Openness – a future challenge of informatics?

Vratislav Podzimek
Faculty of informatics,
Masaryk university,
Brno

10.06.2013

Introduction

When I was about to start working on this essay I searched for some study material for the topic I had originally chosen to cover – Minimum energy computation. However, often I end up on a page providing an abstract for some paper and a nice *"Purchase a PDF for 35 $"* button (for example [1] and [2]. Since it happened so many times and typically with the papers I was most interested in, I decided to switch to a different topic hoping for better luck, but, unfortunately, it turned out to be a naïve hope [3]. Disappointed by the approach of the science community or more probably the publishers of science works, I decided to write and essay on how important the openness is for informatics and science in general and which challenges it poses for the informatics as the driving force for all open resources, communities, ideas and collaborative efforts.

1 Open access, knowledge and data availability

We live in a world and age where science and technologies develop faster than ever before and even faster faster year by year. This, among the other things, means that there are many scientific data, papers and in general research material produced and processed every day which is only possible thanks to the huge development in the information processing and communication science (informatics) and technologies in the last two (20th and 21st) centuries. For example, the Large Hadron Collider (LHC) at CERN produces cca. 25 petabytes of "raw" data per year and this is only a thousandth

of the real raw data produced by the detectors that are filtered to remove "uninteresting" data [4]. These are numbers nobody imagined to be possible to handle few decades ago, but on the other hand they are at the edge of the current technical limitations and delivering followed by processing of the real raw data would be practically impossible (the rate of the real raw data is cca. 300 GB/s) [4]. The obvious (and sad) fact is that a huge amount of gathered data is definitely lost and thrown away. If, at some point, we find out that this data may have not been so "uninteresting" as they seemed to be, there will be no way to get them back apart from repeating the experiments and measurements (that still wouldn't provide the exact same data) which would be vastly expensive. So while it is thanks to the information science and technologies that the data produced by the LHC can be processed somehow, at the same time it is clear that better technologies and more processing power would allow getting more (and possibly better) results.

Another aspect is that not only the results are important, but so are the data, for multiple reasons. First of all, every scientific result has to be verifiable which would be hard without the input data available[1]. Another reason is that even a completely different research may use the same raw data to build on or to compare it with their measurements/data. As it is a generally accepted fact, one of the reasons behind the mankind's success is that we are usually able to build upon the knowledge of the previous generations through keeping that knowledge in a more or less permanent form, typically as books, letters or some other written form. Having access to the knowledge and wisdom of the previous generations is vital for the success of the mankind and in a non-negligible extent the same applies to the raw scientific data. However, this poses a lot more potential barricades that has to be overcome, starting from permanent storage and preservation of the data and ending with the accessibility of the data. For both these areas being so important (and in the same time problematic) there were or still are many initiatives and projects focusing on them, but the key idea is shared across them – it is the usage and further development of the information processing and communication technologies that is crucial for the success in those areas.

To mention a few examples of initiatives and projects focused on permanent storage and access to the (not only) scientific data, let's start with the Alliance for Permanent Access (APA) [5] and the two related related projects – PARSE.Insight (insight to the problematic of Permanent Access

[1]The correctness of the measured data is also crucial, but in some cases (e.g. data from the LHC) it is practically impossible (often because it would be vastly expensive) to repeat the experiment and gather new data.

to the Records of Science in Europe) [6] and APARSEN (Alliance for Permanent Access to the Records of Science in Europe Network) [7] both co-funded by the European Union. These two projects focus on permanent storage and availability of the scientific data. Another hot problem is that even the data is preserved somewhere and reachable, the Universal Resource Locator (URL) used to reference them often quickly become invalid or point to a completely different resource. The Permanent Universal Resource Locator (PURL) [8] is a project focusing on that area by proving a service for registration and administration of permanent URLs. A slightly different approach was chosen by the Digital Object Identifier (DOI) [9] system providing permanent identifiers instead of permanent locators. Then there is the Open Access initiative [10] targeting a rather ironic situation where typically a government funds some research and gives money to the researches to, among the other things, write and publish their papers, but in the end gives additional money to other researchers or academic institutions that want to have such papers available because they are published in science magazines subscriptions of which are very (and getting more and more) expensive. The motivations behind this initiative are greatly summed up in the "comics-style" video [11] and the basic goal is to encourage mainly the academic institutions to provide their science materials in an open and accessible way and to provide an infrastructure supporting such efforts. One of the important parts of this initiative is the usage of so-called the Creative Commons licenses [12] that on their own play an important in permanent and open access to (not only) scientific data by providing an easy way to choose a permissive license for a newly written text, paper, etc.

As the number of the projects (and I could add a lot more of them) focused on these areas show, the permanent and open access to the knowledge base we have already is a challenge for the informatics and will become a bigger challenge in future with growing amounts of data, numbers of users and so on. At the first glance it may seem there is no problem at all, but when putting some more efforts in it one can find out it is not a negligible issue. In contrast to many paintings, texts, letters, etc. that were preserved for hundreds or even thousands of years, we e.g. no longer have the original tape recordings of the first landing on the Moon (after less than 50 years) [13]. Some rumours also say that we already don't have some valuable data gathered in the LHC. And while we have notes about first telegraph message sent, archived written reports give us notion about when each city started to exist, we don't have the first Web page ever created (by Sir Tim Berners-Lee, the inventor of the World Wide Web), less than 20 years ago [14]! And what about the steps that lead to many inventions and discoveries? We have a lot of letters various mathematicians, physicists and other scientists sent to their

colleagues discussing their work and opinions. But do we have any emails, instant messaging conversations, phone calls etc. capturing the recent conversations of that kind? Almost none of them. Again, this may seem to be only a small problem because we still have the results of such scientific work, but there is a big difference between knowing the formula to compute a value of the sine function and knowing it is (and can be) derived from the Taylor's series. Because while the formula can be only used to get some value, the principles behind it can be used to derive many, many more formulas or even improve the principle itself. However, the interesting thing is that even the technologies we already have and understand may be successfully used to address the issues with preservation, preservation and accessibility of scientific data. So maybe it is, in some sense, a challenge for informatics to more focus on issues we already have next to the further development of better and more sophisticated technologies that would allow solving some of the issues, but that would add a lot of new issues as results of the new possibilities provided by such inventions.

2 Open source principles in many areas

Having the development as well as results of scientific work preserved and available is a crucial point for mankind's advancement in science and technologies. Nevertheless, it is only one of the building blocks needed. With the development in these two areas becoming faster and faster and even faster faster, the complexity of the principles and mechanisms used in them usually (there are some exceptions) grows with the same speed. As a result less and less people understand such principles and mechanisms and even less are able to contribute to their further development. Fortunately, there are more and more people contributing to science in general (in what used to be "developing countries" few years ago), so it is possible to keep and even improve the current pace. However, this is probably unsustainable and it may happen that the development will slow down due to the enormous complexity and few people able to contribute to it, especially if even the early and understandable phases of the research will stay available only to small teams of people located in the same place. It's again advancement of information processing and communication technologies that may provide a solution for such potential issue, a solution we already have available, but not (enough) "deployed" in this area – the open source philosophy, principles and maybe a way of life.

Open source, as evolved from its origins closely related to the emerge of the open source software, now works as a methodology in many areas of

human interest ranging from software development to car design, government, education and so on [15]. The truth is that open source development in all these areas leads to remarkable results and very innovative approaches. One, and probably the most noticeable example, for all – the Linux kernel and open source GNU/Linux operating systems (OSs) based on it. Such OSs run vast majority of Top 500 supercomputers, almost all of the biggest stocks and other mission critical systems and uncountably many devices of various types and purposes [16]. Also, without much exaggeration, one can say that open source systems (be it the ones based on the Linux kernel or the ones based on the BSD systems[2]) run the Internet [16], a crucial medium for the today's world. Another prove open source software being very successful is the fact that many big global companies providing very popular and highly-demanded services build upon it their systems (Facebook, Yahoo, Google, various post/delivery services and many others). Also the LHC (and the whole CERN) use the Scientific Linux [4], which is a GNU/Linux distribution derived from the Red Hat Enterprise Linux. And beside stability and reliability it is the openness and related ease of doing modifications and improvements tailored to needs of these companies and institutions that bring a huge advantage over the closed-source solutions.

So the achievements of the open source principles and methodologies are quite remarkable. But what actually is behind them? The main attribute of open (source) projects is that they are done in a highly collaborative way. And as the so-called "Linus's law" [17] states: *"given enough eyeballs, all bugs are shallow"* which, brought in a more general sense means that many people contributing to a project take care of each other's mistakes and overlooks. Moreover many people provide many opinions, suggestions and many points of view that can, when managed in the right way, lead to very innovative, original and practical solutions. And collaboration of people with various levels of knowledge, aims of interests and generally various professional and personal background is another advantage of the open work and environment. History of science has many times shown how important a "look from the outside" may be [18].

Though it may not be obvious from the first sight, it is again the development of information processing and communication technologies that allows such open and collaborative projects exist. Without an effective way of communication it would be impossible for a lot of people to share their ideas, knowledge and work together. Also without a possibility to check and test the progress of the work it would be hard to find and fight issues that appear and, for example a compilation of the Linux kernel needs quite a lot of com-

[2]FreeBSD, OpenBSD, NetBSD, ...

putational performance to be done in a practical time. However, it is again a challenge of informatics to provide better and better tools, mechanisms and in general environment for such open (source) and collaborative efforts, that will likely be needed, even with the enormous number of participants and contributors. A slightly different, but not less important, challenge is the inter-disciplinary work ranging from informatics and physics over social studies to law that will be needed to achieve such goals. Considering the low amount of inter-disciplinary studies, narrow focus of many (if not a majority) of researchers and issues that emerge with communication between two groups focusing on different areas (mostly if one area is some soft science and the other is a hard science) I'm afraid this will be a real challenge in no way smaller than the technical and theoretical advancement.

3 Will we be able to go on without openness?

With all the issues and suggestions presented in the previous sections, there probably is a basic question – Can we afford not supporting and not promoting the open, collaborative way of research, studying and generally science advancement? Maybe yes, but it may happen that the pace of the scientific and technological advancement will gradually slow down (and possibly completely die out) due to a need of gathering the same or similar data again and again, due to a bad preservation of knowledge and findings causing us to reinvent the wheel again and again or due to less and less people capable of understanding it and making contributions to it. I take it we are expected to have some sort of "computer-aided thinking" in the future and technologies beyond our needs, but even with that happening there still will be a need to have a lot of data as well as results of their processing available and shared to prevent enormous waste of resources when getting them again and again from scratch.

One practical example what the difference is – next to writing this essay, I may have been getting knowledge and gaining interest in the area of minimal energy computations, but instead of it I helped to improve a library for reading data about languages, territories, keyboard layouts, time zones etc. [19] by using a different approach of processing data, speeding the library up and lowering its memory consumption. Which one of the efforts would be more valuable is impossible to say. Maybe I would become so fascinated by the minimum energy computations that I would decide to study it further and maybe, some day, I would contribute to the field with some important finding. Or maybe not, maybe I'd leave it be with some basic notions learnt and the improvement to a library that can be used practically by anybody in

the world was the right and more valuable thing to do. But what makes the difference between going one or the either way? I'm fairly interested in both things, however, I don't want to spend 35 $ to purchase a paper I may later find not interesting or too hard to read and understand. Especially when I have a chance to make a contribution anybody in the world may freely use and build upon.

Maybe we can go on and keep up the pace without going the open way, but maybe (likely?) we cannot. However, the stakes are so high that we should really think it through and even for a pity's sake we should start encouraging scientists, researchers, engineers and all experts together with their students to go the open way. The possibilities to do so are definitely available and easily reachable. For example, what prevents people from the academical sector from contributing with their knowledge to the Wikipedia? It is just a lack of encouragement and motivation (also financial, of course) from their funders and colleagues. Almost every month I hear that the information from the Wikipedia shouldn't be much believed and how bad it is to cite those pages in works, but I have very rarely seen those people actually do something with it by correcting what they thing is wrong in some particular article. And the same applies to the computational resources many (not only) academic institutions have and waste with. Nothing prevents universities from connecting their computers and servers e.g. to the Berkeley Infrastructure for Open Network Computing[3] and participating in one of many grid computations. Well, again nothing apart from the funding and costs that would have to be covered, but that could be saved in many other projects, that require and consequently buy additional computational power even though we have a lot of it unused and wasted.

I see it as a big challenge of informatics to play its important and irreplaceable role in making sure we don't go the wrong way and don't overestimate our abilities in "playing in our own playground" and competing with others often in (at least) problematic manners. I believe the openness, collaboration and shared knowledge is the right way [21] to go and also the only sustainable way we can take to keep up the pace with our expectations for the future.

[3]the performance of which is, by the way, very similar to the performance of the best supercomputers [20]

References

[1] Nature Publishing Group. *Experimental verification of Landauer's principle linking information and thermodynamics: Nature: Nature Publishing Group* [online] 07.03.2012, [cited 09.06.2013]. Available at:

<http://www.nature.com/nature/journal/v483/n7388/full/nature10872.html>

[2] IBM. *IBM Journal of Research and Development* [online], [cited 09.06.2013]. Available at:

<http://www.research.ibm.com/journal/rd/441/landauerii.pdf>

[3] Nature Publishing Group. *Controlled drug release by a nanorobot : Nature Biotechnology : Nature Publishing Group* [online] 07.05.2012, [cited 09.06.2013]. Available at:

<http://www.nature.com/nbt/journal/v30/n5/full/nbt.2206.html>

[4] Wikipedia contributors. *Worldwide LHC Computing Grid* [online] 11.03.2013, [cited 09.06.2013]. Available at:

<http://en.wikipedia.org/wiki/Worldwide_LHC_Computing_Grid>

[5] Alliance For Permanent Access. *Welcome to the APA* [online] 2013, [cited 09.06.2013]. Available at:

<http://www.alliancepermanentaccess.org/>

[6] PARSE.insight Project. *About PARSE.Insight: Permanent Access to the Records of Science in Europe* [online] 2013, [cited 09.06.2013]. Available at:

<http://www.parse-insight.eu/>

[7] Alliance for Permanent Access. *About APARSEN* [online] 2013, [cited 09.06.2013]. Available at:

<http://www.alliancepermanentaccess.org/index.php/aparsen/>

[8] OCLC. *PURL Home Page* [online], [cited 09.06.2013]. Available at:

<http://purl.oclc.org/docs/index.html>

[9] International DOI Foundation. *The DOI System* [online] 09.04.2013, [cited 09.06.2013]. Available at:

<http://www.doi.org/>

[10] Wikipedia contributors. *Open access* [online] 05.06.2013, [cited 09.06.2013]. Available at:

<http://en.wikipedia.org/wiki/Open_access>

[11] CHAM, Jorge. *PhD_Comics_Open_Access_Week_2012.ogv* [online] 11.11.2012, [cited 09.06.2013]. Available at:

<http://commons.wikimedia.org/w/index.php?title=File%3APhD_Comics_Open_Access_Week_2012.ogv>

[12] Creative Commons. *Creative Commons – About* [online] 2013, [cited 09.06.2013]. Available at:

<http://creativecommons.org/about>

[13] FOX, Maggie. *Moon landing tapes got erased, NASA admits* [online] 16.07.2009, [cited 09.06.2013]. Available at:

<http://www.reuters.com/article/2009/07/16/us-nasa-tapes-idUSTRE56F5MK20090716>

[14] BRUMFIEL, Geoff. *The First Web Page, Amazingly, Is Lost* [online] 22.05.2013, [cited 09.06.2013]. Available at:

<http://www.npr.org/2013/05/22/185788651/the-first-web-page-amazingly-is-lost>

[15] opensource.com contributors. *Open source is changing the world: join the movement | opensource.com* [online] 10.06.2013, [cited 10.06.2013]. Available at:

<https://opensource.com/>

[16] TEDxTalks. *What the Tech Industry Has Learned from Linus Torvalds: Jim Zemlin at TEDxConcordiaUPortland - YouTube* [online] 15.04.2013, [cited 10.06.2013]. Available at:

<http://www.youtube.com/watch?v=7XTHdcmjenI>

[17] Wikipedia contributors. *Linus's Law – by Eric Raymond* [online] 24.04.2013, [cited 10.06.2013]. Available at:

<http://en.wikipedia.org/wiki/Linus%27s_Law#By_Eric_Raymond>

[18] Wikipedia contributors. *Tommy Flowers – World War II* [online] 04.06.2013, [cited 10.06.2013]. Available at:

<http://en.wikipedia.org/wiki/Tommy_Flowers#World_War_II>

[19] FABIAN, Mike. *mike-fabian / langtable – GitHub* [online] 07.06.2013, [cited 10.06.2013]. Available at:

<https://github.com/mike-fabian/langtable>

[20] Wikipedia contributors. *FLOPS – Distributed computing records* [online] 05.06.2013, [cited 10.06.2013]. Available at:

<http://en.wikipedia.org/wiki/FLOPS#Distributed_computing_records>

[21] Red Hat Community Architecture team. *The Open Source Way* [online] 2009, [cited 10.06.2013]. Available at:

<http://www.theopensourceway.org/book/>