Bottom-Up Creation of Open Scientific Knowledge

By Peter Murray-Rust et al

When OpenForum Academy sent out a call for its second book, we felt the need to contribute a piece on the enormous upwelling of openness in the scientific process. At the Open Knowledge Foundation, we had already published a chapter in the last book and felt this was a good opportunity to present some of our ideas and culture to a readership who would appreciate it.

'Open Science' is too big and multifaceted a term to be defined precisely. It covers at least the spectrum of materials, process, culture, formal specifications and activities. At the Open Knowledge Foundation (OKFN), we have many people interested in Open Science and have a dedicated working group (http://science.okfn.org/), blogs (http://science.okfn.org/blog/all-blog-posts/) and mailing list (http://lists.okfn.org/mailman/listinfo/open-science).

Rather than try to summarise it, we brought together stories under the umbrella of 'bottom-up Open Science'. Several people volunteered and we have included everyone who contributed. We have discussed how free and open source software can make a difference in science making through the enlightening example of the Blue Obelisk, have approached the 'Quantified Self' movement, have addressed libraries as more than ever needed tool for knowledge discovery and organisation, and have summed up the whole through the lenses of 'citizen science' thus proposing a new common denominator for open knowledge.

These stories are very varied but all have the core belief that individuals and small groups, working together, can make a

difference by changing ideas, setting up tools and content and -- most importantly -- by growing communities.

Bottom-up Open Chemistry – the Blue Obelisk

Chemical software and data is a major activity, almost certainly exceeding 1 billion USD per year. However, almost all of it is closed, represented mainly by domain-specific software companies and traditional STM publishers. This is often aggressively protected; when the NIH set up an Open[*] database of chemicals and compounds the American Chemical Society (ACS) lobbied politically to have this curtailed and threatened Wikipedia with legal action for publishing the widely used CAS identifiers for chemicals. A major software producer will take legal action against licensees who publish program output, including bugs.

A number of independent, often unfunded, chemical hacker activities grew up during the 1990's and by 2000 a handful of codes were available but there was little continuity or coordination. We used to meet occasionally at ACS meetings and in 2006 we met in a bar near the large Blue Obelisk in Horton Plaza , San Diego. We felt that we had a consensus of philosophy, that the world undervalued our software and that we had the potential to change the future. We then agreed to loosely coordinate (not pool) our efforts. I suggested the name "Blue Obelisk" and our mantra ODOSOS – "Open data, Open Standards, Open Source ". To support this we created a Wiki, a mailing list and agreed to meet for dinner whenever we had a critical mass. There is no budget, no membership, no formal mechanisms – the mantra is our collective and very powerful DNA.

This has proved extremely successful and might work in other disciplines. We have about twenty projects which are happy to be counted as Blue Obelisk (http://en.wikipedia.org/wiki/Blue_Obelisk) and which fit into our criteria of ODOSOS. Our dinners are open to all – and closed source providers have attended and been relaxed. In 2007 we published a paper outlining our components. Recently we reviewed this in a 2011 paper with about 20 groups as authors.

When someone or organisation does something meritorious (normally an identifiable software product or data resource) I award a quartz Blue Obelisk (remarkably these are common and inexpensive). These loose traditions work. We now have software components in most of the chemical infrastructure for pharmaceuticals and increasingly in materials. The biggest problem is data – chemists do not publish machine computable data (though they should), instead embedding a subset in formal, subscription-access publications. We have machine extraction software but risk being prosecuted for extracting data.

Governance is minimal and we have been blessedly spared from either factionalism or imperialism. Each project is self-contained but uses other Blue Obelisk libraries where possible or more recently runs them as web services. The main language is Java, followed by Python and C(++) – with some historical FORTRAN. There is generally a leader to each project and while the Benevolent Dictator for Life (BDFL) occurs the commonest is "Doctor Who", where the Doctor hands on to a successor at irregular intervals.

Originally dismissed as cranks, we are now taken seriously. Companies such as Kitware, NY and Chemical Computing Group contribute significant amounts of code and as importantly, the critical mass of internal and external confidence. National labs (e.g. Pacific Northwest National Laboratory in US) have been

awarded a Blue Obelisk for collaborating on Open Source. We know that our code is widely used in pharmaceutical companies but we have few metrics on this usage, which is a common problem of Open Source in secretive industries.

As with all volunteer Open Source projects we do not have clear timelines, but progress over the last 5 years has been very good. It's possible to find high-quality components in most subdomains, including unit and regression testing.

The main problems we face are that chemistry (surprisingly) often does not engineer its own solutions but prefers to buy them. This puts a value on shrink-wrapping and hand-held maintenance which gratis Open Source cannot easily provide. Academics producing new code often get little credit and it's worse when they re-engineer existing solutions, even when the result is markedly superior. It's also difficult to get funding ("it's a solved problem"). The fragmented nature of the commercial domain makes semantic interoperability very difficult –companies protect legacy walled garden approaches. The internal messes created by unvalidated variants of legacy files in the pharma industry (e.g. when the result of a merger requires data reconciliation) has probably cost well over 100 million dollars in human effort, while the Blue Obelisk could have provided common semantics.

However I think we are approaching a breakthrough. Chemical software has made few objective advances in the last 10-15 years such that we have now implemented most of the major algorithms as Open Source. For an organisation which takes a responsible view of costs and values innovation, the Blue Obelisk can be an attractive part of a solution.

Sample Size of One

By Bastian Grashake

The Quantified Self (QS) movement is a community of people who perform bottom-up citizen science every day. Many participants of the QS community meticulously collect different kinds of data about themselves: dietary composition, calorie intake, physical exercises, sleep habits, even dreams. More recently, metabolites, genetic variations and the composition of their bacterial communities – metagenomes - have become the subject of their self-surveillance. Such strict monitoring may seem strange and too cumbersome to be performed outside the realms of professional top athletes. However, recent technological advances such as the emergence of wearable consumer-oriented activity trackers (recording, for example, the number of steps the wearer has taken over time) and sleep trackers (monitoring the stages of sleep), combined with the rise of direct-to-consumer (DTC) genetic testing now offer easy ways to collect larger quantities of data about oneself.

It is not mere narcissism, but curiosity and desire to understand one's own body that drives involvement to the QS movement. Which workouts bring the effects I'm looking for? How does my diet not only influence my weight but maybe also my mood? Which drugs work best or have the least side-effects? And such data may also be used to ask more obscure questions: What effect has a shared bed on my sleep quality?[26] How does my butter intake influence my math-skills?[27]

---

[26] http://gedankenstuecke.github.com/blog/2012/09/26/on-getting-sleep/

[27] http://quantifiedself.com/2011/12/butter-and-arithmetic-how-much-butter/

QS participants thus use their data to perform experiments. By design, these are done unblinded and their sample size of one is as small as possible. These features may appear as an obstacle. These experiments, however, are performed by people who deeply care about the questions on hand. While most of these experiments probably will never enter the canon of peer-reviewed science, they are not doomed to fade away unnoticed. Many participants of the QS movement are out-going, and share their experiences and results. They write about their results in blogs such as QuantifiedSelf.com and meet for yearly conferences in the US and Europe.[28] Many cities worldwide now host monthly QS meetings where people share their practices and results by answering three questions: What did you do? How did you do it? What did you learn? This exchange – both online and offline – inspires community members to try similar approaches, to reproduce earlier results, or to modify experiments according to their needs.

Many people who are active in the QS movement are also openly sharing their data with others, thus allowing for experiments that overcome the limitations of the sample size of one. One of the most famous examples of this approach is the study on the effects of lithium carbonate in patients with amyotrophic lateral sclerosis (ALS), performed by users of the PatientsLikeMe community. In 2008, academic researchers published a study suggesting that regular intake of lithium carbonate could be used to slow down the progress of ALS, a currently incurable disease.[29] Following these observations, members of PatientsLikeMe started the off-label use of lithium

---

[28] http://quantifiedself.com/conference/Amsterdam-2013/

[29] http://www.pnas.org/content/105/6/2052.long

carbonate aiming to slow down their ALS. By comparing the disease progression of users who did and did not take lithium carbonate, the PatientsLikeMe data showed that the drug is not effective in slowing down the disease.[30]

Thus, a highly motivated group of patients using the Internet allowed a clinical study to be performed at a low cost and largely outside of academia. Following this success, similar projects have burgeoned: Genomera,[31] a San Francisco-based start-up, offers a community dedicated to the idea of small-scale studies. Users can create new studies, give input on the experimental design and enrol as participants. The non-commercial openSNP project[32][33] is similarly interested in genetics. Users can upload their genetic data along with further physiological details about themselves. For example, daily step counts or weight as collected by QS sensors can be provided, as can information about hair colour and diseases using text fields and images. The goal is to create an open database (the data being released under a Creative Commons Zero waiver) that can be used for studies seeking new associations of genetic variations with diseases and traits.

From people collecting seemingly unimportant data to real studies with medical significance: many participants of the QS movement are already performing science. They share results and data, replicate earlier findings and organise conferences. Most of this is done outside of academia, unfunded and without any central organisation. The collaborations of those highly motivated

---

[30] http://www.nature.com/nbt/journal/v29/n5/full/nbt.1837.html

[31] http://genomera.com/OLIJHOEK

[32] https://opensnp.org

[33] Disclosure: Bastian Greshake is a developer of openSNP.

science amateurs show how science can be performed in a bottom-up fashion and how they can complement research performed in academia and industry. With more and more people joining the QS movement, with new ways of sharing data, knowledge and insights won through self-tracking, and by collaborating the impact of their efforts will rise in the future.

A new role for libraries in open access information management

By Tom Olihjoek

The dissemination of knowledge on a large scale only became possible through the distribution of books and journals by publishers among a growing group of (highly) educated people. Prior to the introduction of the Internet in the 1990s, publishers had built up a monopoly on the production and distribution of knowledge through printed scientific journals and books. Publishers were justifying the ever increasing costs of subscriptions to scientific journals by the increased production and distribution costs. Scientists and research institutions had no choice but to pay.

After the Internet became popular, modern digital reproduction and distribution made these costs almost negligible. The publishers, however, have continued to increase their prices and to shield most publications from free access online. Consequently, scientists, institutes and other knowledge seekers still pay large sums to publishers for a now basically redundant service. Moreover, these developments have forced libraries to drastically cut-back on their subscriptions to scientific journals. The role of libraries is seemingly also undermined by an increasing number of scientists who read and publish work in open access journals, which can be accessed without library

subscriptions. Libraries thus suffer from an 'identity crisis' as they are forced to re-assess their role as suppliers of information.

Noticeably, the past few years have shown a spectacular growth of the number of open access publications. As of April 2013, the Directory of Open Access Journals[34] lists over 8,000 journals publishing articles under the terms of permissive Creative Commons licenses. While the volume of information available online dramatically increases, the difficulty of finding relevant information in the resulting haystack becomes more pressing every day.

This 'information glut' creates a growing need to find ways of making accumulated knowledge easily accessible. Just because open access information is accessible does not mean that specific information is easy to find, let alone that the reliability of the information can be easily determined. While access may no longer be an issue, discoverability is. By discoverability I mean that information should be easy to find and that access sites should be easy to use. The problem with the combined knowledge accumulated on the internet is that information of interest to specific communities is often too scattered and fragmented to be useful for them. Anybody in need of specific information has to dig, find and curate an ever increasing number of sources. All this leads to a general feeling that the information is out there but is "too big to know."[35] Some scientists claim this is not a problem as all information, even scattered all over the internet, will always be easily found with the appropriate indexing and computer search algorithms. In my view, we would be putting too much trust in computers. Even if computers would function flawlessly all the

---

[34] Directory of Open Access Journals (DOAJ) http://doaj.org

[35] David Weinberger (2012) Too big to Know http://amzn.to/11Du4Fd

time (and we know this not to be true), there is always the risk of only finding information that companies or government bodies want you to find. The search algorithms referred to as "personalised search services" are very much on the rise,[36] but wouldn't it be far better to organise information according to topics ourselves?

At first sight, the classical library function of offering access to information may appear to become of lesser importance in a 100% open access situation, but I see a new role emerging where libraries and librarians will start to organise open access content in a way that the public and scientists can use it best.[37][38] One way of doing this would be for libraries to take on the task of organising information around topics. Thus, libraries could reclaim their central role in making information accessible, a role which they have had all along but which has become much more complex after the invention of the internet and the digital revolution that followed. One such initiative has already started: the Open Library of Humanities wants to be a platform for open access publications in the field of humanities.[39]

There are many advantages of organising information by topics. For one, the discoverability of the information would improve, second communities interested in the topic could collaborate with libraries to keep information up to date and third

---

[36] Eli Pariser (2012) The filter bubble. http://amzn.to/11DtR4Y

[37] Bjoern Brembs Blog: open access taking off: Visions2: http://bjoern.brembs.net/comment-n894.html

[38] What role do university librarians play in access to research? http://bit.ly/14qeNVW

[39] Open Library of Humanities http://www.openlibhums.org/

community efforts could shape information in such a way that everyone can find information on his or her level of understanding.

Collaboration between scientists, libraries and communities could be a first step in the creation of an Open Science society, where most science is not kept hidden behind toll-access bars, but has an active role in sharing knowledge between all people on the planet.

The rebirth of the citizen scientist

By Rayna Stamboliyska

In the recent decade, the term 'citizen science' has emerged to define public involvement in genuine research projects. Synonym labels such as 'crowd-sourced science,'

or 'networked science' actually represent a new make-up for an old idea: back in 1982, science theoretician Feyerabend advocated the "democratisation of science." Going more decades backwards in time, Thomas Jefferson used to envision[40] weather stations operated by volunteers as a means for people to be informed and educated thus engaging into self-governance, a dynamics that is currently happening for real.[41]

This Jeffersonian idea illustrates one of the basic and most crucial issues with science as it is currently performed (i.e., through research within official institutions): its isolation. Contrastingly, citizen science operates – by design – free of the constraints inherent to such strongly formalised places. Citizen

---

[40] http://blogs.scientificamerican.com/guest-blog/2012/07/03/life-liberty-and-the-pursuit-of-data/

[41] http://wxqa.com/

science thus not only relocates science, but it also fosters its growth in the mainstream of society. Non-professionals join professionals, thus co- creating knowledge that makes science an integral part of our daily lives and shared human culture.

Numerous examples can be quoted, each bringing its unique colour and shape to the picturesque landscape of citizen science: from birdwatchers illustrating[42] how times of nesting shift as a consequence of climate change to disaster management,[43] from mapping roadkill accidents[44] to producing one's fluorescent yoghurt at home.[45] These projects illustrate a shift in public engagement in science: from citizens being solely data collectors to data analysts, visualisers and generators of new hypotheses. The hacker and DIY movements have widely contributed to the emergence of a true citizen science, i.e. one that fully explores human curiosity in a non-professional context.

Citizen science is in its infancy yet its popularity grows exponentially as the concept is modular enough to reach the humanities and social sciences (HSS),[46] generally overlooked by both professionals from the so-called "hard" sciences, and citizens. HSS are studies of human nature at large. They

---

[42] http://nestwatch.org/

[43] http://www.huffingtonpost.com/w-david-stephenson/citizen-science-disaster-information_b_1321899.html

[44] http://www.huffingtonpost.com/2012/05/11/adventurers-scientists-for-conservation_n_1510048.html

[45] http://www.indiebiotech.com/?p=152

[46] http://blogs.plos.org/citizensci/2013/02/25/science-and-society-voices-from-the-humanities-and-social-sciences/

encounter the same issues as the “hard” sciences: popularisation and communication, policy questions, and a wide range of ethical concerns. Additionally and similarly, HSS have particular theoretical traditions, methodological orientations, and critical interests.

The recent surge of citizen science, greatly assisted by information and communication technologies, thus allows reconsideration of the somewhat artificial categorisations of science domains and naturally involves trans- and interdisciplinary in scientific practise.

These considerations indicate that one does not need a ten-person lab, multimillion-dollar grants and caffeine-intoxicated PhDs in order to perform brilliant science. Citizen systems of participation aimed at collective problem-solving bring, however, two crucial questions: Is citizen science capable of producing reliable data? What guarantees do we have that it is ethical science?

Engaging huge numbers of citizens in a research project means that massive input is generated. Indeed, volunteers already collect data for scientific projects: how reliable is this? Two decades ago, the USA introduced an amendment prohibiting volunteer-collected data to be used in the US National Biological Survey. In the case of a community-based bird species diversity survey, the estimated number of birds correlated with the changes in numbers of observers. Such examples contribute to a stigma associated with citizen science data, which is sometimes labelled 'incompetent' or 'biased.' In a recent piece, John Gollan argues[47] the opposite: “a growing body of literature shows that data collected by citizens are comparable to those of professional scientists.” Although data-integrity issues can occur, Gollan

[47] http://blog.okfn.org/2013/01/23/citizen-science-can-produce-reliable-data/

highlights an important message: “it’s just a matter of honing in on those

particular issues and addressing them if necessary. This can be through training to improve skill sets or calibrating data where possible.”

The second question that springs to mind when opening scientific practiceto non-professionals is ethics. Many have voiced concerns[48] about dubious ethical frameworks in various citizen science projects. The project that caused recent kerfuffle was uBiome, a project to sequence human genome entirely supported through crowd-funding. Indeed, research ethics are not something to play with: thus, every project dealing with human subjects requires the review and approval of an independent committee – generally referred to as Institutional Review Board (IRB) – prior to its start. The uBiome citizen science project was thoroughly criticised for seeking IRB review of their protocols only after the crowd-funding campaign was completed. A similarly strict review framework is de rigour when a research project involves animal subjects. In a recent piece for Scientific American,[49] professional scientist and citizen science advocate Caren Cooper called for community answers to ethical questions as the boundary between hobby practitioners and citizen scientists is too blurry to be defined, and so are the cases in which participants need to be invited to follow official ethics

---

48 http://boundarylayerphysiology.com/2013/02/17/why-im-worried-about-ethical-shenanigans-in-the-citizen-science-movement/

49 http://blogs.scientificamerican.com/guest-blog/2013/03/05/animal-care-ethics-in-citizen-science-my-conundrum/

protocols. As also exemplified by numerous reactions from open and citizen science enthusiasts,[50] IRB approval can be a hurdle for citizen scientists.

Cooper's call-out to the community of both professional and citizen scientists does echo a widely shared concern:[51] is there someone – and if so, who? – to provide oversight of DIYbio/ citizen science practices? By design, both professional and citizen scientists need to urgently address this particular and foundational issue.

None of us can continue standing passive when a threat is posed to citizen science. It fosters our common culture of curiosity and bridges gaps between people whose personal aims and leisure-time activities converge on a desire to advance research and improve human welfare and communities.

---

[50] http://storify.com/PatrikD/is-irb-approval-a-significant-hurdle-for-diybio-pu

[51] http://scio13.wikispaces.com/Session+6B

Peter Murray-Rust is a contemporary chemist born in Guildford in 1941. He was educated at Bootham School and Balliol College, Oxford. After obtaining a Doctor of Philosophy he became lecturer in chemistry at the (new) University of Stirling and was first warden of Andrew Stewart Hall of Residence. In 1982 he moved to Glaxo Group Research at Greenford to head Molecular Graphics, Computational Chemistry and later protein structure determination. He was Professor of Pharmacy in the University of Nottingham from 1996-2000, setting up the Virtual School of Molecular Sciences. He is now Reader in Molecular Informatics at the University of Cambridge and Senior Research Fellow of Churchill College.

His research interests have involved the automated analysis of data in scientific publications, creation of virtual communities e.g. The Virtual School of Natural Sciences in the Globewide Network Academy and the Semantic Web. With Henry Rzepa he has extended this to chemistry through the development of Markup languages, especially Chemical Markup Language. He campaigns for Open Data, particularly in science, and is on the advisory board of the Open Knowledge Foundation and a co-author of the Panton Principles for Open scientific data. Together with a few other chemists he was a founder member of the Blue Obelisk movement in 2005.

In 2002, Peter Murray-Rust and his colleagues proposed an electronic repository for unpublished chemical data called the World Wide Molecular Matrix (WWMM). In January 2011 a symposium around his career and visions was organized, called Visions of a Semantic Molecular Future. In 2011 he and Henry Rzepa were joint recipients of the Herman Skolnik Award of the American Chemical Society.