# A KNOWLEDGE DATABASE

Lukas Nabergall

February 18, 2016

## 1 Principles

This knowledge database application is designed with the following principles in mind:

1. Should contain content which rapidly converges in time towards validated knowledge.
2. Should be relevant and useful to advanced research, development, and academia.
3. Should optimize searchability and quick content recognition and absorption.
4. Should be based upon open content submission, editing, and validation.

## 2 Users

The knowledge database will be completely open to all for viewing and contributing, but users will have the option to create an account and, for transparency, will be encouraged to provide their full name. Logged in users will be able to create new content, have the option to be denoted as authors of pieces of content that they have created or edited, and will be able to validate new edits on all pieces of content of which they are authors. Users who are not logged into an account, or who have opted to remain anonymous as editors of a piece of content, will not have the ability to submit new pieces of content or validate new edits.

## 3 Content

The knowledge database will be populated, primarily via open user submission, with brief entries explaining or documenting some single idea related to or relevant to advanced research and development fields, including mathematics, physics, chemistry, biology, statistics, computer science, programming, advanced technology and engineering, and other areas. Although the emphasis will likely at first be on content relevant to research and development in areas related to the hard sciences, the scope of the database could expand to include the social sciences, the humanities, the arts, and other disciplines, as well as more elementary content which is typically encountered at the secondary school and undergraduate levels, as appropriate.

All pieces of content will be of a single form containing the following items:

- Name — the primary name of the idea to be displayed in searches and lists, may contain LaTeX and other special formatting; e.g. *abc* conjecture, the $W$-trick, Equipartition theorem, Finite Field, etc.

- Alternative Names [Aliases] (optional) — alternate names (aliases) of the idea which may also be used for identification and search.

- Content Type — the type of the idea, e.g. idea, theorem, conjecture, lemma, equation, formula, inequality, code, algorithm, visualization, technique, proof, etc.

- Text — the actual explication/description/explanation of the idea, ~50 characters minimum[1], ~2000 characters maximum,[2] can contain LaTeX, TeX, code, rich text, HTML, and other special formatting;[3] e.g. "Given double occurrence words $w_1, w_2$, define the word distance $\delta$ by...".

- Images (optional) — related images, including all common file types (JPEG, PNG, TIFF, BMP, GIF, etc.).

- Keywords / Subject Designations — keywords and subject designations contained in or related to the idea, approximately 2 or 3 required; e.g. for a piece of content on finite fields, could include field, group, field theory, abstract algebra, associativity, operation, etc.

- Dependencies (optional) — references contained in the text to other pieces of content via their names or aliases, can be manually specified via special formatting and automatically generated[4] by matching words in the text with other pieces of content in the database, will be given hyperlinks on display; e.g. "Given double occurrence words $w_1, w_2$, define the word distance $\delta$ by...".

- Citations (optional) — Including both implicitly and explicitly (in-text) cited works, input with some standard academic citation format[5] (although including less information, e.g. only a url, is acceptable but discouraged), in-text citations are referenced in the text via special formatting.[6]

To prevent mass fraudulent content submissions, an authentication scheme (e.g. CAPTCHA[7]) will be used to verify that a human is indeed making the submission.

1. To be finalized.
2. To be finalized.
3. The input field will be a rich text editor containing buttons which automate much of the formatting process.
4. Possibly with user validation.
5. To be specified.
6. To be specified.
7. In the sequel, it is assumed that CAPTCHA will be used for authentication in the knowledge database.

### 3.1 Editing and Validation

All aspects of all pieces of content will be fully editable by any user of the knowledge database. To combat malicious, inaccurate, or inappropriate content submissions and edits, an essentially democratic validation scheme will be implemented on most aspects of the content.

- Name — All modifications will be validated using the system specified below.[8]
- Aliases — Validated modifications, open (not validated) additions with CAPTCHA authorization.
- Content Type — Validated modifications.
- Text — Validation modifications, including insertions and deletions.
- Images — Validated deletions and additions.
- Keywords / Subject Designations — Validated modifications, open additions with CAPTCHA authorization.
- Dependencies — Validation of automatically (system) and manually (user) specified dependencies.
- Citations — Validated modifications, open additions with CAPTCHA authorization.

In order to facilitate rapid growth and large throughput, content submissions will not be validated.[9] Suppose a user modifies some aspect of a piece of content which requires validation. The user's modification will not be visible until it is validated, and accepted, by the authors of the piece of content.

The validation process will proceed as follows. After a user submits an edit, all authors of the corresponding piece of content will be signalled to the existence of the modification and will have the option of either voting for or against accepting the modification.[10] An author may change their vote at any time prior to the closing of the vote. The edit is accepted and the voting is closed if any of the following conditions are satisfied:[11]

1. A majority of the authors[12] have voted for acceptance of the edit, or
2. if there are $N \in \mathbb{N}$ authors, at least $\lceil N/2 \rceil$ authors have voted and at least $3/4$ of the votes are for acceptance of the edit,[13] or

8. Which will henceforth be assumed when discussing validation, unless otherwise specified.
9. Except via later edits.
10. Or abstaining from (/ignoring) the vote.
11. Note that these conditions are evaluated with respect to the time at which they are satisfied, not with respect to the time of the edit. In particular, users who become authors after the validation process begins for some modification may still participate.
12. Note that each author can vote at most once.
13. Note that these details are not finalized, although currently $N/2$ and $3/4$ are chosen because at least $3/4$ of the remaining $N/2$ authors would have to vote against acceptance of the edit for there to be less than a majority for acceptance. This is in general very unlikely, assuming benevolent authors.

3. at 5 days past modification, at least 2 votes have been submitted and at least 2/3 of the votes are for acceptance of the edit, or

4. at 10 days past modification, at least 1 vote has been submitted and a majority of the votes are for acceptance of the edit, or

5. there are no authors,[14] or

6. no votes are submitted within 10 days.

Otherwise, after 10 days, the edit is rejected and the voting process is closed. Upon acceptance of a logged in user's edit, the user[15] is denoted an author of the piece of content and given the ability to participate in the edit validation process of that piece of content. In this way, there is an essentially *viral authorship* system which ensures that those who validate edits have contributed useful, accurate, and appropriate edits in the past and therefore can be assumed to be relatively benevolent and have some knowledge of the content.

The only other caveat to the editing process is that we will assign a probability of ~1/7[16] to the possibility that CAPTCHA authorization will be required in order for any given user to submit an edit.[17] This will reduce the likelihood of a malicious user successfully using a series of computers to perform mass fraudulent modification of content in the database, while also minimizing the fixed costs of editing content for benevolent human users.

The above procedure covers all cases except when a user has submitted a piece of content that is entirely inappropriate or unsuitable for inclusion in the knowledge database or a group of authors are maliciously or inappropriately rigging the validation process in order to advance an agenda which runs counter to the principles of the knowledge database. In this case, any user will have the option to report a piece of content or a group of authors to the administrators[18] of the database, one or more of whom will investigate and make an appropriate action.

## 4 Application Map

The following is a list of each of the (functional) pages[19] in the knowledge database and, for each page, a list of linking pages, that is, the pages that a user can navigate to from that page. Note that these pages are only defined functionally and may not correspond to separate HTML pages or any page-like interface elements. Furthermore, linked pages may not correspond to a unique button (i.e. they may be dependent on previous navigation history). This generates a corresponding "application graph" or map, although it is too intricate to be effectively visualized. Each page (if applicable) contains a list of contained content.

14. Although it will likely be very rare, this could occur if all the authors delete their accounts.
15. If they were logged in when they submitted the edit and are not already an author.
16. To be finalized.
17. Where we assume that a user is defined by a unique IP address (or some similar metrics).
18. Likely individuals from among the organization which will maintain and develop the knowledge database.
19. Or areas/features.

1. Home — 2*, 3, 4, 6, 7, 12
2. Search Results — 3, 4, 5*, 6, 7, 12, 1

   10 content pieces matching the search criteria arranged in descending order.

   For each content piece: Best matching name, content type, snippet(s) of text with matches highlighted, and all keywords.
3. Sign-Up — 1*, 5
4. Login — 1*, 5
5. Content Page — 8, 6, 4, 3, 2, 1, 7, 9, 11, 12, 13, 14

   Name, alternate names, content type, text, citation, keywords, author user names, links to edits page and, if applicable, content author pages.
6. User Page — 1, 2, 7, 11

   Four separate "tabbed" subpages: recent activity, authored content pieces, edit history, and settings.

   Recent activity: All activity on authored edits and content pieces, including acceptance/rejection of authored edits, acceptance/rejection of edits of authored content pieces, submission of validating edits, and admin report notifications. If there is a vote, a notification will urge the user to vote and link to the authored content pieces tab or directly to the edit depending on whether there is a single or multiple validating edits.

   Authored content pieces: Name, content type, last modified date, notification icon indicating whether a validating edit exists or not and, if so, a "vote" button.

   Edit history: List of authored validating/accepted/rejected edits with the content piece name, content part, submission timestamp, acceptance/rejection timestamp (if applicable), and a quantitative summary of net modifications.[20]
7. Content Submission Page — 1, 5*, 4
8. Content Editing Page — 5*, 1, 2, 3, 4, 6, 9, 12
9. Report Content Page — 5*
10. Report Authors Page — 11*
11. Content Authors Page[21] — 5*, 6, 1, 2, 9, 10, 12, 14

20. e.g. number of characters/words added/removed, total formatting modifications made, and images added/removed.
21. Or content validation page.

A list of author user names, a list of (up to 10) validating proposed edits and corresponding current vote results,[22], and a list of (up to 10) recently accepted/rejected edits with corresponding vote results.

For each validating edit: a "vote" button or "vote submitted" icon will be displayed depending on whether the user has voted or not.

12. Admin Account Page — 5*, 13, 1, 2

13. Admin Action Page — 5, 12

14. Edits Page — 5*, 11

    Two lists:[23] the first contains currently validating proposed edits and the second contains accepted/rejected edits.

    For each validating edit: content part, creation timestamp, user name or IP address, a quantitative summary of net modifications,[24] and, if applicable, the edit rationale.

    For each accepted/rejected edit: content part, acceptance/rejection timestamp, user name or IP address, a quantitative summary of net modifications,[25] and, if applicable, the edit rationale.

An asterisk indicates that this is the primary page to which a user will next navigate.

## 5 Use Cases

Users

- A user must register with a username composed of at least two words separated by a space (with each word containing at least 2 characters),[26] a password containing at least 5 characters, and a unique valid email address.
- A user logs into their account by entering their email and password[27] and are then redirected to their account page.[28]

22. Still need to decide the level of information released prior to closure of the vote.

23. Paginated with 20 edits displayed per page.

24. e.g. number of characters/words added/removed, total formatting modifications made, and images added/removed.

25. See previous footnote.

26. That is, with their full name. These will be stored with underscores replacing the spaces.

27. Since their username and password might not be unique; i.e. the username is functionally only a display name.

28. This behavior could be more customized—a user could be redirected to their account page only if there is a pending vote on a piece of content they have authored and, otherwise, they are redirected to the home page.

- Upon registration, a user is sent an email containing a unique[29] link which they must follow in order to confirm their email address. Confirmation must be completed in order to submit a piece of content or be denoted an author after having an edit accepted.[30]
- A user must be logged in to submit a piece of content or be denoted an author after having made an accepted edit; if they attempt to submit a piece of content anonymously, they will be directed to login or sign up.
- When an edit is made to a piece of content which a user has authored, the user can see the existence of the edit, and be redirected to the author page of that piece of content to inspect and vote on the edit, from their account page.

Content Retrieval

- Each piece of content will have its own unique url which a user can follow to view content.
- A user can search for content via a search bar located on all pages of the application, although the home page will be the primary entry site. Ten entries will be displayed per page in the search results.

Content Submission

- After submitting a piece of content, the user will be directed to a preview page to confirm the content of their submission. Upon final submission, the user will be directed to the page of the posted piece of content.
- Each user will be limited to submitting at most one piece of content every minute.

Content Editing

- Each user will be limited to at most one edit every 10 seconds.
- Upon editing, a user will have the option of including an explanation of the rationale behind the edit which will be displayed to all authors when voting.
- After submission of an edit, all users can see the existence of that pending edit on the edits page.

Content Validation

- When an edit is made to a piece of content which a user has authored, the user is sent an email notifying them of the existence of the edit and urging them to vote. Until the voting ends or they have voted, such an email is sent to the user every 4 days.
- When an edit is accepted or rejected, a log of it is displayed on the edits page of the associated piece of content. From the edits page, users can view the changes made or proposed by all past edits.

29. And unguessable.

30. Primarily because an email is sent each time an edit is made to a piece of content the user has authored.

- When an edit is accepted, the piece of content is updated to display the edit.
- While an edit is being validated, that is, a vote is open, all authors can see how many authors have voted thus far. No other information will be available until after the vote is complete, when a complete summary of the results of the vote are displayed in the authors page (including the votes of all authors).

Admins and Violation Reports

- All users may submit a violation report against a piece of content, possibly calling for the deletion of the piece of content. They will only be required to give a nontrivially detailed report of the alleged violation.
- The authors of a piece of content may also submit a violation report against another author or group of authors. They will be required to list exactly which authors are committing the alleged violation and give a nontrivially detailed report of the alleged violation.
- When a violation report is submitted, an administrator is selected to investigate and resolve the situation. The admin may elect to take no action, send messages of guidance or warning against authors or other editing users, temporarily or permanently suspend user accounts, revoke authorship, or grant authorship[31] in the case[32] where there are no authors.[33][34] In the event that the reporter has requested the piece of content be deleted (e.g. if the piece of content was a duplicate of another piece of content or it contained inappropriate text), then an admin may elect to immediately delete the piece of content or give authors five[35] days notice of upcoming deletion.[36] An admin may also elect to take action against a user who is determined to be abusing the violation reporting feature.

## 6 Technology Stack

The following is a basic outline of the main technologies which will likely be used to implement the knowledge database. The emphasis at first will be on using technologies that are simple,

31. To registered users that have made some number of "good" edits to the piece of content.
32. Likely rare.
33. Or the current authors have become essentially unactive.
34. This may not be an exhaustive list of all possible admin actions.
35. Approximately.
36. Primarily if the piece of content was a duplicate or was only partially unsuitable for inclusion in the knowledge database. This gives the authors time to transport some of its information over to other pieces of content if appropriate. The contents of the piece of content will also be emailed to all authors.

proven, minimal, easy to work with, and ideally familiar.

Database — PostgreSQL, for general storage of content and metadata.
ORM — SQLAlchemy, for interfacing with the database.
Search — Elasticsearch, for searching and matching text content.
Processing — Celery and RabbitMQ, for asynchronous and scheduled task processing.
——————— Custom Python API Layer ———————

Framework — Pyramid, for routing requests and serving webpages.
——————— REST Framework API Layer ———————

Front-End — Bootstrap / Foundation / etc. (HTML/CSS).
Polymer / React / jQuery / etc.
Various JavaScript libraries (LaTeX, rich text editing, etc.).