Metadata Abattoir

Mike Rylander
2016 Evergreen Conference
Friday, April 22
Indexing

HUH! What is it good for?

um ... well ... quite a lot, actually
Out of the box...

- MODS-based
- Various title fields
- Some author-ish fields, and other stuff
- ALL THE SUBJECTS
- One (count 'em, one) series field
- Identifier-ish fields
- The BLOB(TM)
Dipping a toe in the water

Make titles matter
Dipping a toe in the water

INSERT INTO config.metabib_field (field_class, name, label, weight, browse_field, xpath) VALUES ('keyword', 'titleproperkw', 'Title Proper kw', 25, FALSE, '//mods32:mods/mods32:titleNonfiling[mods32:title and not (@type)]');

Tech track, code slide required.
Say, that's great!

Now we will kill the BLOB!!!!!
Faceting for a better tomorrow

Harry, why do you have a tail?
Faceting for a better tomorrow

Here, let me trim that for you...

```
INSERT INTO config.index_normalizer (name, func, param_count)
    VALUES ('Replace by regular expression', 'regexp_replace', 2);
```

```
INSERT INTO config.metabib_field_index_norm_map (field, norm, params, pos)
    VALUES (1, CURRVAL('config.index_normalizer_id_seq'), '[" ;.*",""'], -1);
```

Tech track, code slide required.
The future is ... soon, I hope

Here's a graph full of dots:

http://blog.pgaddict.com/posts/performance-since-postgresql-7-4-to-9-4-fulltext
The future is ... soon, I hope

Ahhhh, that's better!

http://blog.pgaddict.com/posts/performance-since-postgresql-7-4-to-9-4-fulltext
Elbow grease required

- **Similarity search**
  - Available now, but tests full indexed string
  - Good for short tags; subject or short-title matching

- **Word similarity**
  - In 9.6, so, the future
  - Finds best match for substrings based on the query
  - Good for any indexed string length!

- **FTS Phrase search**
  - In 9.6 ... more future
  - Allows one ts_query for phrase matching, much faster than today's method
  - Less complex search == faster search
Let's slay superpages!

- Superpage pros:
  - Searches start faster today
- Superpage cons:
  - Small datasets can hide in the weeds
  - Hit counts start as an estimate for large result sets
Thanks! Now, let's chat.

Mike Rylander
2016 Evergreen Conference
Friday, April 22