pubgem Charter

Ian Dennis Miller

2016-04-10

The Charter is a seldom-changing record of the original conditions for an agile project. The Charter document describes what the finished product might look like and roughly how we'll get there. The Charter document also describes the parties that have an interest in the project. Finally, the Charter document describes the time and budgetary constraints for the project.

Background I

This project came about due to the following events and factors:

- motivations
    - To further Ian Dennis Miller's research interests, it became necessary to perform a literature analysis which required more than leading academic databases provided.
    - During Ian's MA Thesis defence on the topic of Memes and Social Networks, the question of academic citations was raised; specifically, what to make of insular academic communities that cite within their community and shun citations without?
    - Ian spends an undue amount of time collecting citations, and those citations tend to be very messy with lots of errors, requiring extensive cleanup and post-processing.
- existing work
    - The landscape of academic repositories is limited:
        - Google Scholar
        - CiteSeer
        - Microsoft Academic Search
        - Thompson/Reuters/ISI

Background II

  - library-specific search/indexing facilities

- It is notable that a range of citation management tools exist for duplicating the work that every other citation-user must also perform in order to use a given citation. The following tools come to mind:
  - EndNote
  - Zotero
  - Mendeley
  - Papers
  - RefMan
  - BibWorks
- We are at a time with several powerful file formats and namespaces for organizing science, including:
  - BibTeX: an idiosyncratic but brief and ubiquitous format for citations
  - DOI: a globally unique handle for referring to a digital object
  - ORCID: a globally unique ID mapping for academic authors
  - PDF: a cumbersome but sufficient container for formatted text and images

Background III

    - RSS: an XML file format with polling conventions for distributing updates

- create new opportunities
  - Due to the proprietary nature of existing databases, it is not possible to get a database dump suitable for performing system-wide analyses, such as social network analysis (e.g. coauthor analysis)
  - It is typically impossible to alter or update any of the existing citation indexes. However, as BibTeX is a text file format suitable for tracking with Git, it would therefore be possible for community members to submit new or updated citations as pull requests.
  - It is complicated to “resolve” or “expand” a citation into a readable format. In the absolute best case, this takes about a minute per article. The typical case is closer to 5 minutes. In practical terms, this limits the maximum number of articles that can be processed per day to about 60.

Background IV

- If it were possible to predict the filename that corresponded to a citation (e.g. using the pattern “doi.pdf”), it would be possible to separately distribute articles and their citations, while preserving the ability to automatically synthesize the two into a coherent library.

Stakeholders

The following people and groups, both internal and external, have an interest in this project:

- Ian Dennis Miller
- Collective Intelligence Group
- academics and researchers
- journal editors/boards
- publishers of journals and books
- existing academic search resources

Goals I

In an ideal world, when development is fully done, we will have a product that will:

- the open-source Google Scholar: replace proprietary databases: support web-based users with article search facilities; initially title, author, journal and year; later fulltext
- the reddit of academia: support web-based users with a “front-page news” view of scientific updates
- replace citation management: facilitate direct end-user interaction with an open text file format (BibTeX)
- facilitate community curation of a single authoritative database of every scientific citation
- automatically update the database with new citations as publishers release new works
- provide the entire database in an open, downloadable manner

Goals II

- facilitate synthesis of citations with original archival materials (i.e. make it easy to get a PDF for any given citation)
- support analysis of inter-citations between papers, which is not currently supported by BibTeX and which is non-trivial to extract from an article.

Roadmap I

This narrative defines the rough sequence of steps that will attain the vision. The product will not “pop out” in a fully formed state. There will be several phases/releases that will become ever-closer to the vision.

- git repository with curated .bib files per journal
- journal quantity goal: index 100 most popular journals
- system health monitor
- online journal and article search capability
- “front page” presentation
- user accounts, journal subscriptions, and customized front page
- journal quantity goal: index 1000 most popular journals

Time

This project must take into account the following time factors:

- this is nobody's full-time focus
- academic schedule:
  - summer is May-August
  - fall is September-December
  - winter is January-April
- dissertation defence timeline
- conference RFP deadlines for dissemination

Budget

This project has the following budgetary resources and constraints:

- operating expenses:
  - server: compute, storage, bandwidth (digital ocean; ∼$5 monthly)
  - domain name (nearlyfreespeech; ∼$10 yearly)
  - DNS (nearlyfreespeech; ∼$5 yearly)
  - collaboration tool (Asana; free, billed monthly if paid)
  - code repository (github; free, billed monthly if paid)
- possible sources of funding:
  - academic research grant
  - self-funding (currently provides for 100% of expenses)
  - donations

Minimum annual operating expense is approximately $75 USD.

Working space

Team

The successful completion of this project will rely upon the following team members for their particular skills.

- Ian Dennis Miller; Project Lead