Committed to Science

– Scientific publishing in the web age –

A DRAFT for a grant proposal,

written by

Members of the Open Science Community

Abstract

About this project:

- This file serves the collaborative drafting of a project proposal on an **Encyclopaedia of (and GitHub for) science**, as explained in this blog post. The .tex file was started by pasting below the leftovers from the drafting of the above blog post. We then turned this into LaTeX format and continue drafting that way on GitHub (cf. version history; **help available**).
- Unless something is clearly marked as being imported from elsewhere, all of this text is licensed CC0/Public Domain, while the LaTeX code is available under the LaTeX license, with the origin being the Elsevier article bundle for the .cls file and Copernicus for the .bst file .
- You can **get involved** (FAQ).
- **Submission of the proposal is anticipated for the end of July, 2011**.

The real abstract, or at least some potential phrasing for it:
One of the basic rights of society is universal access to knowledge. Information can be transformed into useful knowledge when it is accurate, up to date and freely available.

Scientific knowledge is at the core of social and community development, yet access to up-to date information is limited for those outside narrow academic circles. Even for those of us who do have access to the published information there are still limits on how we can reuse the published literature to maximise the social impact of scientific findings. These constraints result from restrictive formats and licensing terms that characterise traditional scientific publication systems.

Our ultimate goal is to place the existing openly licensed scientific literature in an environment in which it can become dynamic and where it can facilitate public discussion and outreach. We want these formats to be compatible with the needs to:

1. record research as it happens
2. review how it happened
3. review the interpretation of research results in the context of existing knowledge
4. identify gaps in knowledge, infrastructure or methodology
5. stimulate further research
6. stimulate public engagement
7. deliver useful outcomes in local and global communities

Key words:
Science as a wiki, GitHub for science, open access, Creative Commons, defragmentation of science, version control, digital encyclopaedia, digital collection, digital museum

---

Contents

1. To do

- Shorten the paragraphs in section 2 and merge it with section 3.
- Think of potential contributions to the July 16 Barcamp Auckland
- By mid-July: Think about potential projects for Summer of eResearch NZ (perhaps for testing the EoR prototype)

2. Introduction

See also section 3 for an alternative structuring of the introduction.

2.1. A brief history of research publishing as we know it

Scientific research consists of collaboratively exploring and pushing the boundaries of human knowledge through well-documented and contextualized sets of observations. Newly acquired knowledge is shared with the scientific and wider community through published reports, usually in scientific periodicals. In most areas of science, these reports take the form of 'complete stories' consisting of a relatively complete picture that emerges through descriptions of a series of experiments that complement each other. This mode of reporting prevents *new* data (and the knowledge that can be derived from it) from being made known until the entire set of experiments is completed. In some instances, the process of producing a 'complete' paper can take years.

There are several consequences of this ubiquitous practice:

1. While science knowledge within a research group grows in small continuous incremental steps, the community outside the research group does not gain access to that knowledge until the full gathering process for multiple data sets is completed.
2. Unaware of the existence of those intermediate steps, other research groups may end up duplicating the same or similar work.
3. Data generated in the process that may not be justifiably included in a formal scientific report (because it does not contribute to the overall 'story') does not get published and remains unknown to the wider community.
4. The burden of the costs of this orphaned data falls upon the funding agencies, and indirectly, upon taxpayers.
5. The narrative of scientific publications (especially as they continue to increase in number) becomes highly redundant, layering unnecessary burdens on the scientists: time is wasted re-writing text that in itself does not constitute new knowledge
6. The sequence of steps leading to a formal publication (writing, peer review, revisions, copy editing) further delays the availability of the work from its completion to the published state.
7. Once published, further modifications to the work that may be suitable due to new results cannot be incorporated into the existing work without going through an new cycle of publication.

Scientific publications in themselves become poor containers of knowledge because they do not reflect the granular nature of the scientific process nor do they allow further contributions to or enhancements of the original work.

Research publishing can be therefore best described as the business of providing low-resolution snapshots of these steadily evolving processes.

2.1.1. Attribution and Prestige

The prevailing structure of scientific publishing has seen few changes since its origins in the mid-17th century. At a time when communication between different parts of the globe was characterised by long time delays, scientific publishers offered the advantages of the printing press to distribute information between researchers. In 1665, the *Journal des Sçavans* started publishing scholarly articles, followed by the Royal Society of London's *Philosophical Transactions* a few months later (Spier 2002). Under the leadership of Henry Oldenburg, Philosophical Transaction made scientific findings broadly known and established a system (and culture) of attributing the new knowledge to its creators. In the words of (Guédon, 2001):

> In short, the Republic of Science claimed the right to grant intellectual property to scientific "authors" and Phil Trans was its instrument of choice.

It wouldn"t be long before a system of pre-publication peer review helped validate the published findings, adding an esteem factor to the authors of the articles (Spier 2002).

The second half of the twentieth century would see a new system of prestige associated with scientific publishing: the science citation index of Eugene Garfield (1955). The index allows individual articles to be "linked"to other articles that cited them and thereby provides a way of understanding how different articles are related to each other. Despite Garfield"s own warning about the possible abuse of the citation index (Garfield 1963), the system has nevertheless crept into academic assessment systems. Nowhere is the use of the indices more controversial than when the importance of an individual research article is associated with the impact factor of the journal where it is published. The alternative article level metrics, while preferred, still has the danger of using number of citations (impact) as a proxy for quality (importance) (Garfield 1963). Nonetheless, these metrics have become the yardstick by which institutions (e.g., through PBRF) and individuals (e.g., in staffing committees) are assessed and against which funding decisions are made (or at the very least influenced). It is not surprising then to see scientists"aspirations focused on the traditional article and the journal in which it is published.

2.1.2. The ever increasing body of literature

Since its early inception, the printed format has continued to offer widespread dissemination of results and, in the process, strengthened its role in the adjudication of prestige. As the assessment values shift from the data (the currency of science) to the paper (the currency of publishers) (Dobbs 2010, Mietchen et al. 2011) and as the number of research groups increase, so inevitably does the number of articles.

> The number of scientific papers, since about 1750, has multiplied tenfold every 50 years. The number of abstract journals has multiplied tenfold since 1880, every 30 years. The number of computer indexes to the scientific literature, since 1950, has increased tenfold in every ten years. Where will this exponential increase stop? (Glass, 1972)

The number of scientific publications has increased from the few hundreds published in 1774 to a projected 50 millions by the end of 2008 (Jinha, 2010). It is not surprising, then, to see a high degree of redundancy between different articles, in particular in the introduction and discussion sections (Mietchen et al 2011). This multiplication of efforts comes at the cost of time spent re- inventing the narrative wheel instead of producing original research.

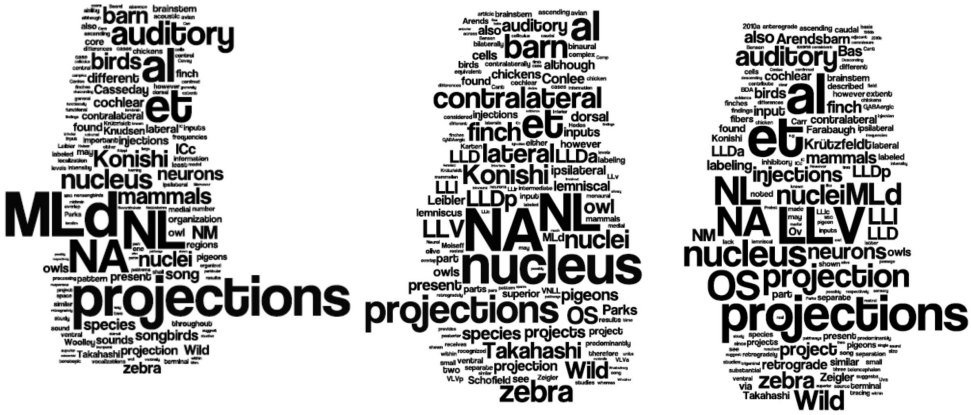

Figure 1: **Redundancy in journal articles.** A Wordle from three articles (Krützfeldt et al 2010a, b and Wild et al 2010) showing the occurrence of individual words in the introduction and discussion sections

Multiple and sometimes simultaneous discoveries characterise science (Merton 1957, 1963). In a climate where publishing in "high impact factor"journals is a priority – and where these journals only publish "novel"findings on a first-come-first-serve basis – researchers find themselves trying to balance the ethos of

scientific openness and collaboration and the need for secrecy prior to publication. Some of these issues could be ameliorated by alternative publishing models that allow new data to be incorporated to new or to existing published artefacts that appropriately attribute the authors.

2.2. The situation now

The rate of increase in the number of groups dedicated to specific areas of research and the accelerated pace at which data can be acquired and analysed due to technological and computational advances will probably continue into the future. The burden of the peer review system on individual researchers (who provide this service free of charge and for which they receive little credit) is becoming increasingly heavier. Further, as research becomes more and more interdisciplinary it is increasingly unlikely that a single researcher would be able to critically examine the range of methodologies that are usually included in a single multidisciplinary research report.

The value of the traditional scientific article as the primary source of research dissemination must be brought into question. Researchers have historically been early and eager adopters of new technology, especially when it accelerates the gathering or analysis of data, that is, when it increases productivity. In contrast, the publishing system has remained virtually unchanged since its inception. Scientific publications have not shown much innovation aside from the opportunity to publish media (e.g., audio and video) that cannot by their very nature be published on paper. Fundamentally, how readers interact with the published article (and how the authors interact with the readers) remains virtually intact despite the new possibilities associated with web 2.0 tools and the rise of social networks.

2.2.1. What do we value in a journal?

Guédon (2001) described scientists as Dr Jekyll and Mr Hyde. As scientists-authors we are eager to submit our work to high-impact factor journals, but as scientists-readers complain when our libraries cannot afford subscribing to them. What do we then value in journals?

> "Scientists often care less about the journal title than the ability to track down quickly the full text of articles relevant to their interests. Increasingly, users view titles as merely part of hyperlinked 'content databases' made up of constellations of journal titles."Butler, 1999

I would love to meet a scientist that has never uttered the phrase "I cannot believe that got published in Nature"or "I cannot believe that could not get into Nature". As scientists-readers we are able to dissociate the value of the science within an article from the perceived value of the journal. If the value of our own work was not measured by the journal in which it is published, what would we value most in a journal?

Before the internet, searching through indexes such as the Current Contents index booklet was tedious and time-consuming and when obtaining a copy of an article required sending a "reprint request"postcard to the author (and hope it wouldn"t get lost in the mail). Publishing in high impact journals made a difference. The articles had a better chance of being discovered and the chances of the journal issues being placed in the "recent arrivals"section of the library were higher. In 1997 the MEDLINE database was made available via the internet through PubMed (Weiner 2007). Access to the indices from personal computers made it easy to find articles related to a topic, regardless of which journal they were published in. A quick email to an author today gets us a pdf copy of the article if this is not available through our libraries. Having lowered the discovery and access barriers to articles it is not clear what the impact factor of a journal provides to authors other than the rewards associated with assessment. And as long as this continues to be the case, scientists-authors and scientists-readers will continue to be Dr Jekyll and Mr Hyde.

In 1991, Paul Grinsparg created the Los Alamos physics archives, a place where scientists could deposit their work, and which became a primary resource in the field (Butler 1999). The success of this alternative form of making research results public reflects a community that the "work"over the "brand". But in many disciplines, self-archiving has not gained enough momentum to replace traditional publishing. The question is then, what attributes should a publications system need to have to attract both the scientist-writer and the scientist-reader? What added value should this new system provide?

The 21st century has seen a rise in the number of researchers that reflect upon whether science is best served by continuing with traditional publishing practices. How can we retain the value of the process of scientific communication and peer review but also take advantage of the technology that has the potential of accelerating (and enhancing) the way in which scientists communicate their results? We have successfully embraced the use of technology to accelerate the generation of data but lagged behind in adopting it to accelerate the communication of these results.

2.2.2. Can we bridge the gap between the cycle of scientific knowledge and the publication cycle?

A good understanding the cyclical and dynamic nature of science is crucial for those incorporating scientific findings into educational curricula or evidence-based policies. Scientific knowledge emerges through a series of cycles that begin at the conception of an idea and lead to communicating the results. At this point feedback from other scientists helps refine and re- define the validity and implications of the findings leading to a new iteration of the research cycle where these new possibilities are explored. Research is therefore a continuous series of cycles where new findings are first put into old contexts and then integrated into new ones through the emergence of new data. Yet the ways in which scientific findings are reported do not reflect this dynamic nature due to the static format of the traditional reporting media nor do they provide an efficient way by which to re- examine the results in the light of new findings. Imagine two groups simultaneously working on a particular research question using two different methodologies, each reaching contradictory conclusions, and both publishing their results at the same time. Unaware of each other"s work, the published reports would not cite each other nor consider the implications of each other"s contradictory results. Under the current publishing system, these two sets of data could not be easily (and publicly) integrated into a new interpretation unless a third publication is produced. As a result, published reports become static snapshots of a single research group"s set of experiments that sit isolated from useful discussions that have the potential of re-contextualising the work to generate new knowledge.

Openly licensed scientific literature has the potential of changing the way we interact with the scientific reports. Creative Commons licences that allow the content to be distributed and reused offer the potential to modify published reports and adapt them in light of new findings. Creative Commons licences (with the exception of those with an ND clause) allow interested parties to place original or modified versions of articles where specialists can comment on the methodologies or interpretation of the results and where the actual text of the article can be modified. The scientific article in this way can potentially become a living document that continues to evolve in parallel with the scientific field of knowledge.

2.2.3. What should the next move be?

Literature that is placed in open collaborative environments can be subjected to post-publication peer review and open discussions offered by experts. These discussions can be centred on the entire document or smaller elements within it.

The individual paper distributed by journals continues to mostly follow the format designed for print - even when journal also have online versions. Seldom does the online version add value: corrections cannot be made directly on the article, specialists identifying flaws with the methodology are unable to share their views through comments, and new interpretation of the data based on new findings cannot be incorporated into the original work. Errors (conceptual or factual) in the original text are often perpetuated in subsequent articles.

In 1995 the group led by Eric Kandel (now a Nobel Prize winner) published a paper in Science (Grant et al 1992). Their results were challenged by a separate group (Huerta et al 1996). A Google Scholar search shows that the original paper has been cited 724 times. Of these 724 citations, 533 occurred after1997, after the challenging article had been published, and there were 5 citations in 2011 at the time of writing this. In contrast, the challenging article by Huerta et al. has only been cited 22 times since it was published. The original paper was never retracted, and no opportunities to reconcile the two contradictory findings were provided by the publishers. This example highlights how challenged results continue to propagate in the literature and why a new way of interacting with the scientific literature is imperative.

"[“] when we assume that people are selfish we build systems that reward selfish behaviours" Clay Shirky (2010)

Unfortunately, as long as the assessment systems continue to use publication metrics as a proxy for quality these issues cannot be properly addressed. The ubiquitous linking of research quality to publication frameworks (and their rules) provides very few incentives for researchers to contribute to the scientific flows outside of the traditional formats since they do not immediately translate to peer-recognition and are therefore not part of the reward system.

2.2.4. What do we propose?

"When you get the right combination of motivations and incentives you change the way that people interact with each other in fairly fundamental ways" Clay Shirky (2010)

We propose that placing articles in online formats that have a way of tracking individual contributions will lead to a mode of interacting with the literature which will:

1. Prevent the dissemination of erroneous facts that arise from typos or omissions
2. Clarify methodological issues that may be inadequately or incompletely represented in the article
3. Challenge the validity of the articles through post-publication peer review
4. Allow the incorporation of orphaned data from other groups that may enhance or contribute to the published article
5. Incorporate new findings to the interpretation of the presented results (i.e., continuous updating)

Once the articles are placed in these dynamic formats they can become living documents that work in parallel with the scientific community. At the moment, any enhancement (or criticism) of any article continues to occur at lab meetings, journal clubs, etc. These online platforms would allow having these discussions in open forums that are blind to geographic or temporal barriers. The resulting discussions will form an integral part of the scientific work "the author and reader become reader-author and author-reader.

In a more general sense, these platforms have the potential of creating broader benefits. There are several uses for such aggregates that can potentially have a wider impact.

1. Health articles about specific regional diseases can be translated and delivered to the local communities where they will have their most impact
2. General articles can be enhanced by accompanying plain English summaries that can be used by science reporters and educators as OERs.
3. Museums could use these articles to link to them enhancing digital collections.

Knowledge is deeply rooted in context and collaboration. We propose that publication systems should be built so that they facilitate and encourage the expression of those attributes. A record of the context and collaboration is the centre stone of the preservation of the cultural heritage of science.

3. Scientific practice in light of the interactive web

An alternative structure for the introduction is outlined in this version. See also section 2 as well as this comparison of paper-based and wiki-based research publishing.

Scientific research consists of collaboratively exploring and pushing the boundaries of human knowledge through well-documented and contextualized sets of observations. New research results are shared with the scientific community, such that their interpretation can be reviewed in the context of existing scientific knowledge, infrastructure and methodologies as well as of gaps therein. This interaction ultimately leads to new research.

Information is shared within the cultural contexts of our global society. Over the last two decades sharing has moved almost entirely to the web. What are the key aspects of communicating research, and how could these be – or already are – affected by taking place on the interactive web.

3.1. Context of existing knowledge

Relevant for all steps of the research cycle, though typically neglected when actually generating data. Currently distributed over a myriad of articles in thousands of journals that make it practically impossible to keep track of relevant developments outside a very narrow area, and very hard to find out whether some specific research question has already been addressed. Plus access barriers, synonyms and inconsistent usage of terms.

3.1.1. Duplication of effort

Duplication of research efforts may happen because: (a) Different groups may not be aware of each other's work. Science communication is characterized by inefficient ways of identifying who works on what, and secrecy is the norm rather than the exception in most fields of research. (b) Even when the individual teams are aware of each other's activities, the highly competitive environment does not provide sufficient incentives to share on the final attribution of the work. Instead, researchers tend to increase their secretive practices, independently invest in developing similar methods. Those who would prefer to find synergies and share on the work often find disciplinary and national boundaries associated with funding that can discourage - or prevent - more open and collaborative attitudes.

perhaps worth mentioning, if only for some of the numbers involved: this US Republican report has a section on duplication of effort from the side of US funding agencies

The following subsections are currently just outlines of what could go in there. The order is also likely to change. Only some parts of section 2 will fit in.

3.2. Ethics

Typically dealt with at an institutional level, behind closed doors. Different jurisdictions have different subtle requirements. Reported to the public more or less in a binary fashion (has been approved by the IRB).

- informed consent
- privacy of certain kinds of data (personal information, location of endangered species etc.)
- animal experiments
- outright fraud or data manipulation

3.3. Materials and instrumentation

Hard but not impossible to share. Example: Personal Genome Project, remote-controlled telescopes. The australian

3.4. Budget

Seriously under review as part of the funding application process, little oversight during the research, some review following reports to funders, no review from the wider scientific community.

3.5. Protocols

Can be easily shared, but strong links to ethics, materials and instrumentation usually imply that different variants will be in use, even though there might be just a single formal publication.

3.6. Code

Standard is to report algorithms in mathematical notation or pseudocode, though making code available is more and more popular, and indeed a requirement with some journals or conferences.

Typical problems: Cross-platform interoperability, web-based version, long-term maintainability, sometimes need institutional access to institutional servers (hard to get access to my server if you are not in my uni, for eg.), software specific formats (e.g., in microscopy, when you go from the 'microscope' software stack to a more open stack you lose the metadata)

3.7. Data

Typically, a publication only provides a summary of the data acquired, and not necessarily of all of it - negative results in particular remain underreported.

3.8. Peer review

In the present system, it acts at three levels - grant peer review decides whether a project is to be funded (though decisions are editorial by the committee - also different grants reviewed by different people, no fair way to compare the given ranks) - this is also behind closed doors because you cannot release the details of a grant that was not funded, and even for those that were funded only the general public summary is released), journal peer review whether an article is to be published, and post-publication peer review (and citations, use and derived works) determines the long-term value and impact of the research reported in the article. (Garfield of SCI fame says to be careful that impact (citations) says very little about value - citations do not contain the context of how the work is cited).

Peer review takes place very rarely when it matters most - when the details of the research are being determined, controls selected and so forth. But that is when it would be most helpful to have insightful commentary (cf. Polymath project). This only happens when a paper is being submitted, imposing further delays to the communication of the results.

Similar for data analysis and interpretation.

Pre-publication peer review is slow.

3.9. Outreach

Currently limited to "scientists say" stories - could be "let's see how they are trying to find out" stories instead, or in addition.

See also Why we watch reality TV

3.10. Reputation

Reputation of a researcher is currently mainly determined on the basis of the articles they have published (how many, in which journals, oh and about what topic, using which methodology, with what results?) and the amount of grant money they have been awarded (typically on the basis of their article-based reputation). This is an unfair system - the more you publish the more money you get the more you publish and so on - It hinders innovation because 'new ideas' are hard to fund unless someone powerful is happy to integrate them into their work (this will be difficult if the new ideas challenge those of the powerful groups).

Incidentally, these are the two steps in the research cycle that have traditionally involved formal peer review. Wouldn't it thus make sense to assume that allowing for peer review at additional steps in the cycle would provide for additional reputation systems to develop?

This way, the reviewed would gain new dimensions of reputation, and their research would get better if feedback is suitably structured.

3.11. Discoverability

Even a very comprehensive search of the formally published literature does not provide a guarantee that all relevant research (even that part that eventually ended up published) will be found.

An encyclopaedic structure (with redirects for synonyms, and disambiguation for homonyms) has discoverability built in. Semantic enhancements would increase it.

3.12. Discourse

As #arseniclife has demonstrated, blogs can be much faster and way more efficient than traditional discourse See also the 123 steps to publishing a comment (or counterexample)

3.13. Notes

- record research as it happens
- review how it happened
- review the interpretation of research results in the context of existing knowledge
- identify gaps in knowledge, infrastructure or methodology
- stimulate further research
- stimulate public engagement
- deliver useful outcomes in local and global communities

Science in the age of the interactive web. Every step within the research cycle can in principle be published, not just the "final" result.

The project aims at a bit of all of these:

- Encyclopaedia of original research
- Science as a wiki - i.e. collaboratively updatable database of interlinked articles
- GitHub for Science - distributed version control (recent example: E. coli O104:H4 analyses on GitHub); problem to solve: WISIWYG for Git (otherwise wider adoption unlikely)
- CC-BY - reusably licensed; forkability
- Lab notebook - science as it happens, rather than "scientists found out"

Perhaps best to use the above shorthands as titles for the sections of the aims and goals section? Probably not - they are not aims in themselves, tehy all seem to be the same aim really. IT seems more appropriate in the definitions of 'desired' and 'needed' features

Encyclopaedias.

- Encyclopaedia Britannica, Wikipedia, Scholarpedia
- review articles

Lab notebooks.

- OpenWetWare

Version control.

- Centralized version control (SVN, CVS)
- Distributed version control (GitHub, Mercurial, Bazaar)

Collaboratively editable databases.

- most wikis, especially Wikipedia
- many scientific databases, like GeneBank

OA-to-wiki export.

- images and taxon treatments from ZooKeys/ Plazi to Species ID and Wikispecies
- Images from PMC OA subset already being uploaded to Figshare (recently reviewed by (Singh, 2011))

Semantic enhancement.

- Semantic MediaWiki

Reputation system.

- StackOverflow

Notes. Room for "fishing expeditions" (data-driven research not directed at particular hypotheses; cf. (Botstein, 2010)) in addition to hypothesis-driven research

4. Aims, goals and objectives

See also the draft over at Species ID
Use SMART(ER) approach: Specific/ Measurable/ Agreed/ Realistic/ Time constrained.
See also the funding criteria at the Gates Foundation.

4.1. Turning science into a wiki to make research communication more efficient

We want to render scientific information more efficient by adapting it to the age of the Web. Specifically, we want to explore the potential of openly licensed scientific information to provide a basis for systematic reuse both within and beyond research contexts.

4.1.1. Encyclopaedic structuring of knowledge instead of flood of journal articles

"So what would you rather have?something checked by three experts over six months to a year, or something checked by 1,727 people in the first 100 hours? [..] Also remember that the refereed journal article is fixed at a moment in time, and beyond that any errors or new developments aren?t included." - see also this comment
early mention of "science as a wiki"

4.1.2. Collaborative updatability

4.1.3. Forkability

4.1.4. Contextualization of research findings

Paves the way for **Journal of Research Proposals** (demo) and **Journal of Science Contests**

4.1.5. Semantic enhancements

4.1.6. Reputation schemes compatible with collaboratively edited versioned documents

See this blog post

4.1.7. Major hurdles to overcome

Traditional publications do not reflect (nor have they room for) the continued updating and revising of the published material. This contradicts the nature of science where ideas and results are under constant revision and where interpretation of results are adaptable to new findings.

Furthermore, even if research communication would take place in a versioned encyclopaedic environment as envisaged here, integration of non-versioned legacy publications still provides a number of challenges, even if they are already digitized and their data made available.

4.2. Illustrating the potential of open licenses for reuse in new academic contexts

Re-use as a measure of impact "When PubMed Central has permission from the copyright holders, it makes articles libre OA and allows bulk downloading. It calls this the Open Access Subset of PMC. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/ But only 10% of PMC belongs in the libre OA subset. The other 90% is gratis OA, not libre, and PMC is obliged by the rights-holders to block bulk downloading. (Thanks to PMC's Ed Sequeira for these details.) Because BioMed Central offers libre OA to its whole corpus, it can offer its whole corpus for bulk downloading. http://www.biomedcentral.com/info/about/datamining/ In this sense, libre OA removes custody barriers that gratis OA may leave in place. The difference between gratis and libre OA isn't limited to permission barriers; permission barriers can create downstream possession and custody barriers."

4.3. Illustrating use cases of open scientific information beyond scholarly contexts

4.3.1. Health on a stick – medical information for rural areas in the developing world

Figure 2: **Sugar on a stick. Left**: The One Laptop Per Child project (OLPC) project provides children in the developing world with computers, so as to support their learning with Open Educational Resources and to prepare them for a life in an interconnected world. Source: http://wiki.laptop.org/go/File:Bucolico-full.jpg. License: CC-BY-SA. **Right**: Sugar – the Fedora-based operating system specifically designed for OLPC computers – is available on USB keys to facilitate development, testing and outreach. Source: http://wiki.sugarlabs.org/go/File:SugaronastickMirabelle.png. License: CC-BY. We want to create a variant of such sticks that hosts medical information distilled from the Encyclopaedia of original research in a way compatible with inclusion in OLPC deployments.

This subproject is concerned with medical information contained in the Encyclopaedia of original research, which shall be distributed to rural areas in the developing world.

The word "encyclopaedia" originally referred to a collection of contemporary knowledge curated for the same purpose that XO laptops were designed for: educating children. We want to add a spin to this classical educational paradigm and use XO deployments with Health-on-a-stick versions of the Encyclopaedia of original research to reach out across age groups in remote populations.

Further thoughts:

- openZim for slicing up wikis for offline use; see also Wikipedia's WikiProject Wikislice
- somewhat similar functionality on Sugar: Info Slicer. See also Make your Own Sugar Activities handbook (CC-BY-SA).
- distributed version control via WikipediaFs (on something like GitHub)? The documentation states "For example, it would be possible to use WikipediaFS to perform a massive content migration from an existing site to a Mediawiki". See also other tools for using Wikipedia offline
- language issues

- WikiPack
- mention How Khan Academy Can Help OLPC (perhaps in conjunction with Teach A Class) and Hesperian Foundation, which already supported Where there is No Doctor
- "Patients participate" project to provide synopses of medical research papers
- "I suggest authors must submit for review, and scientific societies be obliged to publish two versions of every journal. One would be the standard journal in scientific English for their scientific club. The second would be a parallel open-access summary translation into plain English of the relevance and significance of each paper for everyone else. "

Previous (manual) slicing of Wikipedia for educational purposes: 2008/9 Wikipedia Selection for schools - too much effort to keep doing manually

Potential side project: FigShare/ GBIF-IPT for OLPC. Here, kids could upload images and sensor data as they explore their environment; sample case (summary) for possible scholarly reuse; biodiversity data could be ideal. Also relevant for #altmetrics.

MapUganda - OLPC sensor use for mapping Uganda

4.3.2. Museums of the future

See also Museum3 and Museum digital

4.4. Documenting in public the process of writing a grant proposal

4.4.1. Collaborative drafting

4.4.2. Feedback from the public

5. Timeline

5.1. Notes

- Discussion of Gantt chart tools
- Runway

6. Sustainability of the project

6.1. Sustainability of content

Key aspect: open licensing; starting with CC-BY. Also important: content curation.

6.2. Sustainability of code

Key aspect: open licensing; starting with the GNU General Public License (GPL) under which Mediawiki is available.

6.3. Sustainability of platform

the example of FigShare

6.4. Sustainability of proposal

The whole proposal has been drafted in public, so as to provide an example anyone can use to learn or teach about grant writing, to invite others to come up with similar proposals, to test the potential for public pre-submission peer review, and to stimulate the debate about doing science in the open.

7. Project team

7.1. Applicants

7.2. Partners

We have not defined any formal partnerships yet, but the following are amongst those we are considering:

- enspiral - partner for software development
- polar bear farm UI for editing from mobile devices (to lower adoption barriers)
- PubMed Central - content partner for seeding of the platform with content

8. Description of work

8.1. Work packages (subtasks)

8.2. Timeline

8.3. Deliverables

9. Resource requirements

Estimating about 60K articles available as 'libre' - (about 4

10. Budget

11. Acknowledgements

People
Anyone who helped in one way or another

Tools
LaTeX, GitHub, Wikiversity, Species-ID, Google Docs, any other tools we used. Creative Commons, P2PU.

12. References

References

Botstein, D.: It's the Data !, Molecular Biology of the Cell, 21, 4 – 6, URL http://www.molbiolcell.org/cgi/content/abstract/21/1/4, 2010.

Guédon, J.: In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers, and the Control of Scientific Publishing, in: Creating the Digital Future: Association of Research Libraries 138th Annual Meeting, Toronto, Ontario (Canada), May, pp. 23–25, URL http://www.arl.org/resources/pubs/mmproceedings/138guedon.shtml, 2001.

Singh, J.: FigShare, Journal of Pharmacology and Pharmacotherapeutics, 2, 138, doi:10.4103/0976-500X.81919, URL http://dx.doi.org/10.4103/0976-500X.81919, 2011.

13. Figures

- Call for ideas
- three blind men
- the research cycle in a classical variant (publication as the last step) and a web one (continuous publication of the progress of a project)

14. Possibly useful quotes

See also this collection.

"Before the wiki revolution, each time science advanced a new generation would bring out a new generation of textbooks. With the wiki revolution the bits of your work that are superseded will be replaced, and as language changes other bits will be rephrased. But those of your words that are still valid for future generations are likely to be read long after other works have come out of copyright."

""Thirty years ago, you didn"t know what was going on in a different field and you did not have Google. It could take you months to figure out that an idea was a good or bad one. These days, you can get a good sense of that in a matter of minutes because information is much more accessible. That"s really, really huge. It makes it much easier to move from one field to another.""

Knowledge Blog: "So what could be more fitting than to revamp science through a platform explicitly built to be revised, commented on, and updated?"

Yesterday I asked one of my students if she knew what an encyclopedia is, and she said, Is it something like Wikipedia?

Jason Priem: "'How cool would it be to fork articles, a la Github"

Marcio van Muhlen: "We need a GitHub of Science."

15. Notes from earlier stages of the drafting process

*Nice piece by David Dobbs about Jonathan Eisen's attempts to publish his father's papers online - could be cited on several points, including Knowledge Blog ("So what could be more fitting than to revamp science through a platform explicitly built to be revised, commented on, and updated?") and the four essential functions of science that are currently wrapped up in the scientific paper: registration, certification, dissemination, and preservation

Also cites Antonio Panizzi and mentions ADNI and Mendeley ("Many of the metrics and connections between papers aren?t accessible on the desktop, presumably because they require the server?s data and processing power, and finding them on the web interface feels vaguely opaque.").

15.1. In focus: The encyclopaedia of original research

. merging research projects, and linking them with each other as well as with the concepts and methods behind them.

¡del¿and with any info about the giants, on whose shoulders they have been built¡/del¿

¡del¿*Copying, forking (e.g. of this draft), ¡/del¿ ¡del¿*Mention "Science as a wiki" (including blog repository) and Wikis in scholarly publishing and "Towards threaded publications" ¡/del¿ ¡del¿;possibly embed Larry Lessig's talk at CERN, 18 April 2011 ¡/del¿ ¡del¿:Lessig's talk is licenced under CC-BY; could be used to highlight issues of license stacking and reuse, also with respect to the default license of the EOR¡/del¿

*EOR: earlier version

:comment on the EoR being open and on it being a federation of wikis

15.1.1. Motivation

Including definition of goal. Follow SMART scheme.

15.1.2. Aims

15.2. Zooming in

Discuss what could become of the project ideas that won't end up in the final proposal, and how we plan to go about this decision.

15.3. Zooming out: Testing open vs. traditional science

*Do we need a ? (see also Panton Principles and Altmetrics manifesto)

*What online tools do scientists wish existed to facilitate their work?

:ORCID-coupled cross-platform reputation system

:see also this Nature News piece

15.4. Notes

15.4.1. Quotes

"See also Collection of "science as a wiki" quotes."

- Sandra Bajjalieh: "All of these issues, including the trend towards judging scientists on where they publish instead of what they publish, would be solved if NSF/NIH provided and serviced a highly searchable website onto which people posted results as they obtained them. Search engine capabilities make this entirely feasible. The following features would make the system far superior to the current one of publishing in journals. 1. The comments of interested readers would be added to the posting. Thus there would be peer review. 2. Additional data and revisions that respond to comments could be added. 3. Entry time stamps would solve any issues of priority. 4. The number of "hits" and downloads a link got (similar to the information PLoS One provides for each paper) could serve as a measure of it's interest. This solution is so obvious and the benefits so numerous (NIH program directors would have current updates of research progress, no more publication costs, the ability to imbed movies and animations....)that it's really difficult to understand why there hasn't been more of a move to implement it. Do we, as a community, really want a few people regulating the flow of scientific information?" (Sandra Bajjalieh)

- Paulo Freire: "At the point of encounter there are neither utter ignorance nor perfect sages; there are only people who are attempting, together, to learn more than they now know."

- "The internet allows for a much more powerful system than the current journal system, much more powerful than even an open journal system. Some things I'd like to see in a unified online open system

  - Hyperlinking between papers
  - Discussion threads for papers
  - Collaborative mark ups of papers, so that difficult papers can be communally dissected and fleshed out, or so that students can work through a paper and provide a mark up to ease the reading for other students
  - An ongoing wiki for every subfield, detailing current outstanding problems, papers to read to get up to speed, most recent progress, etc, as well as curating accepted knowledge. Wikis should also be able to be marked up by students, so that difficult material can be broken down and fleshed out for the sake of other students."

  TheEzEzz

- Larry Lessig's talk at CERN, 18 April 2011: http://vimeo.com/22633948 (Lessig's talk is licenced under CC-BY) :"copyright is a regulation by the state intended to change a regulation by the market; it's an exclusive right, it's a monopoly right, a property right granted by the state which is necessary to solve an inevitable market failure." ... in a more colloquial nutshell by Alex Pasternack: Copyright isn't just hurting creativity: it's killing science :: Notes: If we get above the din of this battle is that both sides agree that copyright is necessary for creative works - There is a place for sensible copyright policy but, however, not only artists rely upon copyright. Publishers do too rely upon copyright - the economic problem for publishers is different from that for the artists. We've been fighting a battle where copyright is essential but not on science where copyright is not essential. There is a trouble that few see - How accessible is information for the public? What does it mean for info to be available on

the internet? It is only freely accessible if you are part of the 'elite'. Here copyright is placed to benefit the publishers - not the authors - no author has a business model that is built around profiting from this copyright. Does this limitation serve any of the purposes of copyright? What is the publishers objective? To disseminate knowledge or to profit from it? ::JSTOR archive: has become increasingly criticized because of the cost involved in accessing the articles in the archive. ::Lessig asks: Can we do better? ::Open access self archiving movement ::Open access publishing movement: ::Some open is free (as in free speech) some open access is free as in you dont pay for it but other copyright rules apply. ::Science Commons: "broader strategy for producing the information architecture that science needs" as per the 4 principles of Open Science (check on site).

- Read write creativity / read-write communities
- "Sharing is at the core of the architecture of the net" ::Note by Claudia Koltzenburg: [by whom?] Barbara van Schewick points out: On the Internet architecture level, due to corporatism, "enclosures" are rampant. The principle of network neutrality that characterized the Internet in its beginning, thirty years ago, has been put at risk particularly by profit-making interests of network providers, observes Barabara van Schewick. The effects of this amount to what economists who think in terms of traditional market economy would call a "market failure". Van Schewick holds that we (and the regulators) need to protect the factors that allowed widespread application innovation in the past (modularity, layering and the end-to-end arguments). These factors made for the openness at the core of the Internet until the early 1990s. Van Schewick recommends let users choose, and practice as much 'application agnosticism' as possible. Internet users today are mostly controlled by flatrate offers and application bundles that leave no alternatives to choose from openly. Van Schewick's argument says that users should indeed be allowed to get a sense of how much they need for what they want to do on the internet ??? and yet maintain a predictability of one's bills. – see Barbara van Schewick. Introduction. In: Internet Architecture and Innovation. Cambridge, Massachusetts/ London, England: MIT Press, 2010, 1-15. (see also Internet Architecture and Innovation) – """in this vein, what is the "flatrate" in academic/scholarly/scientific publishing today that lures into control?""' – Claudia Koltzenburg 12:12, 1 May 2011 (CEST)
- "In the academy [..] we need to recognise an ethical obligation [...] which is at the core of our mission which is universal access to knowledge." Entails: work needs to be free (this should be an ethical point) - We do not need (and should not practice) exclusivity about our work.
- models of access that block access except to a paying elite and discourages innovation.
- Dorothea Salo: "At the risk of sounding all commie and stuff: we work toward a collective openness, or we die off one by one as the business model sustaining us as well as publishers crumbles to bits."
- Douglas Rushkoff: As soon as a network is in the hands of policy makers and their funders, this network loses its power to effect change. His conclusion is: "Create new forms that exist beyond any authority's ability to grant them protection", The Next Net. 1 March 2011. Shareable - Sharing by Design.

  Vannevar Bush. As we may think. The Atlantic Monthly, 1945
- "There is a growing mountain of research. But there is increased evidence that we are being bogged down today as specialization extends. The investigator is staggered by the findings and conclusions of thousands of other workers - conclusions which he cannot find time to grasp, much less to remember, asthey appear. Yet specialization becomes increasingly necessary for progress, and the effort to bridge between disciplines is correspondingly superficial"
- " Professionally our methods of transmitting and reviewing the results of research are generations old and by now are totally inadequate for their purpose.:

- ". The summation of human experience us being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships."

- " A record, if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted"

- "Thus far we seem to be worse off than before - for we can enormously extend the record; yet even in its present bulk we can hardly consult it. This is a much larger matter than merely the extraction of data for the purposes of scientific research; it involves the entire process by which man profits by his inheritance of acquired knowledge."

15.4.2. Potential problems

- Barriers to expert participation

- *Giving credit is key, and if wiki contributions (or any other science 2.0 activities) would be recognized in academic career terms (#altmetrics; requires), scientists would be willing to reallocate their time accordingly

- "'What does "publishing" mean in a wiki context?"' The current use of the term "publishing" in itself can be taken as an illustration of how commercial codes and practices seeping into academic culture have not been counteracted successfully since the invention of the Web. In academic CVs, research output in print only (and in electronic but non-open access format) still figures as a "publication", even though the meaning of "to publish" as "make generally known" and "disseminate to the public" has seen fundamental and indeed groundbreaking changes with the Web as a publishing platform. Indeed, "the public" itself has changed fundamentally because today, a "publication" can be made accessible on the web "'without"' . Had academic institutions been more interested in the benefit offered by such opportunities, publishing openly would be much more widely accepted today. In this light, nothing should be claimed to be a "publication" any longer unless it is open, maybe even "'in the sense of Open publishing"': "Open publishing is a process of creating news or other content that is transparent to the readers. They can contribute a story and see it instantly appear in the pool of stories publicly available. Those stories are filtered as little as possible to help the readers find the stories they want. Readers can see editorial decisions being made by others. They can see how to get involved and help make editorial decisions. If they can think of a better way for the software to help shape editorial decisions, they can copy the software because it is free and change it and start their own site. If they want to redistribute the news, they can, preferably on an open publishing site."

15.4.3. Notes

- Wiki stats tools: Article-level traffic stats, Trending topics, Edit stats, Edit stats for new pages :See also MyADS in astronomy

- virtuelles Museum

- Museum fish MRI & Tierstimmenarchiv

- Micropayments for culture - similar for science

- using Google docs

- Open Science Games? Any equivalent to open vs. public peer review?

- eResearch talk by Mark Gahegan

- , via Twitter

- Filipe Cruz, in Skype chat of May 6, at 19:11 - superfabs: there are a few sites dedicated to harboring science papers and journals in digital free for download formats. would be nice to do a list of them atleast, to analyze and figure out how to better complement them? its similar work i think :Link provided a bit later: http://xdatelier.org/2010/12/11/open-access-repositories/
- More from that chat: Fabiana Kubke 19:20 @scann Ah, I see - I was thinking about that earlier today - how do I phrase this to say "as a first step we will do this in this area" - I thought concentrating on one specific (I was thinking Chagas would be a good candidate - I am more familiar with some implications)
- https://creativecommons.org/weblog/entry/23831
- Jamendo - company based on distributing CC-licensed music
- FigShare et al. as virtual museum?
- Community building in ecology
- Use case (from Wilbanks talk): Reverse Causal Reasoning
- Open science article in the Guardian
- "open-access repository for all research findings, which would let scientists log their hypotheses and methodologies before an experiment, and their results afterwards, regardless of outcome"
- Bruce Alberts "Our goal as teachers and educators should be to expose our students to the discovery process and to excite them about challenges at the frontiers of knowledge." (B. Alberts, "A Wakeup Call for Science Faculty", Cell, vol. 123, 2005, pp. 739-741. DOI:10.1016/j.cell.2005.11.014) Also: "Old habits die hard, and I have been disappointed to discover that this is especially true in academia.", from the same source
- Wiss-ki
- couple things to ORCID
- Increasing the impact and visibility of your research
- FirstMonday - interview with Linus Torvalds on motivations behind open-sourcing Linux
- real-time visualization of Wikipedia edits:Wikistream
- Candidate for article-level metrics: Wikipedia links Linkypedia
- 50 million scholarly articles have been published so far
- Kete - a collaborative platform worth a closer look
- COASPedia:FAQ
- OUP problem, OUP solution
- How Funders Practices Inuence the Future of Digital Resource
- Liquidizer - non-linear version control (on GitHub)
- Sakai on BeSTGRID
- Summer of eResearch NZ
- Strategic Reading, Ontologies, and the Future of Scientific Publishing – mentioned in Geoffrey Bilder's talk at e-publishing: expresses great enthusiasm for web-based science
- Note to self: For basic help with GitHub, see http://help.github.com/git-cheat-sheets/ .

16. Potential funding schemes

Help with finding funders: Foundation Center (US-centric; examples)

16.1. Calls for proposals

- Ian Sullivan: http://grants.gov/ in the US has a comprehensive listing for all grant opportunities open at the federal level

16.2. Funders with good match in scope

- Shuttleworth Foundation - Fellowship scheme (deadline Nov 1, 2011; sample fellowship pitch video)
- MacArthur Foundation (co-funder of Encyclopedia of Life)
- Sloan Foundation (co-funder of Encyclopedia of Life)
- Gordon & Betty Moore Foundation (co-sponsor of Beyond the PDF meeting)
- http://www.skollfoundation.org/
- Howard Hughes Medical Institute (cf. How to fund research so that it generates insanely great ideas, not pretty good ones.)
- JISC (cf. Increasing the impact and visibility of your research)
- VolkswagenStiftung
- see also supporters of P2PU
- Human Frontiers Science Program
- Hesperian Foundation, especially relevant to section 4.3.1 (Health on a stick), as they already supported Where there is No Doctor
- Ewing Marion Kauffman Foundation ("our researchers must determine what we know, commit to finding the answers to what we don"t, and then apply that knowledge")
- Qatar Foundation?

16.3. Prizes and competitions

- NIH reuse app challenge

16.4. Microfinancing

Invite crowd-sourcing, with link to a description of a project already funded by that source and ideally with some overlap to the current proposal. What role do funders have to play in bringing research into the web age?

- Startl - startup support for socially responsible businesses
- Bitcoin; option for mining
- KulturWertMark
- Kickstarter - could the subprojects perhaps be submitted there, or to similar places (indiegogo, or rockethub)?