The Anatomy of a Large-Scale Human-Computation Engine

Shailesh Kochhar, Stefano Mazzocchi, Praveen Paritosh

Freebase August Meetup
1: Freebase & Human Computation

2: Example – Stanford Library

3: RABJ

4: Consensus
Freebase
Structured database
12 MM entities, 300 MM triples/facts
Where does the data come from?
Community contributions
Mass Data Loads
Human Judgments Improve Both
Community
Simplify contribution through games
Is this an Ethnicity?

**Gogo people**

The Gogo (or Wagogo or Chigogo) are a Bantu ethnic and linguistic group based in the Dodoma Region of central Tanzania. In 1992 the Gogo population was estimated to number 1,300,000. The Gogo have historically been predominantly pastoralist and patrilineal (tracing descent and inheritance through the male line), but many contemporary Gogo now practice settled agriculture, have migrated to urban areas, or work on plantations throughout Tanzania. Gogo music has achieved an international reputation. Their name was invented sometime in the 19th century by the Nyamwezi caravans passing through the area while it was still frontier territory. Richard Francis Burton claimed a very small population for it, saying only that a person could walk for two weeks and find only scattered Tembes. There was and remains the problem of inadequate rain for crops and humans, the rainy season being short and erratic with frequent drought. In the 18th century the Wagogo were mostly pioneer colonists from

**Tajakant**

The Tajakant (also Tadjakant) is a Sahrawi-Moorish tribe of Arab-Berber origins. They speak Hassaniya Arabic. They traditionally lived in Western Sahara, Mauritania and Algeria. They are Muslims, belonging to the Malik school of Sunni Islam. Some sources claim they are
Community
Simplify contribution through games
Enable QA for Gridworks loads
The For Real Filmfare Best Movie Awards by jon

Aug 18, 2010 Freebase Meetup
Mass Data Loads

Precision: QA for >99% accuracy
### Is this book edition information correct?

**The Duelling Handbook, 1829 (Dover Books on History, Political and Social Science)**

*by Joseph Hamilton*

<table>
<thead>
<tr>
<th>Edition of</th>
<th>The Duelling Handbook, 1829</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pagination</td>
<td>167</td>
</tr>
<tr>
<td>Publisher</td>
<td>Dover Publications</td>
</tr>
<tr>
<td>Pubdate</td>
<td>2007</td>
</tr>
<tr>
<td>ISBN-13</td>
<td>9780486446489</td>
</tr>
<tr>
<td>Status</td>
<td>Created</td>
</tr>
</tbody>
</table>

### Instructions

Each box on the left represents a Book Edition topic in Freebase that was created or modified by the OpenLibrary book edition load. Please examine the data for any errors or inconsistencies and report them using the comment box. You should use the links provided to check this data against external sources. Here are some things to look out for:

- **Edition of** field should represent the book this is an edition of. It is OK for the book title to be slightly different than the edition title (i.e. missing a leading "The" or the full subtitle starting from a ":") as long as they are obviously the same book.
- When filled in, **Publisher** should be a Freebase topic for a publisher of books, and the right publisher for this edition.
- **Pagination** should be the correct number of pages for this edition.
- **Pubdate** should be valid for this edition.
- **ISBN-13** should be valid for this edition.

### Status: Created vs. Existed

Please pay special attention to the **Status** field, which indicates whether the load created the book edition or simply reconciled to an existing one. This is a potential source of error, and in the event of a bad reconciliation, you may see muddled data.
Mass Data Loads

Precision: QA for >99% accuracy

Coverage: Manual reconciliation
1: Freebase & Human Computation
2: Example – Stanford Library
3: RABJ
4: Consensus
Reconcile Stanford Library Catalog with freebase.com
Stanford Library Catalog

4.4MM book editions
1.3MM *English* book editions
1.2MM *English* books
600K authors
For freebase, identity is key
match books, match authors
Automatic matching insufficient
Trained judges needed to decide hard cases
How to get this?
RABJ
Redundant Array of Brains in a Jar
What?

Abstraction

Powers human judgment (HJ)

applications

3.1MM judgments in 16 months
Provides primitive elements for more sophisticated applications
Questions
Judgments
Queues
Agents
Design Constraints
Content-agnostic
Dynamic data
Low latency
Aug 18, 2010
Freebase Meetup

Architecture
Questions contain metadata, pointers to dynamic content

Questions added to queues

Metadata allows slicing and dicing
Acre applications pull questions from RABJ

RABJ matches judge to available tasks

Acre renders question, sends judgment back
Declarative consensus
Yes: 3, No: 3, Skip: 4, Invalid: 3, Max: 6

RABJ notifies agents when consensus is reached
2.3 MM questions
3.1 MM judgments
500+ queues
20+ applications
1: Freebase & Human Computation
2: Example – Stanford Library
3: RABJ
4: Consensus
Always have leftovers
Perfect Consensus? Not!

Aug 18, 2010

Freebase Meetup
Evaluating QAers
Explore
http://rabj.freebaseapps.com/explorer

Create

Reference
http://wiki.freebase.com/wiki/RABJ_API/
Questions?