Toward Automatic Metadata Assignment in the Texas A&M University Digital Repository

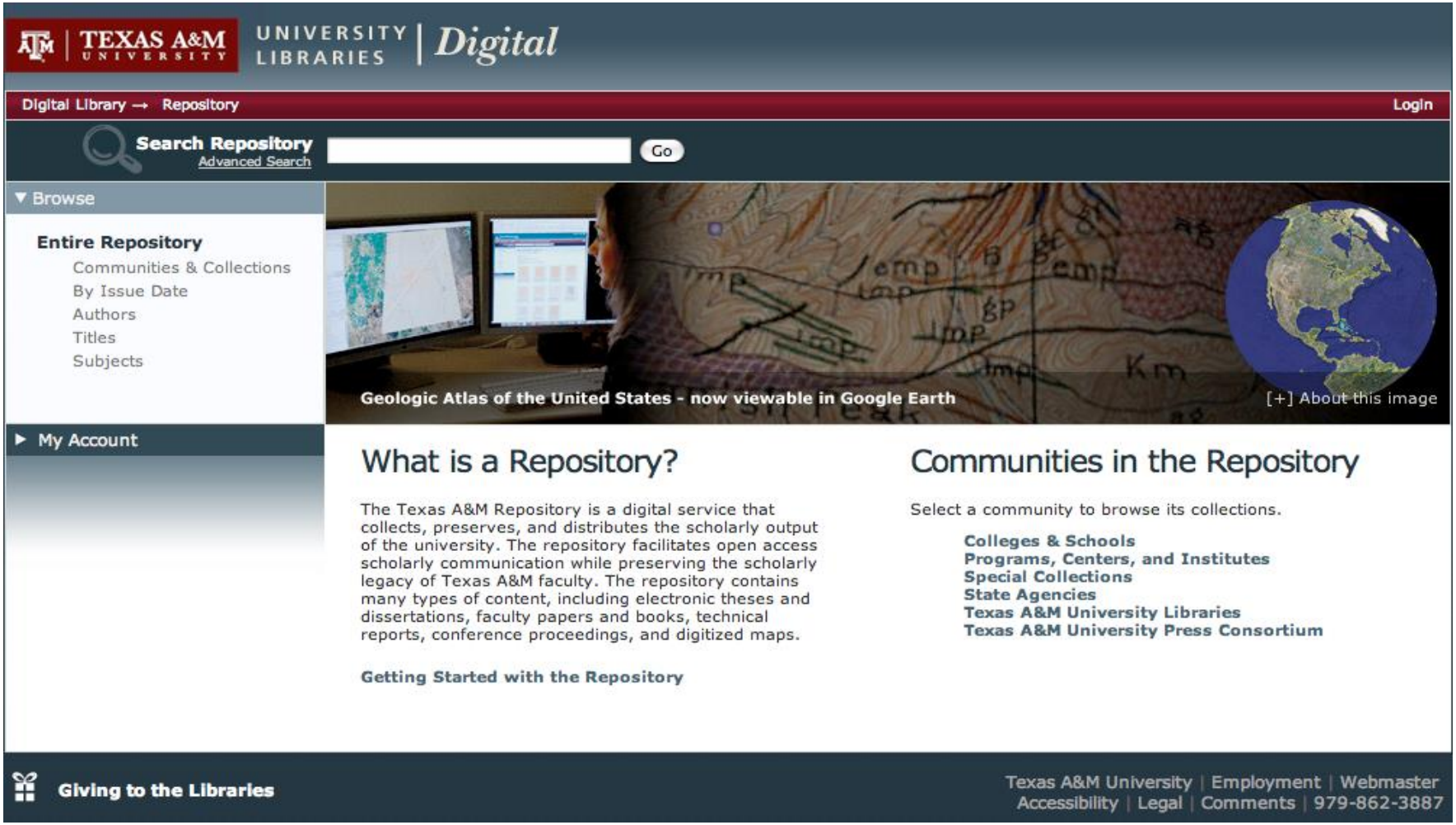
James Creel, Alexey Maslov, Adam Mikeal, Scott Phillips, and John Leggett

Texas Conference on Digital Libraries
May 27, 2009

Overview

- *The Texas A&M Digital Repository*
- *Theoretical Consideration of Metadata*
- *Technical Consideration of DSpace*
- *Research Status*
- *Future Work*

Texas A&M Digital Repository



Theoretical Consideration of Metadata

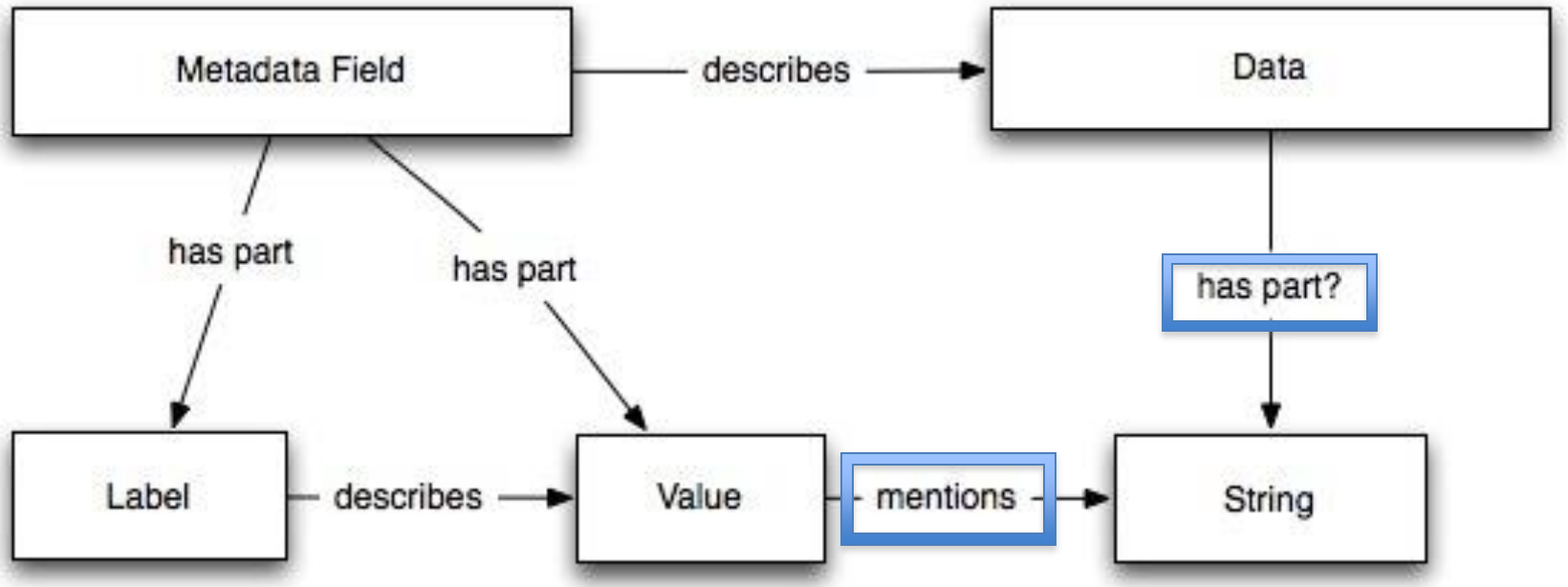
- *Use/Mention distinction*
- *Semantic Structure of Metadata*
- *Metadata Assignment*

Use/Mention Distinction:

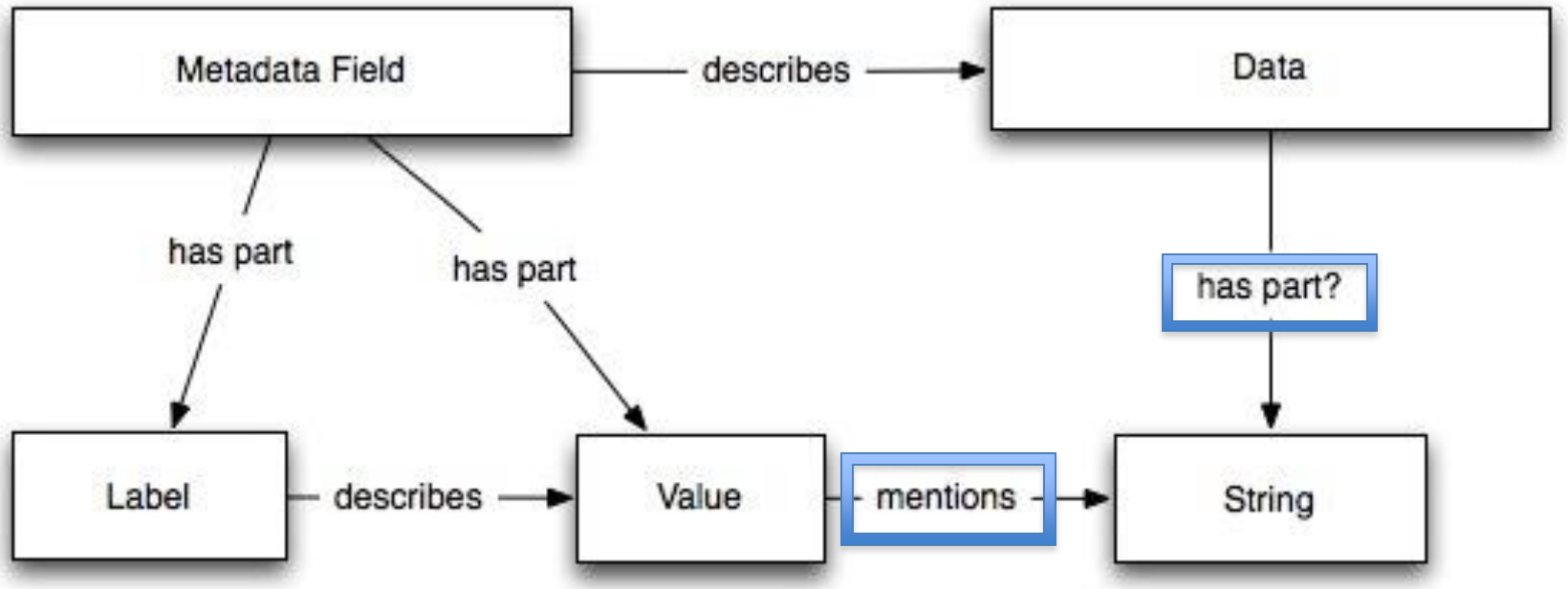
Examples of use and mention of the word “tough”

- Ex. of word use
  - That Ben Bernanke is a real tough guy.
- Exs. of word mention
  - Tough may refer to both the character of a man and the texture of a steak.
  - Tough is a 5-letter word.
- May 12 (Bloomberg) -- Asian stocks slid from a seven-month high, led by banks and mining companies, as HSBC Holdings Plc said 2009 will be a “tough” year and metal prices fell.

Semantic Structure of a Metadata Field



Extraction vs. Generation



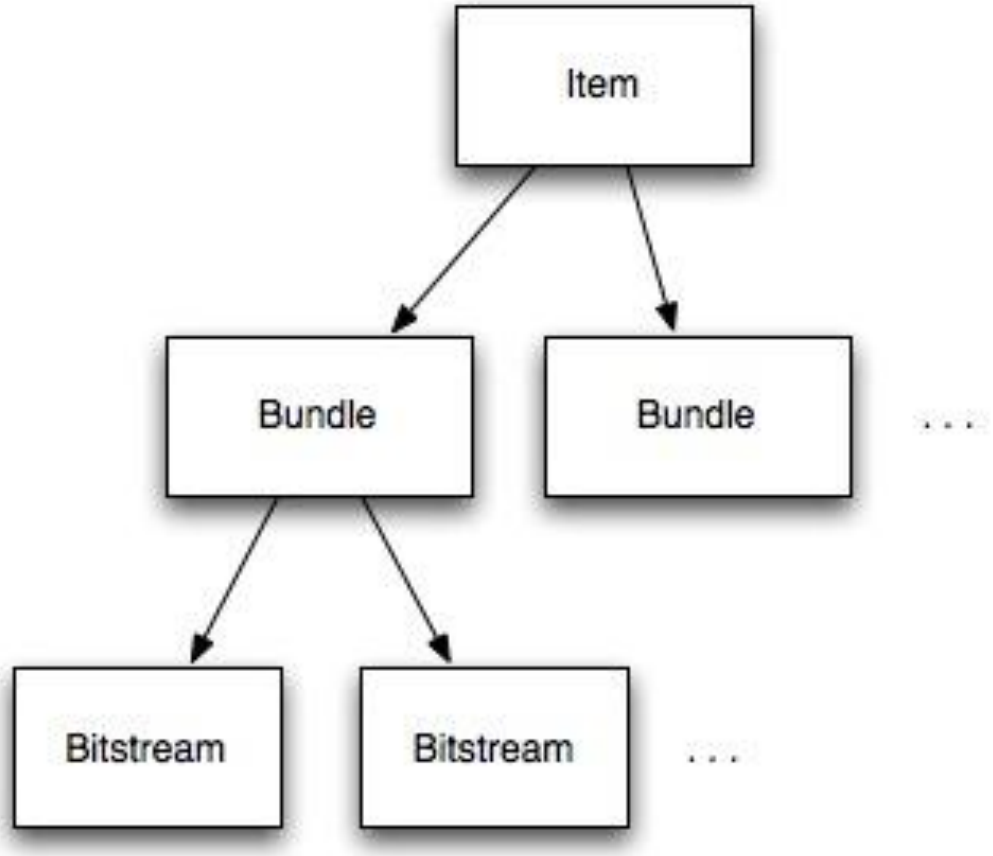
Q: Does an assignment entail metadata extraction or metadata generation?

A: It depends on the relation of the Data to the String mentioned by the Value of the Metadata Field.

Technical Consideration of DSpace

- *Where do the metadata go?*
- *How are metadata assigned?*
- *Item metadata*

Where do the metadata go?



DC-style metadata are applied only at the Item level.

Bundles have only a name.

Bistreams have name, description, and format.

Community/Collection hierarchy is another matter…

How are metadata assigned?

- Item metadata: the main concern.
- Bundle names: determined automatically for internal bookkeeping; hardly a consideration.
- Bitstream names: merely filenames.
- Bitstream format: using file extension, but UNIX file command would be better.
- Bistream descriptions: very challenging!

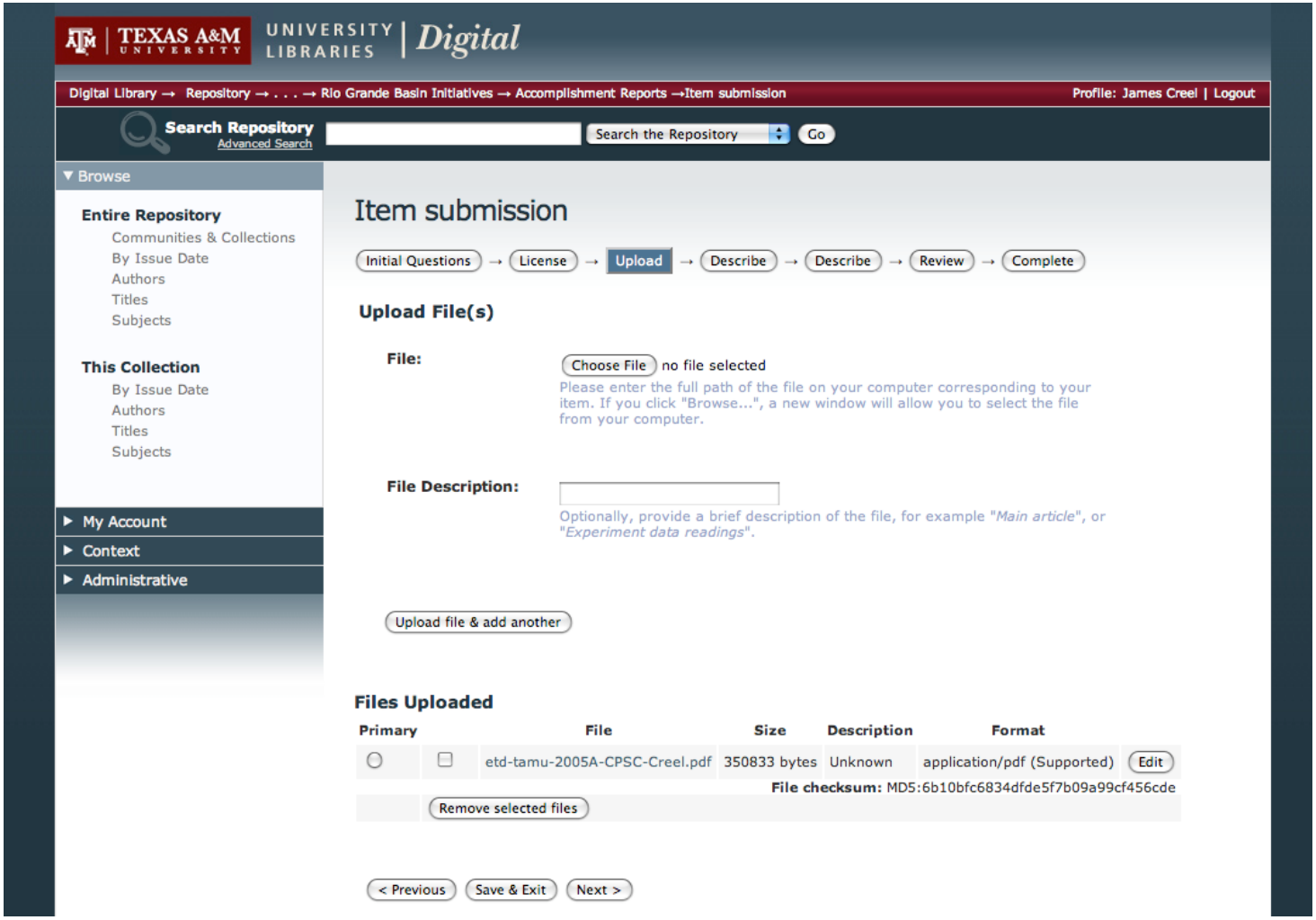
Item Metadata

- Item metadata field labels have three parts (in the style of Dublin Core)
  - Schema name (usually “dc”)
  - Element (Ex: “description”)
  - Qualifier (optional. Ex: “abstract”)
  - Ex: dc.description.abstract
- No built-in support for deeper hierarchy

Research Status

- *Manakin Submission Workflow*
- *Syntactic Extraction*
- *Statistical Extraction*

Manakin Submission Workflow



Syntactic Extraction

Extracting predefined fields from TAMU ETDs

- Title
- Author
- Abstract
- Committee members
- Degree level
- Subject area

Subject Extraction

- Current focus
- Obtains subject keywords from text
- Maximum entropy
  - a model for training computers to classify data
  - Case in point – named entity recognition (visit openNLP at http://opennlp.sourceforge.net/)

Name Disambiguation

- Future focus
- Disambiguate homonyms
- Related work: Author name disambiguation is addressed using Bayesian probabilistic models of text (Efficient topic-based unsupervised name disambiguation. Yang Song et al., JCDL 2007)

Future Work

- Support for deeper hierarchies of metadata labels
  - Create sub-structure for local qualifiers?
  - Semantic graph based item browser
- Long-term
  - Citation metadata
  - NLP+KB for deeper metadata generation and NL interfaces