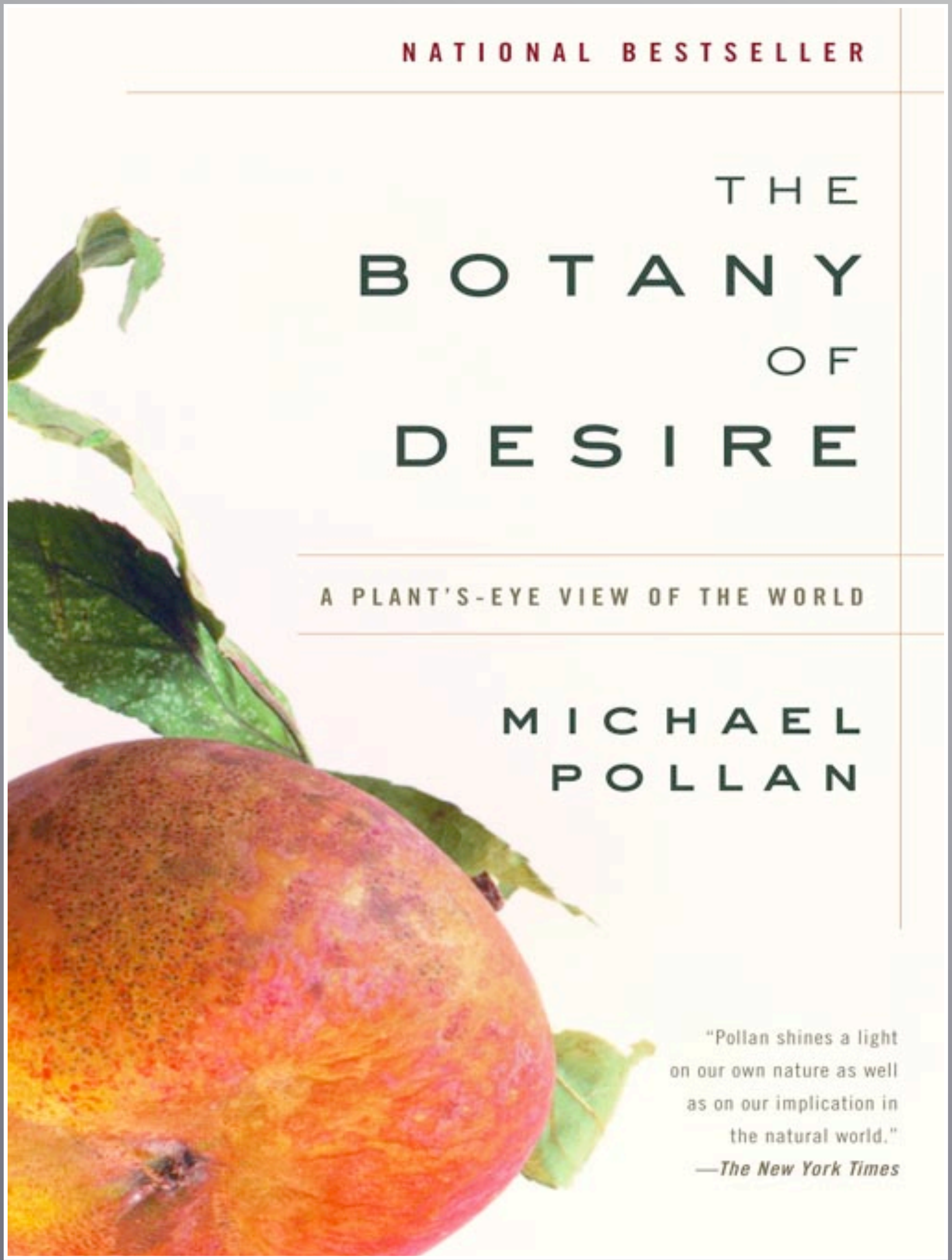
Xapian and Search Integration

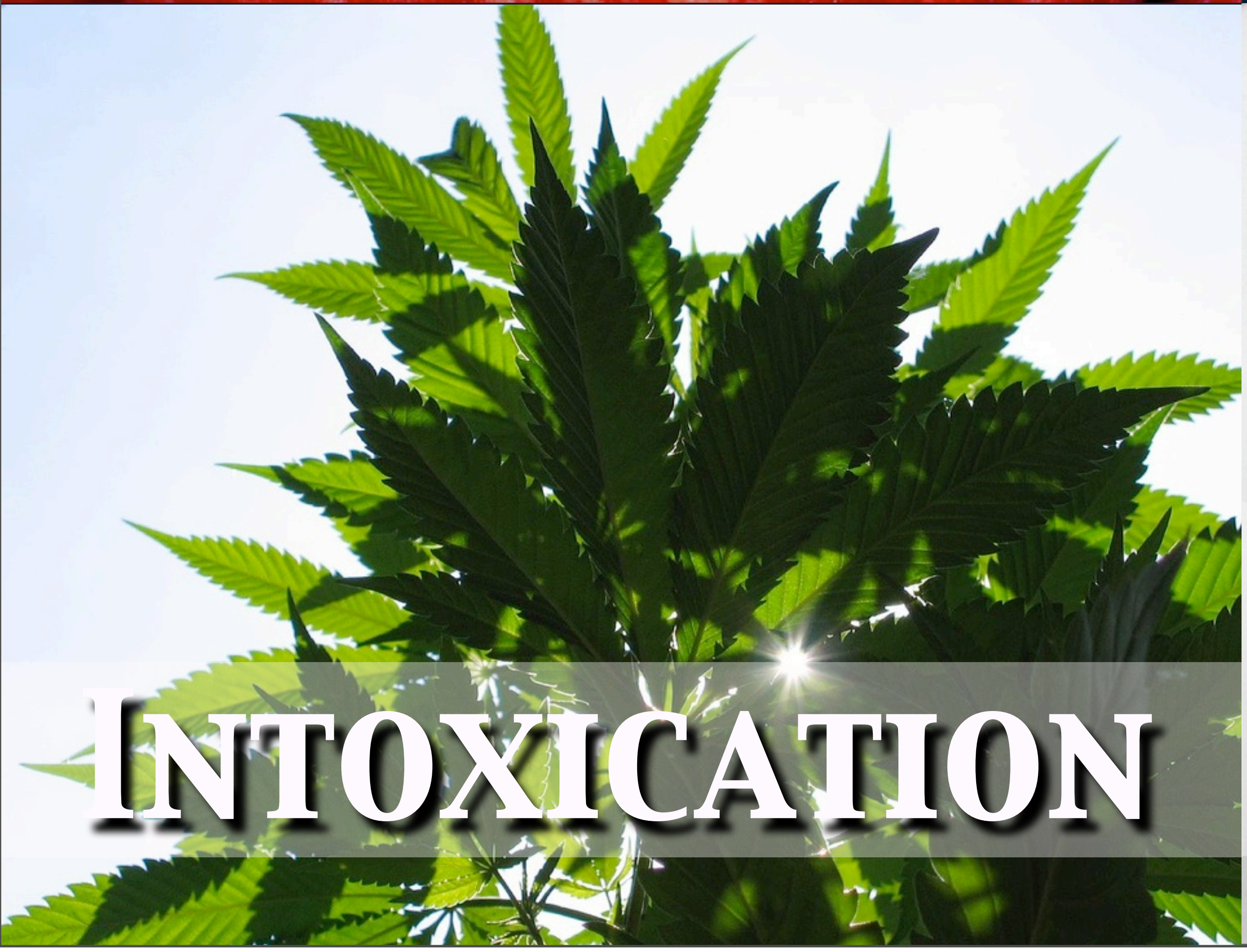
Sidnei da Silva
Enfold Systems, Inc.


NATIONAL BESTSELLER
THE BOTANY OF DESIRE
A PLANT'S-EYE VIEW OF THE WORLD
MICHAEL POLLAN
"Pollan shines a light on our own nature as well as on our implication in the natural world."
—The New York Times

Start with a short story about the book. Don't talk about the fruits yet.
- We drove to NOLA listening to this in AudioBook form.

Sweetness

Beauty


INTOXICATION

Control

The author Michael Pollan explores the nature of domesticated plants from the dual perspective of the plants and humans.

Proposes an interesting question: did we domesticate the plants, or did the plants domesticate us so that they could spread around the world?

- Initially plants were not attractive, but they developed traits to attract us. Apple & Johnny Appleseed.
- The Oak is not interested in attracting our attention, but it's interested in attracting squirrels.
- The Oak in fact tries to be unattractive to anyone but squirrels.

Plone tries to attract us, just like the apple. It's master plan is to spread and proliferate all over the place. Paul Everitt says that "Plone is Sexy".

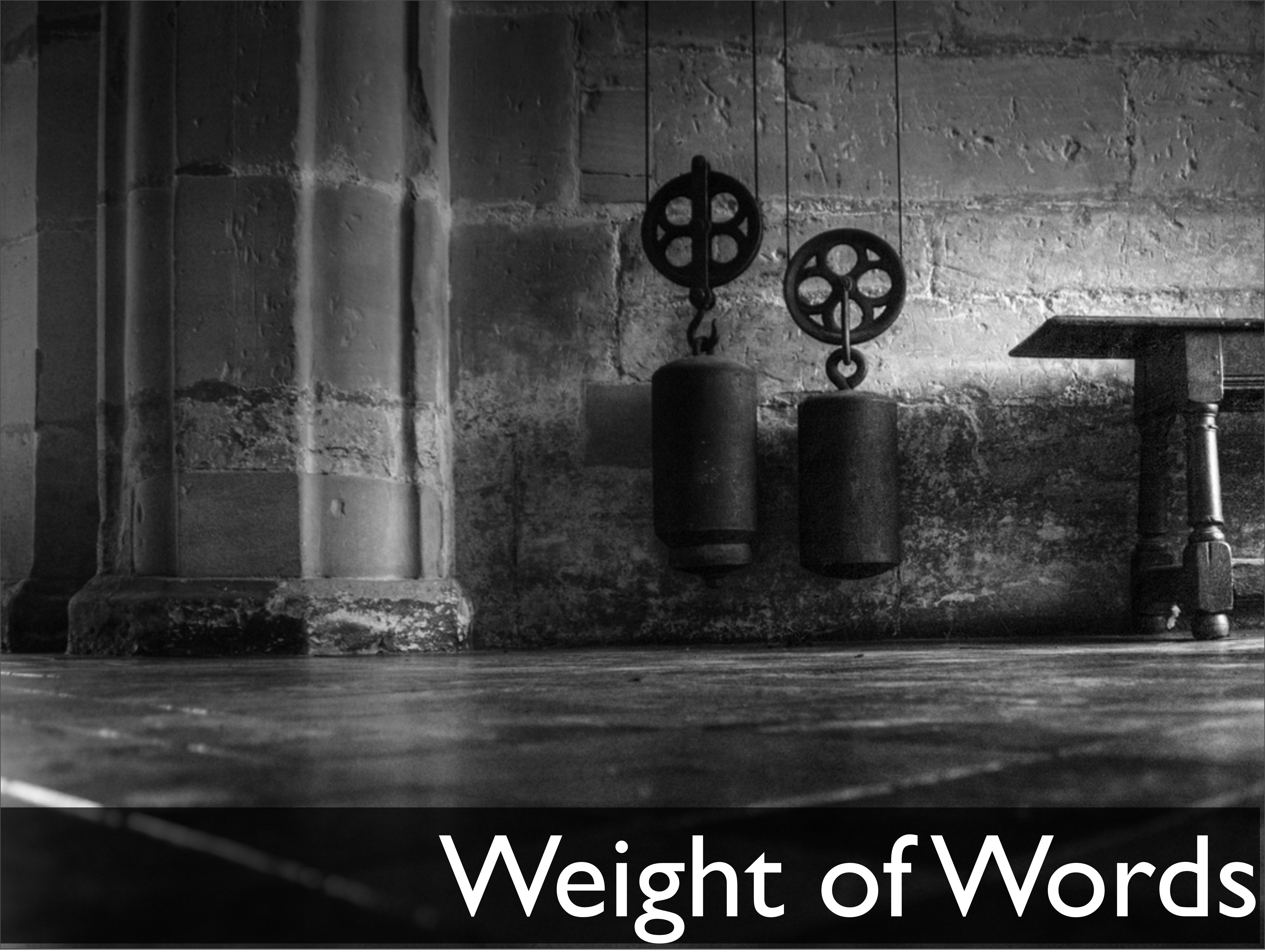
Hidden Cyanide

The apple seed actually contains a little bit of cyanide, which keeps the squirrels away from it because it has a sour taste.

- Something unexpected, leaves a bad taste in your mouth.
- Plone has hidden traits that can leave a bad taste in your mouth. You might not feel when you’re scratching just the surface, but as you get past that and reach into the core there are “cyanide filled seeds” there.
- One of the things that people stumble upon once they get past a certain amount of content is searches tend to get slower.
- Another thing is computing relevancy (quality of the results).

Speed

Enters Xapian. Super fast search results.



Weight of Words

- Ranked probabilistic search - important words get more weight than unimportant words, so the most relevant documents are more likely to come near the top of the results list.
- Relevance feedback - given one or more documents, Xapian can suggest the most relevant index terms to expand a query, suggest related documents, categorize documents, etc.
- Weighting is configurable to a very fine-grained level (OSI project)

Through the Woods

On the OSI project, we've got a handful of features implemented in Xapian.
- Partial matching (incremental search)
- Exact matching (keywords)
- Custom sorting (sort on arbitrary data types by providing custom comparison function)
- Windows support for Remote Database Connection

We've tried to use Xapian as a Relational Database. It worked very well for our use-cases but could still be improved. Exact count for batching was a big deal.

Strong Ties

Using Xapian as a library is the easiest way to get started (enfold.xapian, pyxapian, xappy). However, it means you can’t easily scale to multiple processes.

- We are still working on a solution that can give us the best of both worlds.
- Text extraction is better done on a separate process. Instability issues (memory leaks, locking up).
- Text indexing could potentially be done out-of-process. Metadata indexing would happen right away.

Down the Road

- Out-of-process text extraction
- Out-of-process text indexing (metadata indexing in process)
- Memory-friendly, better metadata extraction
- Keeping queue on disk instead of in memory
- Event ‘reduction’/‘consolidation’

Look at collective.indexing for a good start. Still a lot of stuff ahead to be improved.
Look at collective.solr, based on enfold.solr, which was an improvement over enfold.xapian.

Credits

- Flickr Users:
  - markusschoepke
  - jasmic
  - danielstarrason
  - pipistrula
  - massenpunkt
  - thomashawk
  - dcdead
  - dropbear_au
  - lollyman
  - cdnphoto
  - monster
  - laserstars

- Book Review:
- http://www.nytimes.com/books/01/06/03/reviews/010603.03bilgert.html

Credits for the pictures go to Flickr users above.

So long, and thanks for all the fish!

More information:

http://www.enfoldsystems.com

http://www.enfoldsystems.com/solutions/search

http://dev.plone.org/collective/browser/collective.solr

http://plone.org/products/collective.indexing

http://code.google.com/p/pyxapian

This picture is actually mine. It’s Alan’s preparing a Red Snapper. Taken with a cheap Lomography FishEye camera.