solrizer

pragmatically connecting search, management and indexing in a repository solution

Matt Zumwalt
Open Repositories 2010
Madrid, España

solrizer

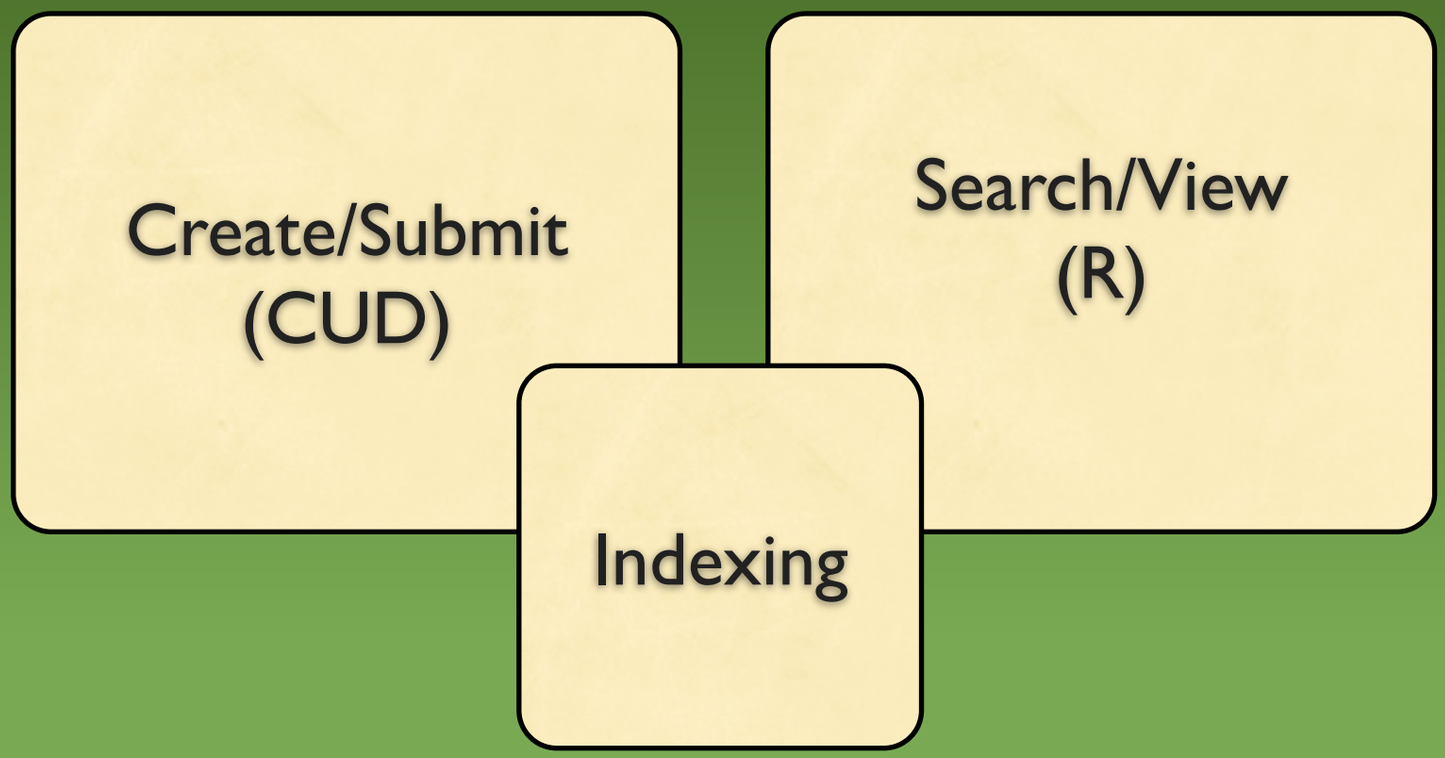
define indexing logic in scripting language (ruby)

allow indexing approach to evolve continually (iterative development)

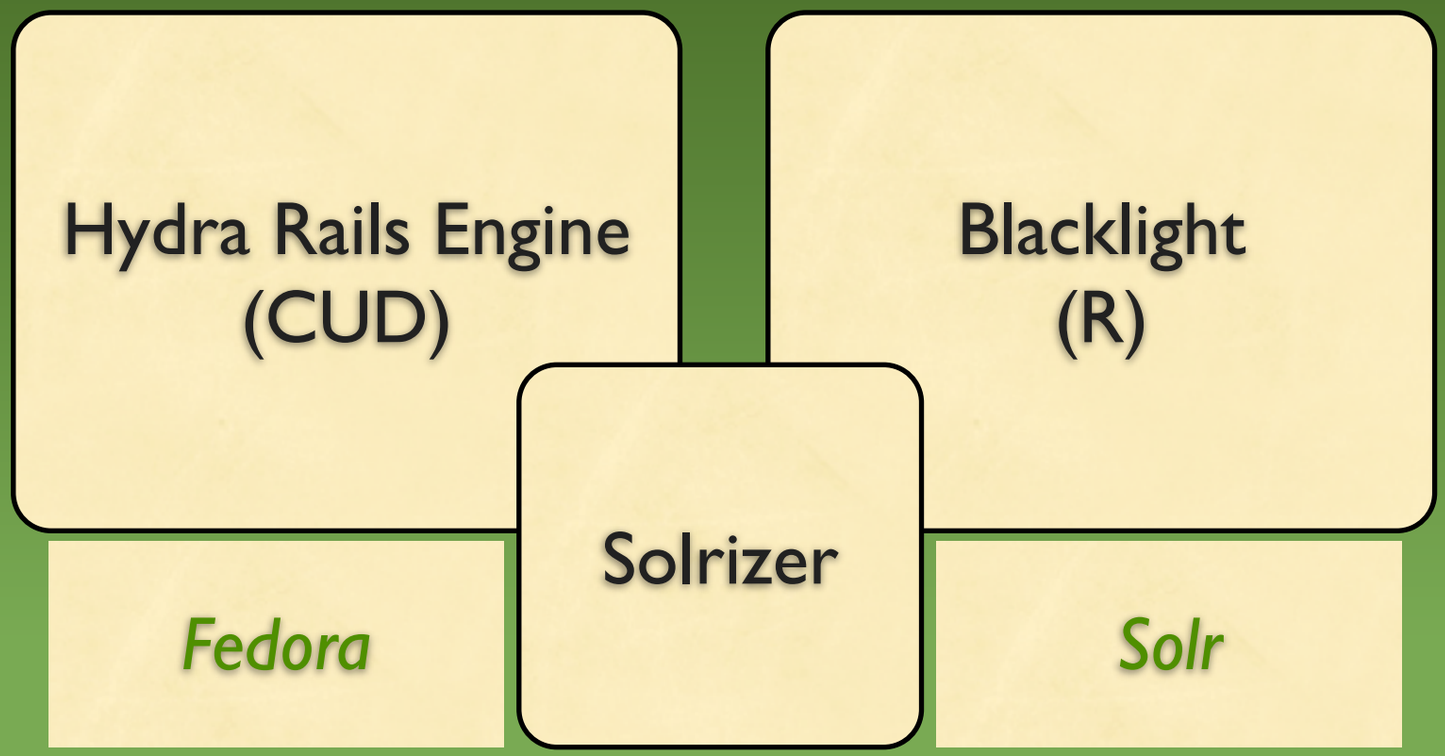
track indexing algorithms within models (DRY)

Bonus: new paradigm for mapping xml to application vocabularies!

CRUD in Repositories



CRUD in Hydra Heads



methodologies for indexing

RDBMS

RDBMS + solr

fedora + gsearch + solr

fedora + ~~shelver~~ + solr

fedora + solrizer + solr

RDBMS alone

- search index bound to data model
- no full-text indexing

RDBMS + solr/lucene

+ fulltext index

+ index is separate from data model

- tools often underestimate conceptual differences inherent in solr-style search

fedora + gsearch + solr

+ index full text and xml metadata!

+ freestanding tool specifically for indexing

+ created & maintained by Gert Pedersen

- XSLT (transformations vs. logical processing)

- lucene-oriented rather than solr-oriented

fedora + ~~shelver~~ + solr

(this is what we did in SALT)

+ ruby based

+ freedom to iteratively refine indexing logic

- code sprawled -- fails DRY principle

http://bit.ly/bJr4O3

http://github.com/sul-dlss/salt/blob/master/lib/shelver/indexer.rb

fedora + solrizer + solr

+ freedom to iteratively refine indexing logic

+ models define their own indexing (DRY)

+ could index anything - not just fedora objects

http://github.com/mediashelf/shelver

solrization process

1. retrieve object by pid (unless an object was already passed in)
2. look up all *known models* for the object
3. with each model, load the object & call to_solr
4. save cumulative document to solr

what's currently fedora-specific

index_objects assumes you want to search in fedora for objects to ingest

the lookup for *known models* assumes fedora objects & active-fedora models

.. both of these could be refactored to be generic.

Bonus: Opinionated Metadata (OM)

allows you to declare mappings between (arbitrary) xml and your application vocabulary

active-fedora uses these to provide a default indexing algorithm (*zero code*)

empower users, free your content

http://mediashelf.us
http://mediashelf.eu
http://yourmediashelf.com/blog