better handling of “deviant” part numbers
- Federal Register logbook
- Solr search (offers faceted search, “more like this” features)

Development roadmap: next next release

- Updating features
- Crowdsourced links
- Formal feedback and correction mechanisms

Pie at higher altitudes

- Taxonomies
- Representation of related documents
- Empirical work
- Data-management practices

Taxonomies & thesauri

- SKOS representations of FR subject heads; NAICS product codes
- Application of EUROVOC and AGROVOC taxonomies
- Applications:
  - Query expansion
  - Navigation aids
  - Cross-jurisdictional retrieval

Representation of related documents

- PTOA makes a good example
- Unavailable as XML
- Not as rich as it might be
- Useful test case showing the tension between formal systems and practical approaches

Empirical work

- Census
- Fault detection: errors, inconsistencies, and noise
- Analysis of different classification approaches and their effectiveness as finding aids
- Usage patterns and user needs

Data management

- What happens when the official source has related collections or augmentations?
- Work with layered architectures that offer different, remixable perspectives on the data

In conclusion...

- For us, an excellent opportunity to learn from those who know the most about the data

- For all, a chance to learn about the needs of different communities.