Proposal

Abstract:

There are many tools to find popular information. To make new discoveries and invent out-of-the-box technologies, however, scientists need something different: the most important information that is not yet popular. The insights underpinning a scientific breakthrough are often “diamonds in the rough”, appreciated only by a few highly specialized minds and ignored by others, sometimes languishing in obscurity for years. Currently, there is no tool that supports and accelerates the crucial act of sharing unique, expert knowledge among networks of scientists from different domains. Beagle, named after Charles Darwin’s ship, is a real-time social sharing tool, which we are designing with the unique needs of scientists in mind. Our ultimate goal is to create Beagle as a scalable open source platform to facilitate the rapid, well-incentivized sharing of insights and annotations that live “on top” of the scientific literature. In this proposal, we will create a secure, distributed social exchange for scientific insights to underpin Beagle, and make it instantly accessible via an elegant browser plugin which operates on top of HTML and PDF documents. To do so, we will create an annotation sharing system that is secure, federated and socially decentralized. We will utilize a secure distributed file system (IPFS) for scalable storage of shared annotations, in a manner that does not require centrally-controlled servers or proprietary software. Furthermore, to incentivize new users by making it trivially easy, as well as safe, to begin sharing insights, we will employ authentication mechanisms that build on existing identity providers while maintaining security throughout the pipeline and maintaining compatibility with the state of the art interoperable standards for web annotation generally. Together, these innovations will comprise a powerful platform for recording and sharing scientific insights.

Research Goals:

As Newton famously pointed out, scientific progress takes the form of insights that build on top of what has been done before, i.e., critiques, extensions or combinations of elements in the existing scientific literature. The goal of the Beagle software platform is to accelerate this process. Beagle is designed to capture such insights quickly at the exact times they are generated and to route them to others who can build on them. It does this by using a tool researchers already have: their browser. Beagle is currently a Chrome extension (extensions in other browsers are planned) which means that researchers looking at PDFs or websites online -- according to their usual research workflows -- can instantly access Beagle's features. When reading a scientific paper, the user can highlight text or images, and share an insight about this snippet with other scientists in their trusted collaborative network. The insights can take the form of questions broadcast to a network of potential experts for feedback, commentary about the paper, or simply a flag for others to pay attention to a particular detail. In this way, a network of collaborating scientists can dynamically route expertise to where it is needed. The system is designed to incentivize the exchange of ideas, and to build a pool of shared commentary that will allow newcomers to understand the true state of expert knowledge in a field, easing barriers to cross-disciplinarity. If the recipient is not already a Beagle user, they simply receive these insights in the form of emails, from which the recipient can instantly join the conversation, thereby growing the collaborative network.

Incentives are crucial in changing the behavior of the scientific community to enable faster and richer sharing: scientists not only must trust the colleagues with whom they share insights, and the privacy of their pre-publication ideas, but they must also be socially rewarded for sharing, and given credit for insights they share publicly. Accordingly, we have hypothesized that a social annotation tool for science will need to satisfy several properties:

1) It must minimally disrupt scientists' existing workflows.

2) All annotations must be timestamped, and stored securely and robustly.

3) The scientist must have control over who sees their insight, and whether it is made public.

To satisfy requirement #1, we are designing Beagle as a browser plugin that is available at one click "on top" of the paper the scientist is already looking at. We will enable minimal workflow disruption in the context of a social platform by creating user profiles that ride on top of existing cloud services, via OAuth. These user profiles will also allow us to satisfy requirement #3 and to readily integrate Beagle with existing services like gmail and Twitter. To satisfy requirement #2, we will implement storage of shared insights using a secure distributed file-system known as IPFS [1], which was developed by our collaborator Juan Batiz-Benet. Integrating these features will enable Beagle to become a powerful platform for scientific sharing, designed with the unique workflows and incentives of scientists in mind.

Prior Work:

Funded by a MetaKnowledge grant from the University of Chicago [2], we have been prototyping Beagle's basic user interface and software structure. As an alpha, Beagle is already available as a Chrome extension. The technical stack uses open source node modules and browserify to bundle the back end code into the extension, and then React.js to display the front end code. PDFs are rendered using PDF.js (developed as an open source project by Mozilla), as the native Chrome PDF viewer does not have a usable API, and annotations and highlights are created using the PDF.js API and stored in a PouchDB database that syncs to a CouchDB instance on AWS. The goal is to swap out this PouchDB / CouchDB database with IPFS, so that the data is distributed and doesn't pass through a proprietary silo. Mails are sent using Mailgun, and screenshots of complex snippets (images, graphs, etc) are gathered using the Chrome Extension native API. Beagle itself works as a sidebar, which is either programmatically injected as an iFrame on HTML pages, or as part of the PDFjs viewer for PDFs. The code is extensible, with the goal that eventually users may be able to see the sidebar on a site without needing to have the Chrome extension installed. Beagle also accesses several external APIs, such as AltMetrics -- for data about how the paper has been shared on social media -- PLOS, and we have plans to access pubmed, ArXiv, and other scientific databases.

Proposed Work:

Aim 1: Create a secure, distributed annotation storage framework for Beagle

Beagle must store annotations securely and privately, as well as be flexible to users' data storage constraints. Thus, beagle should by default encrypt users' data to ensure privacy and allow selective sharing. Moreover, though for convenience beagle will "just work" with data storage servers we will host, beagle *must* allow users to opt to store all their data in storage servers under their control. This flexibility is often required by security-sensitive labs, particularly those working on long-running research projects in crowded fields. Addressing these constraints requires designing and implementing a simple -- but critical to get right -- data access cryptosystem for beagle. We will use well-established cryptographic protocols to ensure users' data is encrypted at rest, and only decryptable by the authors, and the audiences they explicitly select to share insights with. The Beagle data storage layer is designed to be swappable, so that users may setup their own private storage servers easily, on top of commodity cloud storage systems, without depending on the security properties of those systems -- e.g. storing only end-to-end encrypted data. We will employ standard practices, as well as a new versioned distributed file system (IPFS), to make implementation of this system feasible for a small team.

Aim 2: Enable social annotation based on existing cloud services and annotation standards

Beagle annotations will be stored as JSON objects in a simple key-value store. Currently, the schema used has been developed for this specific project, but interoperability with other annotators is planned. For instance, the work of OpenAnnotator and Hypothes.is may be useful in the future, especially as their annotation schemas have been built to conform to bleeding-edge specs written by and with the W3C annotation working group. Each annotation is saved using either the fingerprint of the PDF - a unique identifier - or using the URL of the website where the annotation was made. Since the annotations are persistent, and can be loaded by anyone with access to Beagle, anyone who looks at a PDF which has been annotated can see another person's annotations just by nature of being able to look up the PDF fingerprint. Our aim here is to enable privacy, by using the Google OAuth API to enable user accounts and restricting sharing. Users would be able to choose who they want to share their annotations with by accessing the Google Contacts API and emailing their annotations or the PDF or URL to their colleagues or students. Annotations would have multiple tiers of privacy, like Google Docs; with the ability to view or edit, and with limited or unlimited restrictions on the users on the other end (i.e., public annotations or not). Furthermore, users are able to share annotations on Twitter and other social media. This is enabled for the purpose of helping to bootstrap Beagle's user base.

Aim 3: Implement social sharing models

A crucial aspect of Beagle is the use of social networks to incentivize participation. Building on the basic infrastructure developed in Aim 1 and Aim 2, we will implement and test multiple social interaction structures. One model is similar to Google Plus or to an access-controlled version of Twitter: users are incentivized to share by getting feedback from colleagues, and users can follow annotations on specific papers or by specific colleagues who grant them access (or who share publicly). Another model is for the annotator to directly share annotation threads only with specific trusted colleagues, similar to the mechanism of sharing a Google Drive or Dropbox folder. A third model is a Wikipedia-like joint curation of annotations, with monitored annotations and volunteered edits in a large open forum. We anticipate that different types of users will prefer different sharing models, and that these sharing functions can be implemented modularly inside a single plugin architecture. For example, a group of students learning a new field may prefer the Wikipedia-like model, whereas a professional workgroup doing research in a competitive field may prefer the targeted sharing model for their core research, while also using a Twitter-like model to keep up to speed with developments in more peripheral fields of interest.

References

[1] IPFS - The permanent web: https://github.com/ipfs/ipfs

[2] http://www.knowledgelab.org/news/detail/1.4_million_in_grants_awarded_to_metaknowledge_projects