Abstract:

Much valuable scientific knowledge is "implicit", nowhere written down and often never shared outside individual laboratories. These insights that lie "on top" of the scientific literature include: which findings are interesting, which statements deserve more scrutiny, which approaches are socially sanctioned and which are radically divergent, which factors or biases might have motivated a study and what might be the most interesting thing to do next. Because this knowledge is implicit, it is tremendously hard to aggregate, parse, and analyze, in the quest for greater and more systematic scientific progress. We thus wish to convert as much as possible of this implicit knowledge to an explicit form, allowing scientists to search the pool of collective meta-insights, in real time, as they browse the literature. Under-appreciated but valuable directions or opportunities might then be brought to light more quickly, because connections between formerly implicit ideas could be made.

To create a sharable meta-layer on top of the existing bodies of scientific knowledge, and to enable such a layer to achieve widespread use among scientists, we propose to leverage a remarkable "Web 2.0" technology: the ability to overlay content "on top" of any webpage or paper viewed by the scientist in their browser. Specifically, using open-source tools such as *Hypothes.is*, we will create a "browser plugin" that detects when a user is reading a paper or performing a scientific search query, and then overlays high-resolution (but appropriately attentionally unobtrusive) meta-information on top of the paper or search result. The plugin is called *Beagle,* after Charles Darwin's ship, from which he collected diverse observations that collectively underwrote the theory of evolution, while roaming the world. Beagle will allow the user (scientist) to enter "annotations" and notes about the paper or query, either for their own use, or for sharing within defined networks of trusted colleagues (to protect the user's desire for confidentiality of her most potent insights), or for sharing publicly. Extension ("apps") can be built on top of this browser plugin, to enable more specialized forms of annotation of the literature – e.g., comparing quantitative features of the various optogenetic proteins – as well as to enable meta-knowledge research by mining the pool of (e.g., publicly available) annotations.

3. **Project Description:**

Beagle is a software platform designed to assist scientists as they navigate the massive and overwhelming body of scientific literature. Beagle will primarily be used via an interface that enhances the experience of reading a paper. It will make important information immediately more easily accessible to the user, and unobtrusively incentivize user participation in the semantic annotation of the scientific literature, thereby accelerating the scientific process. The crucial contribution of the Beagle project will be to develop a unified, modular and unobtrusive user interface that scientists will want to use, hopefully enabling many thousands of scientists to contribute on a daily basis.

Beagle is an early example of a "scientist-as-end-user" meta-knowledge tool that simultaneously benefits scientists, accelerates research, and grows the pool of one of the crucial resources for meta-knowledge research: the explicit documentation of otherwise tacit knowledge. Because of its modularity, Beagle will embody not just a discrete application but also an extensible platform for the deployment of a wide range of meta-knowledge applications in the future. Thus, Beagle and it descendants could form

an important part of the core infrastructure for "plugging in" a wide community of scientists to the meta-knowledge endeavor.

By allowing scientists to more easily access and codify the "context" surrounding any given document or search query, and by incentivizing scientists to share this context with one another, Beagle will promote a humble approach to knowledge in which emphasis shifts from an individual paper to the totality of collective knowledge – as well as uncertainty – surrounding that paper.

In what follows, we first explain the problem of scalably accessing tacit scientific knowledge and describe the incentive constraints that shape the adoption of end-user meta-knowledge tools in general. Based on this framework, we then derive the design principles of Beagle and detail its architecture and implementation plan.

Tacit scientific knowledge

Today, scientific knowledge is stored largely in the form of publications, which report the results of research after the fact. The use of publications as the primary medium of explicit written scientific communication has consequences for the structure of science and for the rate of scientific innovation. Because publications hold a monopoly on the communicative media of science, much of our civilization's extant scientific knowledge – indeed, nearly any aspect of science that that doesn't explicitly make it into a publication – is currently "implicit" or "tacit", nowhere written down and often never shared outside of individual laboratories. This body of tacit knowledge includes, among other things, the types of insights that lie "on top" of the scientific literature: which findings are interesting, which statements deserve more scrutiny, which approaches are socially sanctioned and which are radically divergent, which factors or biases might have motivated a study and what might be the most interesting thing to do next.

Because this knowledge is implicit, it is tremendously hard to aggregate, parse, and analyze, in the quest for greater and more systematic scientific progress. We thus wish to convert as much as possible of this implicit knowledge to an explicit form, *allowing scientists to search the pool of collective meta-insights, in real time, as they browse the literature*. Under-appreciated but valuable directions or opportunities might then be brought to light more quickly, because connections between formerly implicit ideas could be made.

An illustrative anecdote can be found in a long-ignored 1971 paper of Gobind Khorana, which described in lucid detail – in the last paragraphs of the discussion section – the steps of the polymerase chain reaction (PCR). Since no experiment was shown, Khorana's ideas were ignored and quickly forgotten, and PCR was re-invented 14 years later, leading to a Nobel prize and a pervasive transformation of biomedicine. While the history surrounding a major scientific discovery is inevitably complex, examples like this raise tantalizing questions. Are there similar gems hidden in today's literature, which could be brought to light by "connecting the dots" among ideas latent in existing papers? Might the 14-year delay have been reduced by an online tool that encouraged the sharing and discussion of meta-level insights on top of the literature, allowing shared annotation at sentence-level resolution?

Even an interface that simply enabled scientists to tag parts of papers as "interesting" or "novel" might have highlighted the pivotal last few sentences of Khorana's analysis,

perhaps bringing attention to the need to test those ideas experimentally sooner rather than later, or perhaps preventing them from being ignored and forgotten for so long. Perhaps more sophisticated tags, emerging in an online conversation among interested experts, could have pointed towards the specific experiments that would need to be done to validate the intriguing suggestion, or towards the people who would enjoy trying such experiments, who could (with the expediencies of today's internet) have been swiftly brought into the conversation.

Towards explicit encoding of tacit knowledge through semantic tagging

Exposing and leveraging this tacit knowledge – the knowledge that connects the dots between scientific studies and embodies the "direction" in which science is heading – is not a task that can, at the present state of artificial intelligence, be fully computer-automated. Making such knowledge explicit requires ability to incorporate *human* insight, individual human perspective, and diversely trained scientific *understanding* into the large-scale analysis of the literature.

While typical queries to a scientific search engine simply search for the conjunction of keywords or phrases, we would like to enable a "semantic" search capability. There are many semantic questions that aren't answered by any one paper, and that wouldn't naturally be the subject of any one paper. In some cases, a human scientist, out there in the world, may already know the answer to a question; in other cases, the answer may never have been thought of in an explicit way. In some cases, like the question "which genes are implicated in the etiology of schizophrenia", aggregating statements over a collection of relevant papers in some simple way may produce an answer; in other cases, no simple list or tally will suffice. To expose such underlying semantic content, large-scale automated computer-based "mining" of the literature – i.e., automated science analytics – is, for the foreseeable future, insufficient. Instead we must generate a "semantic web of science" through crowd-sourcing the knowledge of trained scientists. In particular, we would like to crowd-source the human annotation of scientific papers with structured meta-information and hyper-links that "tag" features at multiple levels.

At the level of entire papers, a system could allow scientists to tag: papers that are surprising or unique, papers that are relevant to a given research direction, pairs of papers that are instances of same idea, papers that are replications of other papers, pairs of papers that make opposite claims, papers that solve problems posed by other papers or papers that clarify the results of other papers. At the level of sub-sections of papers (such as paragraphs or sentences), tags could endow statements with simple semantic labels like: "method", "tool", "hypothesis", "refutation", "confirmation", "dataset", "analysis", "explanation", "physical limit", "meta-analysis" or "finding with P-value [x]".

Once such an initial semantic tag set was partially or fully created by a network of taggers (human participants), the taggers could be queried in a second round, to fill in details and deepen the level of semantic structure. At this point, the engine could ask the taggers questions like: "which reference cited by this paper contains the hypothesis that it is refuting?" or "which reference cited by this paper contains the method that it improves upon?" Taggers could also, through a democratic process, nominate new tags to be added to the mix. Error checking could be performed by cross-validation across taggers.

The feasibility of very-large-scale semantic annotation:

But is it really feasible to make tacit semantic knowledge, currently stored only in scientists' minds, explicit and accessible? Is there even enough human effort to go around, to semantically tag the entire literature? Can semantic taggers ever hope to keep pace with the constant creation of new knowledge? A quick order-of-magnitude estimate suggests that this is likely possible, at least in principle.

Roughly 1.5 million papers are published per year, with a 2.5% annual growth rate. This exponential model predicts a total number of scientific papers since the 19th century on the order of 5.8×10^7, consistent with estimates of about 50-60 million academic papers in existence. For comparison, the NCBI PubMed database has 23,723,205 entries as of April 2014, of which about 4 million (<10%) have free full text (and hence freely accessible reference lists, as well, necessary for defining the "citation graph" of the literature). About 3 papers are published per minute at current rates and the number of papers published per year works out, perhaps revealingly, to roughly one paper per PhD student per 5 years.

There are thus many scientific papers, but the total amount of text is small and slow-growing compared to other types of ``big data''. It is conceivable to use manual, crowd-sourced human annotation to tag many or all of the existing papers in a reasonable period of time, and to keep up in real time with the influx of new papers. We thus tentatively conclude that user-participatory systems (in addition to fully automated systems) could, in principle, scale to the level of the entire literature.

The problem of user incentives

The major problem with the idea of using human-produced semantic annotations to convert tacit scientific knowledge into an explicit form is not the sheer scale of the problem, in terms of the number of papers needing tagging, but rather the following conundrum: how do we *incentivize* a large number of scientists to participate? Not only are scientists busy, but they are also often highly conservative in choosing which new software platforms to adopt. Furthermore, scientists are rewarded for producing new published discoveries – they cannot get a PhD, or grant funding, or tenure, by annotating the literature, no matter how impactful that activity might be, i.e., no matter how important or influential their tacit insights might be, once made explicit and accessible to others. Especially in fields with a long latency between the initial theoretical formulation of an idea and its conclusive experimental validation, scientists are often afraid of "getting scooped": what if another group takes our insights, we often ask, and uses them to produce an experimental result before we can, making months of work appear redundant and hence un-publishable? *Thus, any software tool that aims to extract, codify and publicize insights from individual human scientists must be designed very carefully, or else scientists will feel that they have absolutely no incentive to use it, and, in fact, every incentive to steer clear of it.*

The contexts in which scientists will participate in semantic tagging

Given these incentive constraints, we can identify three core requirements for an end-user interface. First, the tool must minimally perturb the scientist's existing workflow. For example, if the scientist must go to a special website to upload a paper every time they want to share or discuss it with their colleagues, this would perturb their existing

workflows undesirably, whereas if paper sharing could be built into scientists existing emailing routines, that would perhaps be less perturbative and more likely to be adopted.

Second, the tool must provide immediate and obvious value to the user, saving the user time on the activities that she already does, and that she already knows she wants to do, rather than adding additional tasks. For example, scientists are already sharing links to papers with their colleagues, sometimes privately by email, and sometimes publicly on social networks, are already taking notes on papers, looking up citations, and browsing the literature through the "cited by" feature on sites like Google scholar. A tool should make those tasks noticeably easier before it forces new tasks upon the user.

Third, the tool must accelerate the systemically-incentivized value-production in a scientist's work – publishing peer-reviewed papers, hiring students, acquiring grants, receiving accolades, filing patents – without creating any career-related risk to the scientist. For example, if a tool requires scientists to share their proprietary ideas and data with the public, before they submit a paper for publication, many users would fear "getting scooped" and decide not to participate, even if they appreciate, in the abstract, the importance of open science and open data. This holds for semantic tags and tacit ideas as well: often, scientists will only share their best ideas with small trusted groups, from which they can expect to derive useful feedback without incurring the risk of getting scooped, such as their own laboratories, former lab-mates, close colleagues and existing collaborators on grants and papers. Even the list of which papers a scientist finds interesting, important or surprising – or those she finds incorrect or misconceived – can be a valuable currency that is often shared only within trusted networks; perhaps this is why "social bookmarking" tools, which publicly display a list of a scientist's favorite papers, have not yet gained widespread adoption among biologists.

<u>The interplay between supporting default behavior, and changing behavior</u>

At the same time, it is not productive simply to preserve the status quo. The reality is that not only the community as a whole will benefit when scientists share more and more quickly – and when they adopt novel software platforms in order to do so – but individual scientists will usually benefit strongly as well, despite their fears to the contrary. Importantly, though, scientists *will only start to see and appreciate the value* of new knowledge-sharing tools once they are already actively using them and feeling safe doing so; only then can they receive the tangible rewards of participation – such as valuable new collaborations and fun new interactions, a higher rate of learning, a greater frequency of interesting surprises, richer context for motivating and selling their work, and a more fertile crossover between fields – while all the time feeling comfortable. Thus, on our view, the path to adoption of new open-knowledge tools is not primarily one of activism – which emphasizes taking risks and incurring costs to support an abstract cause – but rather one of bootstrapping the adoption of tools that lead to a gradual expansion of scientists' collaboration networks, and of the range of content that they comfortably share on a routine basis.

<u>Deducing an optimal incentive structure for end-user metaknowledge tools</u>

In light of the above, there are various options for the incentive structures that culd be embodied by a tool aiming to enable large-scale semantic tagging, but we can readily see that most of them don't hold water. For example, one model would be "annotate your own papers, for fame and glory". But this doesn't work because you, the scientist,

have lots of papers, and particularly lots of old papers, but very little time. Furthermore, the ideal is for a wide diversity of researchers to annotate any given paper, not just self-annotation. Another model would be “get access to everyone’s annotations after you annotate 10 papers”, but this does not provide the user with immediate value, and incurs a bottleneck in the takeoff phase, when there are initially few annotators and hence only a small amount of total content. Yet another model would be to give all users access to the full system from the beginning, and to rely on a core group of motivated contributors to seed the initial content. Anecdotally from small experiments (such as our attempted creation in 2011 of SynBioOverflow.org, a synthetic biology question and answer site), even tight-knit and motivated groups are unable to seed sufficient content to get such a site rolling organically.

It is therefore essential to reduce the “transaction costs” for the user to near zero, by making it a trivial one-click operation to add common annotations. This is similar to what the company LinkedIn did for the crowd-sourcing of knowledge about people’s professional expertise: LinkedIn provided users with you a guess as to their friends’ expertise, and then the user merely had to endorse this suggestion, or not, with a single click of a button unobtrusively hung near the top of the page. It is important to note that LinkedIn’s endorsement system effectively relies on an implicit notion of social reciprocity among friends, to motivate the button clicks. Likewise, for semantic tagging, low transaction costs are not enough on their own – the platform also needs a “carrot" for the user: a reason to draw users to the system in the first place, as well as a social motivating factor, however minimal, to actually commit a semantic tag.

As described above, it is also essential to build the desired operations (e.g., semantic tagging or idea sharing) into something the user wants to do already. In this regard, we first considered an email add on, which would simply allow users to add semantic tags through a quick form-like interface while they are already emailing a link to a paper to their friends. After much discussion, we have converged on a proposed optimal design of Beagle as a browser add on, which pops up while the user is reading a paper or searching for papers, providing useful information while enabling one-click tagging, note-taking and social sharing. Thus, Beagle is conceived as a socially enabled browser add on that also provides immediate non-social utility.

Summarizing the above, the major design constraints on Beagle are: 1) Should provide immediate value out of the box, enough to justify use, solving problems people know they have immediately, 2) Should minimally perturb workflow, and minimally burden attention, 3) Tagging or other participatory activities should be socially rewarding, just like email, Facebook or Twitter 4) Should have privacy setting such that people feel comfortable sharing with a variety of user-defined group sizes and group compositions and 5) Should leverage scientists’ existing networks, e.g., labs, mentors, students, colleagues.

A sufficiently widely used platform for idea-sharing and paper annotation effectively becomes a “scientific social network”. The existence of such a network would have wide-ranging direct benefits for scientists, as well as benefits for meta-knowledge research.

Solution: Beagle platform

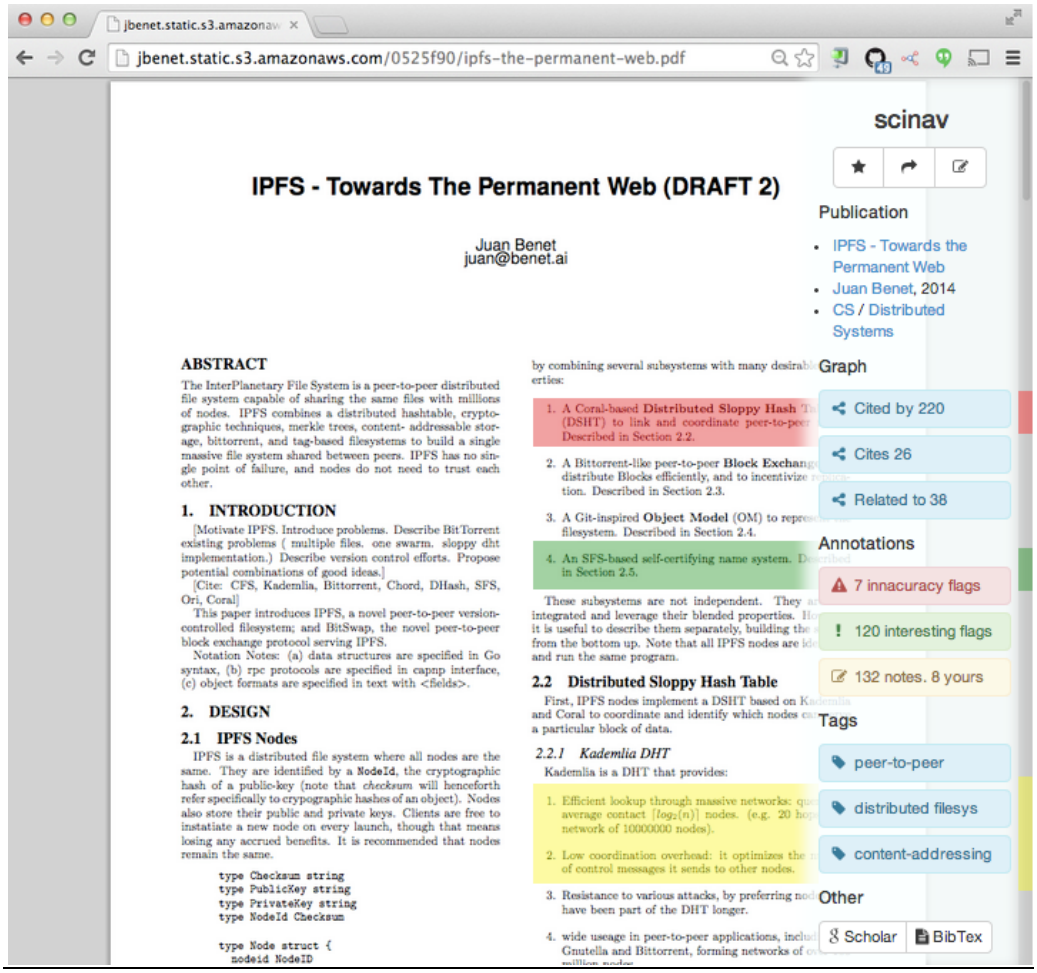

Figure 1: Preliminary mockup of the Beagle interface

To create a sharable meta-layer on top of the existing bodies of scientific knowledge, and to enable such a layer to achieve widespread use among scientists, we propose to leverage a remarkable "Web 2.0" technology: the ability to overlay content "on top" of any webpage or paper viewed by the scientist in their browser. Specifically, using open-source tools such as *Hypothes.is*, we will create a "browser plugin" that detects when a user is reading a paper or performing a scientific search query, and then overlays high-resolution (but appropriately attentionally unobtrusive) meta-information on top of the paper or search result. The plugin – with a preliminary mockup depicted in **Figure 1** – is called *Beagle,* after Charles Darwin's ship, from which he collected diverse observations that collectively underwrote the theory of evolution, while roaming the world. Beagle will allow the user (scientist) to enter "annotations" and notes about the paper or query, either for their own use, or for sharing within defined networks of trusted colleagues (to protect the user's potential desire for confidentiality of her most potent insights), or for sharing publicly. Extension ("apps") can be built on top of this browser plugin, to enable more specialized forms of annotation of the literature – e.g., comparing quantitative features of the various optogenetic proteins – as well as to enable meta-knowledge research by mining the pool of (e.g., publicly available) annotations.

Such a tool could provide great utility to scientists right out of the box: overlaying information "on top" of existing papers and websites, exactly where and when scientists they need it, yet without burdening their attention. We therefore believe that it could see widespread adoption. In fact, over time, Beagle could naturally incorporate and supersede a variety of existing services like reference collection and sharing (Mendeley, Epernicus, ResearchGate), social commenting on scientific articles (Twitter and Facebook), browsing of the citation network (Google Scholar), alternative metrics (e.g., AltMetrics), post-hoc peer review (e.g., PubPeer, SelectedPapers), and specialized search (e.g., "advanced" PubMed queries), because *it enables idea "crystallization" right when people are most interested -- when they are confronting novel concepts and synthesizing them with their own research directions.*

Initial "apps"

Basic Info	**Publication Graph**	**Annotations**	**Social**	**Algorithmic Parsing**
Title	*Cites*	Hypothes.is annotations	User mentions	*Extracts suggestions for a paper's semantic tags using NLP*
Author	*Cited by*	Carry a flag: inaccurate, interesting, refuted by [x]	*Tweets*	*Statistically improbable phrases present in the paper*
Year	*Related papers*	*Notes to self, and flags for one's own use*	Friends who have read it	
Journal	*Sub-field (automatically determined)*	Question	Friends who have flagged it	
Copy BibTex or other citation to clipboard	*View in citation graph partitioned by sub-field*		Friends who might be interested	
Copy DOI to clipboard	*View in co-authorship graph partitioned by sub-field*		Questions about it from your friends	
Find/request open access copy	*Similar authors*		Users who viewed this paper also viewed [x]	
	Similar papers authored by people in your network			
	People in your network who are connected to these authors			

Table 1: Meta-information that can be incorporated into the Beagle interface for papers. The elements in *italics* can be present even before the tool receives wide adoption, i.e., in the absence of "social" features.

Early modules to incorporate into the plugin fall into two categories: 1) "apps" that overlay information on top of scientific papers as they are viewed in the browser and 2) "apps" that overlay information on top of scientific search queries on sites like Google Scholar or PubMed. In the first category, initial apps could include: recent papers by same author, related papers, which papers cite a paper, which papers a paper cites, get citation and email a link to this paper (with a Beagle wrapper). These are summarized in **Table 1**. In the second category, possible apps include: segment search results

hierarchically by sub-field, re-order search results by relevance to the user based on their browsing history or publication history and search query diversification.

The detailed software architecture of Beagle

The Beagle system will need multiple software components; the first core version of the system will include the following breakdown.

Beagle Navigator: The Navigator is the main software component and comprises the bulk of the development work. The Navigator itself will be primarily a loader for smaller modules. It will do some common things (get, parse and store data; render the navigation bar), but delegate the bulk of the interesting functionality to smaller independent modules.

The Beagle Navigator's interesting features are implemented via Modules, software components that implement pieces of functionality. Think of them as mini-applications. Modularity is key for (a) simplifying the software system, (b) isolating concerns, (c) simplifying building interesting subcomponents, (d) leveraging a community of contributors. Additionally, we can imagine a wide range of apps being created in the future, for specialized user bases, as modules on top of the core Beagle system.

Capabilities of a module include:

- Can display a section on the side navigation panel
- Can open larger dialogs
- Can render on top of the paper (highlighting, etc)
- Can store and access their own data
- Can access common data (paper metadata, paper text, etc)

Characteristics of modules include:

- Are developed independently
- Have their own small codebases
- Use common UI elements (so that everything feels uniform), but can render whatever they want (e.g., visualizations)
- Can call into a beagle API to get/store data
- Can be turned on and off by the user
- Occupy positions in the main panel that can be reorganized by the user

The software architecture of the Navigator should be as decoupled from its mode of deployment as possible. Beagle will be deployed primarily as a browser extension. But, its codebase will not be extension-specific, so that we can also readily deploy it via other means (e.g., website, desktop client).

Browser Extension: Embeds the Navigator on web pages (particularly PDFs). Derived in part from Hypothes.is, but with a large amount of interface customization. The core contribution of Hypothes.is is the annotation backend and workflow, whereas their user interface is a lightweight layer on top of this. We will re-do the interface layer to be maximally modular (for incorporating non-annotation-based “apps”) and maximally appealing to scientists (for example, we may decide to incorporate some content at a thin bar on the bottom of the page rather than on the side).

Beagle Data Backend: A service to store Beagle data. Beagle will need a way to store some data on behalf of the user and the modules. Because the data types are simple, we only need a key-value store (leveldb in server, pouchdb in browser). We will use either a very simple REST API server (node + leveldb), or potentially a hosted service.

The annotation data will follow the Open Annotation format (what Hypothes.is uses), and may be subject to strong privacy concerns. Some users may require their annotation content be hosted in their own servers. Hypothes.is will soon release an annotations backend server, and ideally we can just use that. Teams for which privacy is critical can host their own server, and direct Beagle to store their group's annotations on it.
We need to make it very easy for people to host their own private data-stores, or else we will ultimately need to have a large, secure, impartial entity in charge of everyone's data.

Beagle Website: A very simple website that describes the tool/project and lets people install. This can be hosted directly on github (gh-pages). Later on, this website could also load and display some of the user's Beagle data (e.g., links to starred papers, annotations, notes, questions).

Navigator Wrapper Web-app: A simple app that "wraps" a link to an existing PDF with the Beagle navigator on the side, without requiring that the user has already installed the Browser Extension. The purpose of this is to allow sharing of annotations with people who have not yet downloaded Beagle, e.g., when papers and their annotations are shared with non-users via email, thereby engaging them in the system for the first time. That way, users can try Beagle out before signing up. From an implementation standpoint, this is simple, and can be hosted directly on github pages. This will also require a "browserified" version of Beagle.

Later versions of the system may also include:

1. **Beagle Mobile App**: a simple mobile application to access beagle data.
2. **Beagle Desktop Client**: a simple desktop client to browse PDFs with the beagle helper. In actuality, the application is just a browser, but embedded in a standalone desktop application with access to the local filesystem.

4. **Project Outputs, Impact & Timeline:**

The outputs of the proposed project include:

- Beagle open source software codebase: and open, well-documented GitHub repository with the full software for the browser plugin and associated infrastructure

- Beagle software: the actual browser plugin

- Beagle website: a website describing the project, where the software can be downloaded

- Beagle API: a software specification that makes it easy for researchers or other to create new Beagle modules

Timeline:

The basic timeline for development and deployment of Beagle is as follows.

Fall 2014	Initial prototyping of Beagle, interactive user testing among scientists and engineers in the Synthetic Neurobiology Group and its close affiliates. Software development by skilled professional coders under the guidance of neurobiologists and meta-knowledge researchers.
Winter 2014-2015	Pre-beta release, testing in other biology groups at MIT and elsewhere. During this phase, help on software development will also be actively solicited from the broader open-source community.
Spring 2015	Scalable deployment platform (file storage, server load, backup). Adding more advanced or ancillary features. Begin engaging development partners (e.g., companies and foundations, as well as the government).
Summer 2015	Organizational development and deployment, including finding a community-trusted host for the data, and securing continued funding and software support.
Fall 2015	Full release to the community.

Impact on meta-knowledge research:

The impact on meta-knowledge research will be far-reaching. First, if successful, Beagle will generate a large dataset of semantic tags for scientific papers, which are naturally of enormous interest to meta-knowledge research, as well as an active user base of connected scholars willing to share tacit insights in a codified form. Because the meta-knowledge network is plugged into this effort from the beginning, it will be possible to study the growth and outputs of the system during the early prototyping phases – informing future efforts at developing end-user meta-knowledge tools; likely, certain types of studies on the Beagle network (e.g., appropriately anonymized) will be possible even after its full deployment to the community, subject to acceptance by the user base of such "analytics" taking place on top of the platform.

Perhaps even more importantly, we argue that Beagle could provide the first viable platform for end-user tools that both generate and apply scientific meta-knowledge on a large scale, incentivizing many scientists to participate directly in meta-knowledge efforts. The incentive problems that we described above, with respect to the semantic annotation of papers, are in fact generic to any meta-knowledge tool targeted to the scientist as both an end user (client) and as a generator of content (worker). Why should the user go to a new website or download a new piece of software, to benefit the growth of useful meta-knowledge, if the benefits to them personally are remote? In the present environment, these incentive problems are not being taken seriously enough, and consequently we currently lack a viable platform for sharing scientific insights outside of

traditional peer-reviewed publications. Indeed, there is a proliferation of tools offering different services, but each provides little value; this is exacerbated by users having the burden of choosing between and/or signing up for many competing such services.

One of the primary motivations of Beagle, in contrast, is to solve the incentive problems only once, rather than trying to solve them again and again, every time a new tool or feature is deployed – first, get people to sign on to a platform, and then allow apps to be easily plugged into that platform. Therefore, Beagle is explicitly designed not only according to the above-outlined principles of user incentivization, but also to be structurally modular and open source, facilitating its use as a broader platform for end-user meta-knowledge tools. Even in the worst case (which we will work hard to avoid), where Beagle itself does not achieve widespread community adoption, it will still provide a valuable open source toolkit and codebase for annotation and content extraction from scientific papers, as well as the software base for a smooth user interface for future meta-knowledge Brower plugins and other related tools.

Because Beagle is emerging jointly from the meta-knowledge network and from a network of practicing end-user biological scientists (as well as from a vibrant software development community in Silicon Valley), with the dual, interconnected goals of a) developing a popular tool with widespread adoption and utility to end users and b) facilitating the growth of meta-knowledge codification and mining capabilities, we suggest that supporting the emergence of this platform could be a beneficial investment for the meta-knowledge community.

5. **Project Team Description:**

The collaborative team seeding the Beagle project consists of a unique, synergistic mix of expertise.

The Principal Investigator on this project, Ed Boyden, is a renowned neuro-technologist whose large, active lab collaborates with hundreds of other groups worldwide, in an omni-disciplinary fashion, driving innovation in fields ranging from microbial genetics, to robotics and microscopy, to neurosurgery and data science. Ed is a prolific inventor and scientist with who understands how innovation in biological science occurs, and how to accelerate that process. In the mid-2000s, Ed co-invented rhodopsin-based optogenetics, been described as “the most significant advance in the last 40 years of neuroscience”.

Collaborator Adam Marblestone is a multi-disciplinary scientist with experience in physics, biophysics, nanotechnology, neuroscience and more recently with mining of the biomedical literature. He has recently spearheaded efforts to make neuroscience more collaborative, such as through a recent study on the physics of brain activity mapping that brought together 14 different groups in a single analysis. Adam also has experience in developing software for science, and is deeply interested in tools to broadly facilitate scientific discovery.

Collaborator Konrad Kording is a computational neuroscientist and statistician with an active research program on the “science of science”, as well as on pure neuroscience. He is a member of the meta-knowledge network.