Measuring outcomes and impact from the BioHackathon Europe

Leyla Jael Castro¹, Corinne Martin², Georgi Lazarov³, Dana Cernoskova⁴, Terue Takatsuki⁵, Jennifer Harrow², and Dietrich Rebholz-Schuhmann¹

1 ZB MED Information centre for life sciences, Gleueler Str. 60, 50931 Cologne, Germany **2** ELIXIR Hub, Wellcome Genome Campus, CB10 1SD, Hinxton, United Kingdom **3** Rheinische Friedrich-Wilhelms University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany **4** Scientifika, Purkynova 91, 31600, Brno, Czech Republic **5** Database Center for Life Science, Research Organization of Information and Systems, Kashiwa, Chiba 277-0871, Japan
corresponding author: ljgarcia@zbmed.de

BioHackathon series:

[BioHackathon Europe 2020](#)

Virtual conference 2020

[BioHackathon Outcomes project](#)

Submitted: 02 Aug 2021

License:

Authors retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC-BY](#)).

Published by [BioHackrXiv.org](#)

Abstract

One of the recurring questions when it comes to BioHackathons is how to measure their impact, especially when funded and/or supported by the public purse (e.g., research agencies, research infrastructures, grants). In order to do so, we first need to understand the outcomes from a BioHackathon, which can include software, code, publications, new or strengthened collaborations, along with more intangible effects such as accelerated progress and professional and personal outcomes. In this manuscript, we report on three complementary approaches to assess outcomes of three BioHackathon Europe events: survey-based, publication-based and GitHub-based measures. We found that post-event surveys bring very useful insights into what participants feel they achieved during the hackathon, including progressing much faster on their hacking projects, broadening their professional network and improving their understanding of other technical fields and specialties. With regards to published outcomes, manual tracking of publications from specific servers is straightforward and useful to highlight the scientific legacy of the event, though there is much scope to automate this via text-mining. Finally, GitHub-based measures bring insights on some of the software and data best practices (e.g., license usage) but also on how the hacking activities evolve in time (e.g., activities observed in GitHub repositories prior, during and after the event). Altogether, these three approaches were found to provide insightful preliminary evidence of outcomes, thereby supporting the value of financing such large-scale events with public funds.

Introduction

BioHackathons bring together a relatively large network of bioinformaticians and life scientists, to collaboratively work on hacking projects typically requiring a range of expertise which are most efficiently found in a team setting, rather than held by single individuals. Before pandemic-imposed travel restrictions, BioHackathons were traditionally in-person events requiring a physical setting, such as conference centres with high-speed internet, along with accommodation and catering, often supplemented by facilities for social side events (for the purpose of team building and networking). Although these costs are no longer relevant to the virtual versions of BioHackathons that have occurred during the global pandemic, significant labour costs remain, both in terms of logistics ahead of, and during, the event, as well as that of the participants themselves. All things considered, BioHackathons are expensive, usually public-funded and generally unlikely to generate income. As a result, it is essential that the outcomes and impacts of these events are better understood, measured where possible, and disseminated

to relevant audiences including funders. In the case of the BioHackathon Europe, [ELIXIR](#) ("ELIXIR A distributed infrastructure for life-science information," 2019, pp. @harrow_elixir-excelerate_2021), the research infrastructure for life-science data, is the main sponsor, providing funding (e.g., venue, accommodation, catering) and in-kind (e.g., staff time/labour) support.

There is, however, no simple recipe for outcome and impact evaluation, and its evaluation is more a process than a methodology, and this manuscript presents three approaches which will inspire other public-funded (bio)hackathon organizers to demonstrate value to their funders. The three approaches correspond to post-event surveys, publications arising from (or being linked to) the event, and analysis of [GitHub](#) ("Build software better, together," n.d.) repositories.

One of the easiest ways to gain insights into what an event has enabled is to use post-event perception surveys, whereby participants are asked about how well the event was prepared and run, along with how they benefited from attending it. Benefits range from having made concrete progress on so-called 'hacking projects' worked on during the event, to other less tangible personal and professional achievements. Surveying directly after the event usually brings higher response rates as well as clearer insights into the effect the event has had, thereby resulting in more solid data to work from.

However, the outcomes and therefore the impact can also be observed long after the event itself. For example, projects can grow or derive in new (sub)projects, results can be published and new collaborations formed. Such longer-term effects are difficult to measure via surveys, as contacting the participants once the event has finished is not always easy, and response rates are typically low for those successfully reached. Relying on data-driven assessment, gathering public data from code, data and publication repositories is an alternative that can complement shorter-term surveys. Publishing results from a BioHackathon is not as straightforward. BioHackathon immediate results commonly correspond to ongoing work, a sort of publication welcomed by some conferences but not necessarily by preprint servers or journals. To overcome this difficulty, organizers from the [National Bioscience Database Center \(NBDC\)](#) ("National Bioscience Database Center - NBDC," n.d.) and [Database Center for Life Science \(DBCLS\)](#) ("DBCLS - database center for life science," n.d.) NBDC/DBCLS BioHackathon in Japan, and the ELIXIR ("ELIXIR A distributed infrastructure for life-science information," 2019, pp. @harrow_elixir-excelerate_2021) BioHackathon Europe established a preprint server, [BioHackrXiv](#) ("BioHackrXiv Preprints," n.d.), dedicated to host publications from BioHackathons, sprints and code fests on the Life Science and Health Care domain. BioHackrXiv can hence be used as a source to identify publications arising from (or being linked to) specific hacking events, alongside other more mainstream sources of publications (Peer reviewed or not).

Code repositories together with their activities, contributors and metadata are a third aspect of our analysis of outcomes from BioHackathon projects. GitHub is one of the most common code repositories used during BioHackathons. During the BioHackathon Europe 2020, a project to analyze metadata from GitHub repositories was accepted. The initial version of this project, used for this manuscript, gathers metadata mainly on commits and licenses and provides an overview of repository activity before, during and after a BioHackathon.

In this manuscript, we report on these three complementary aspects to assess the BioHackathon Europe outcomes. The first approach consists of using above-mentioned post-event perception surveys, whilst the second focuses on published outcomes. The third approach consists in gathering data directly from the repositories used to work on the BioHackathon participant projects. Whilst these three approaches are independent from each other, they remain complementary and together, they offer a more useful picture to assess the outcomes brought about by the BioHackathon in support of a broader evaluation of its value and longer-term impacts.

Measuring outcomes

Survey-based measures

To date, the BioHackathon Europe has run annually since 2018, twice as an in-person (physical) event (2018, 2019) and once in virtual form due to pandemic-related travel restrictions (2020). Each time, a post-event survey was used to capture participants' perceptions of event quality and effect. Although the set of questions has evolved over the years to reflect learning on behalf of the organizers, questions remain broadly similar and comparable across the three occurrences. Beyond logistical performance (i.e., how well the event was prepared and run, and overall satisfaction level), respondents were also asked about outcomes, which included technical ones (related to the hacking project(s) they worked on) as well as personal and professional ones.

In particular, one question (*"Without the BioHackathon, how long would it have taken a single person with all the skills to reach the same outcome?"*) aimed at estimating the "acceleration effect" of the event on progress towards outcomes of hacking projects, against the counterfactual scenario (no BioHackathon Europe). In essence, this question aimed to test the hypothesis as to whether getting bioinformaticians and life scientists together in a focused environment does lead to increased progress, that might have happened far slower (if at all) without the event. Such an acceleration effect relates to one of ELIXIR's impact areas: *research efficiency* (e.g., faster, easier, more integrated, etc) (Martin et al., 2021). To ensure sufficient 'return on investment', i.e., value, to those who 'fund' the event (cash and in-kind), the time- and hence cost-savings linked to the event was compared to the estimated overall cost of holding the event, encompassing venue, catering, accommodation, and labour (i.e., staff time, effort) before and during the event.

Another set of questions, scored along an agree/disagree Likert scale, was used (from 2019) to gain insights into any personal and/or professional outcomes that participants felt were enabled through their participation in the event. Personal outcomes included perceived improvements to one's *technical skills and/or knowledge* (and *whether these would likely be used in the next 6 months*), improvements to one's *capacity to work as a team*, and one's improvements to *present and communicate work* and to *understand other technical fields/specialties/sectors/cultures*. Professional outcomes were focused on whether attendance at the event helped make one's *profile more attractive to recruiters*, and helped *usefully broaden one's professional network*. Altogether, this set of questions relates to two key impact areas of ELIXIR: *relationship capital* (e.g., benefits of working together, facilitated knowledge-sharing & cooperation) and *human capital* (e.g., better skills for users and service providers).

Publication-based measures

In research, generating new knowledge is pointless if it is not shared for others to use and build upon. In the world of research infrastructures and particularly for ELIXIR, increased awareness of new knowledge related to bioinformatics resources is an impact area (Martin et al., 2021). For this manuscript, we explored the potential of measuring the events' scientific legacy by simply carrying out manual searches in BioHackrXiv, using keywords such as 'BioHackathon' and 'ELIXIR', followed by manual curation of results, specifically looking at the "series" of the preprint (linked to given events) and the Acknowledgement field. The resulting list was complemented with other relevant publications from a list maintained by the ELIXIR Hub (the coordinating secretariat of the infrastructure).

GitHub-based measures

In addition to the above presented measures, there are some values that can be obtained directly from the code hosted in well known repositories such as GitHub, one of the favourites by attendees to the BioHackathons. We extracted some of these values from GitHub with the

purpose of providing a set of impact measures based on the activity carried on before, during and after a BioHackathon. We used the projects from BioHackathon Europe as a proof of concept for our approach.

The BioHackathon Europe offers a [repository](#) (Europe, 2021) hosting a folder per project and initialized with some basic information taken from the original project proposal and saved to a file `readme.md`. Project leads and participants can decide whether to add their code or to some other repository. For instance, our project hosted its code on its [own repository](#) (Lazarov & Castro, n.d.) while the project titled “bio.tools integration and sustainable development” used [the BioHackathon repository](#). It was necessary for us to cover both options.

Our strategy for past BioHackathons was simple: Go through all the folders inside the BioHackathon repository, read the `readme.md` file and search for anything that looks like a GitHub repository link; including the BioHackathon repository as well. For the BioHackathon 2020 we asked participants to tag their repositories with the label “biohackeu20” and used a [GitHub API](#) (GitHub, n.d.) functionality to retrieve them all. Both approaches have flaws, e.g., neither all GitHub like-links in the `readme.md` files nor all of those tagged as “biohackeu20” will correspond to BioHackathon projects; however, our primary goal was extracting information out of the repositories, rather than having the selected repositories right from the beginning. For all of the identified repositories, we used the GitHub API to retrieve some metadata: license, creation date and number of commits.

Results

Across the three BioHackathon Europe events held to date, over 500 bioinformaticians and life scientists collaboratively worked on just over 100 hacking projects, see Table 1. Overall satisfaction levels were consistently high, with almost all participants agreeing that they would recommend the event to a colleague.

Table 1. Key information relating to the 2018, 2019 and 2020 BioHackathon Europe events.

Year	2018	2019	2020
Format	Physical event	Physical event	Virtual event
Location or platform	Seine Port, France	Seine Port, France	Zoom, Remo, Slack
Attendee number	147	149	297
Proportion female attendees (based on registration)	19%	18%	31%
Number of hacking projects	29 (list)	33 (list)	41 (list)
Hacking days	5	5	5
Response rate to the post-event survey	44% (65 responses)	75% (112 responses)	48% (144 responses)
Overall satisfaction level (“Would you recommend this event to a colleague?”)	95% (yes)	99% (yes)	100% (yes)

Year	2018	2019	2020
Overall usefulness of the event ("Has the BioHackathon advanced the outcomes of projects you have been involved with?")	100% (yes)	87% (yes)	98% (yes)

Survey-based measures

In both 2019 (shown in Figure 1) and 2020, around 65% of survey respondents estimated that, without the BioHackathon Europe, it would have taken a single person (with all the required expertise) between 1 and 6 months to reach the same outcome as was achieved due to the event. Both years, just above 10% even stated that the same outcome would never be reached without the event having taken place.

For the year 2019, labour costs (effort) during the event were estimated at approximately 178k euros, based on the number of projects (33 projects), number of participants per project (on average 4.5 participants per project), event duration (5 days), and the [average hourly labour cost in Europe](#) ("Wages and labour costs," n.d.). Based on survey responses and following a similar approach, labour costs associated with each counterfactual category (1 week, 1 month, 2 months, 6 months, 1 year) amounted to just over 1 million euros in additional labour costs, leading to the event having an acceleration effect of 5.6 (Figure 1). Even when venue/accommodation/catering costs are accounted for, along with an estimated one full time equivalent per year to prepare and run the event (i.e., altogether a cost of 341k euros to prepare and run the event), the BioHackathon Europe accelerated progress toward outcomes by a factor of 3. Time- and hence cost-savings linked to the event were hence significant.

Figure 1: Acceleration effect of the BioHackathon Europe in 2019, as perceived by participants (2018 and 2020 results are similar). The effect of the event was to accelerate overall progress towards outcomes by a factor of 3 to 5.6 compared to the counterfactual (i.e., not holding the event)

In both 2019 and 2020 (shown in Figure 2), good insights were gained into personal and/or professional outcomes that participants felt were enabled through their participation in the events. Results were broadly similar across the two years when the question was asked, which is interesting in itself since 2020 was in virtual format. Respondents overwhelmingly agreed that they had improved their capacity to work as a team, increased their understanding of other technical fields and specialties, as well as usefully broadened their professional networks (though not in terms of career prospects with recruiters in mind). Most also agreed that they had improved their technical skills and/or knowledge, and that they would use these within the

next 6 months, thereby giving insights into the longer-term outcomes of the event. One area where the events did not appear to have an effect related to the understanding of other sectors and their needs (e.g., industry if the respondent was an academic) - this was likely related to the small number of hacking projects submitted by private sector participants. One notable difference between the 2019 and 2020 results related to the understanding of other cultures, with 53% (strongly) agreeing that the 2019 physical format event helped them, whilst this was only 34% in 2020 (virtual format).

Figure 2: Personal and professional outcomes reported by participants of the BioHackathon Europe in 2020

Publication-based measures

To date, twelve publications have arisen from and/or can be linked to the BioHackathon Europe series of events (Table 2). They represent one facet of the scientific legacy of the event, keeping in mind that it is not possible to ascertain direct causality. Yet, it is clear that the events contributed to the generation of new knowledge, which was then shared with others via these publications. In due course, their subsequent citation levels will give insights into the usefulness of the work to others, which is an outcome in itself.

All publications corresponding to preprints have been published at BioHackrXiv, with the number of publications increasing from year to year, while there are only two publications for the BioHackathon Europe in 2019, six are available for the 2020 edition (up to the end of July 2021). In addition to the publications corresponding to the BioHackathon Europe, BioHackrXiv also hosts five publication from the COVID-19 Virtual BioHackathon held in 2019, one from the NBDC/DBCLS BioHackathon held in 2019, and one more from other BioHackathons. Although the rate of submissions to BioHackrXiv is still low regarding the number of participating projects (around 6% in 2019 and 15% in 2020 for the BioHackathon Europe), organizers expect to see an increase in the following years with the goal of having one publication per project or topic (being a topic a merge of multiple yet related projects), thus contributing not only to dissemination but also to Open Science.

Table 2. Publications, Peer reviewed or not (i.e., preprints) have arisen from and/or can be linked to the BioHackathon Europe events.

Publication	Type	Comments
Several BioHackathon	Europe events	
Garcia et al., 2020, Ten simple rules to run a successful BioHackathon, Plos Computational Biology DOI:10.1371/journal.pcbi.1007808 (L. Garcia et al., 2020)	Peer reviewed	Acknowledges participants and organizing committee
BioHackathon Europe	2020	
Gray et al., 2021, Exploiting Bioschemas Markup to Populate IDPcentral. DOI:10.37044/osf.io/v3jct (Gray, Papadopoulos, Micetic, & Hatos, 2021)	Preprint	Thanks organizers and participants. "BioHackathon Europe 2020" series of BioHackrXiv
Labra-Gayo et al., 2021, Knowledge graphs and wikidata subsetting, BioHackrXiv. DOI:10.37044/osf.io/wu9et (Labra-Gayo et al., 2021)	Preprint	Thanks organizers. Included in the "BioHackathon Europe 2020" series of BioHackrXiv
Savojardo et al., 2021, SB4ER: an ELIXIR Service Bundle for EpidemicResponse, BioHackrXiv. DOI: 10.37044/osf.io/b34vm (SAVOJARDO et al., 2021)	Preprint	Thanks organizers. Included in the "BioHackathon Europe 2020" series of BioHackrXiv
Groom et al., 2021, Connecting molecular sequences to their voucher specimens, BioHackrXiv. DOI:10.37044/osf.io/93qf4 (Groom et al., 2021)	Preprint	Thanks organizers. Included in the "BioHackathon Europe 2020" series of BioHackrXiv
Zobolas et al., 2020, Linking PubDictionaries with UniBioDicts to supportCommunity Curation, BioHackrXiv. DOI:10.37044/osf.io/gzfa8 (Zobolas, Kim, Kuiper, & Vercruysse, 2020)	Preprint	Thanks organizers. Included in the "BioHackathon Europe 2020" series of BioHackrXiv
Suchánek, 2020, Progress on Data Stewardship Wizard during BioHackathon Europe 2020, BioHackrXiv. DOI: 10.37044/osf.io/9mnkb (Suchanek, Hooft, & Bourhy, 2020)	Preprint	Thanks organizers. Included in the "BioHackathon Europe 2020" series of BioHackrXiv
BioHackathon Europe	2019	

Publication	Type	Comments
Bono and Kasukawa, 2020, TogoEx: the integration of gene expression data, BioHackrXiv DOI:10.37044/osf.io/esrc9 (Bono & Kasukawa, 2020)	Preprint	Acknowledges ELIXIR as founder of the event. Included in the "BioHackathon Europe 2019" series of BioHackrXiv
Gawron et al., 2020, Disease and pathway maps for Rare Diseases, BioHackrXiv. DOI:10.37044/osf.io/gmbjv (Gawron et al., 2020)	Preprint	Acknowledges ELIXIR as founder of the event. Included in the "BioHackathon Europe 2019" series of BioHackrXiv
BioHackathon Europe 2018		
Garcia et al., 2019, Biotea-2-Bioschemas, facilitating structured markup for semantically annotated scholarly publications. Genomics and Informatics. DOI: 10.5808/gi.2019.17.2.e14 (Garcia Leyla, 2019)	Peer reviewed	Acknowledges support
Garcia et al., 2020, Ten simple rules for making training materials FAIR, PLOS Computational Biology DOI:10.1371/journal.pcbi.1007854 (B. A. B. Garcia Leyla AND Batut, 2020)	Peer reviewed	Acknowledges the event
Bresso et al., 2021, Investigating ADR mechanisms with Explainable AI: a feasibility study with knowledge graph mining. DOI:10.1186/s12911-021-01518-6 (Bresso et al., 2021)	Peer reviewed	Acknowledges participants

GitHub-based measures

We run our jupyter notebooks for projects corresponding to the BioHackathon Europe 2019 (E. Europe, n.d.-a) and 2020 (E. Europe, n.d.-b) and for the [Virtual BioHackathon Covid 2020](#) ("Virtual-biohackathons/covid-19-bh20," 2021). A preliminary analysis was performed from the data obtained via the GitHub API for the BioHackathon Europe 2020 corresponding to 17 projects as of identified on the fourth afternoon of the hacking days, i.e., one day before final presentations. Table 3 shows the 17 repositories sorted by the second column corresponding to the license. We found four repositories with no license, two with a file corresponding to a license but with no enough metadata for the GitHub API to identify the actual license, and eleven projects with an identifiable license: four with MIT license, two with GNU general public license, two with GNU AFFERO license, and three with Apache license. Finding projects with no license at all was very surprising for a BioHackathon, as we were expecting some common software best practices, e.g., recommendations for research software (Jimenez et al., 2017), to be followed in this event (particularly given its hacking/coding nature). One of the projects

with no license was the BioHackathon repository.

From the 17 analyzed repositories (Table 3), twelve seem to have been created for the BioHackathon as the creation date was during the hacking days or very close to it (less than 1 month). Three repositories correspond to established projects running for more than one year: V-pipe has been running for about four years, Panoptes for one year and The Turing Way for two years. The other two repositories, BioHackathon-projects-2020 and vsm-pubdictionaries, have been running for less than 5 months. As expected, the number of commits increased during the hacking days with an average of 6 commits per day considering all of the 17 projects. While analyzing the number of commits, we realized that some projects showed a low number of commits, some even zero commits, e.g., V-pipe. A further investigation showed that we were targeting the wrong branch. We only gathered data from the main (previously known as master) branch while some projects, following good practices, created a dedicated branch for the development during the BioHackathon that later would be integrated into the main branch.

Table 3. Preliminary results from the GitHub-based analysis for some of the identified projects for the BioHackathon Europe 2020. Repositories created during the event will exhibit a negative number of active days before it

Repository	License	Active days before event	Average commits before event	Average commits during event
cbg-ethz/ V-pipe	Apache License	1309	0.13	0.00
ds-wizard/ h2020-dmp- template	Apache License	-1	0.00	0.40
zbmed/ BioHackOutcomes	Apache License	5	0.20	2.60
biohackr/iv/ bhxiv-metadata	GNU AFFERO GENERAL PUBLIC LICENSE	-2	0.00	2.00
babelomics/ MechACov	GNU GENERAL PUBLIC LICENSE	26	0.12	1.20
elixir- luxembourg/ BH2020-galaxy- dpmapi	GNU GENERAL PUBLIC LICENSE	0	0.00	4.40
danielabutano/ intermine- R2RML- mapping	GNU LESSER GENERAL PUBLIC LICENSE	7	0.71	8.60
collaborativebioinformatics/ GrOMOP	MIT License	0	0.00	11.20
JervenBolleman/ rs-handlegraph- ffi	MIT License	-2	0.00	2.40
kg-subsetting/ biohackathon2020	MIT License	0	0.00	13.40
panoptes- organization/ panoptes	MIT License	364	0.13	2.40

Repository	License	Active days before event	Average commits before event	Average commits during event
elixir-europe/ BioHackathon- projects-2020		140	0.64	29.40
matdillen/ 33_molseq		-1	0.00	7.60
NuriaQueralt/ covid19- epidemiology- ontology		0	0.00	0.40
UniBioDicts/ vsm- pubdictionaries		135	0.03	2.40
alan-turing- institute/ the-turing-way	# LICENSE	739	7.53	15.00
yocra3/ epimutacions	YEAR: 2020	13	0.54	21.00

We carried out a follow-up analysis on the 20th of April 2021 as shown in Table 4. After the final presentation, more repositories added the tag “biohackeu20” and were picked up by our code. We got a total of 24 projects, eight of them tagged after the final presentations. One of the new repositories, corresponding to a controlled vocabulary, had a license commonly associated to data more than to software. We also noticed that one project removed the “biohackeu20” tag. Most of the repositories kept active for about 3 months after the event but none of them presented much activity in terms of commits with one exception corresponding to the [Bioschemas website repository](#) (Community, 2021), a very active repository supporting the web pages for the [Bioschemas project](#) (Community, n.d., p. @gray_potato_2017). This low activity, of course, could be due to a mismatch regarding active branches as, once again, we only analyzed the master branch.

Table 4. Post-event results for some of the BioHackathon Europe 2020 projects.

Repository	License	# days before event	Average commits before event	Average commits during event	Average after event
UniBioDicts/ vsm- pubdictionaries	### GNU AFFERO GENERAL PUBLIC LICENSE	135	0.03	3.40	0.28
cbg-ethz/ V-pipe	Apache License	1309	0.13	1.60	0.06
ds-wizard/ h2020-dmp- template	Apache License	-1	0.00	0.40	0.02
zbmed- semtec/ BioHackOutcomes	Apache License	5	0.20	3.20	0.01

Repository	License	# days before event	Average commits before event	Average commits during event	Average after event
biohackrxiv/ bhxiv- metadata	GNU AFFERO GENERAL PUBLIC LICENSE	-2	0.00	3.80	2.67
babelomics/ MechACov	GNU GENERAL PUBLIC LICENSE	26	0.12	20.60	0.20
elixir- luxembourg/ BH2020- galaxy- dpmaps	GNU GENERAL PUBLIC LICENSE	0		5.80	
danielabutano/ intermine- R2RML- mapping	GNU LESSER GENERAL PUBLIC LICENSE	7	0.71	9.20	0.07
collaborativebioinformatics/ GrOMOP	MIT License	0		19.20	
JervenBolleman/ rs- handlegraph- ffi	MIT License	-2	0.00	3.40	3.00
kg- subsetting/ biohackathon2020	MIT License	0		15.20	0.12
panoptes- organization/ panoptes	MIT License	364	0.13	1.60	0.15
elixir- europe/ BioHackathon- projects- 2020		140	0.64	34.60	0.48
matdillen/ 33_molseq		-1	0.00	19.20	0.60
NuriaQueralt/ covid19- epidemiology- ontology		0		0.40	0.43
yocra3/ epimutations Identified after final presentations	YEAR: 2020	13	0.54	24.80	
HW-SWeL/ BMUSE	Apache License	580	0.39	1.80	0.18

Repository	License	# days before event	Average commits before event	Average commits during event	Average after event
edamontology/ edamontology	Attribution-ShareAlike 4.0 International	2336	0.41	0.20	0.22
intermine/ intermine- R2RML- mapping	GNU LESSER GENERAL PUBLIC LICENSE	-4	1.25	9.20	0.06
BioSchemas/ bioschemas.github.io		1721	1.38	0.20	1.47
BioSchemas/ specifications		1841	0.25	0.40	0.24
bio-tools/ biotoolsConnect		1740	0.05	1.40	0.01
bio-tools/ content		654	0.56	8.20	0.66
ResearchObject/ ro-crate-py		385	0.71	3.00	0.60

We performed an additional analysis on the follow-up data regarding the number of commits per repository before, during and after the event, see Figure 3. Out of thirteen repositories created for the BioHackathon Europe, ten of them were very active mostly during the event but commits declined after it while the remaining three showed a good number of commits even after the event. Although all of the participant projects are of interest for the Life Sciences community and bring in advances and developments worth sharing, we want to highlight these three cases as they correspond to projects (or at least code repositories) which started during the event and grew after it, i.e., they showed some post-event continuity, a desirable, although not always possible, effect on the BioHackathon projects. These three projects correspond to (i) a Data Management Plan template part of the [Data Stewardship Wizard](#) (Pergl, Hooft, Suchanek, Knaisl, & Slifka, 2019), a well established part of the ELIXIR Europe network, (ii) [PubDictionaries](#), part of the [PubAnnotation](#) tools ecosystem (Kim et al., 2019), a text mining platform for Bioinformatics offered by DBCLS, and (iii) an ontology related to COVID-19 epidemiology and monitoring. The latter started as a research project at the COVID-19 Virtual BioHackathon held in 2019 and, since then, has continued evolving. The project MolSeq exhibits more commits during the event (56.47%) but not by far regarding commits post-event (43.53%). Some other projects that bring some attention are those located at the very left of the graphic as they do not show much, or not at all, activity after the BioHackathon. Although this could indicate that the projects were not continued after the event, it is also possible that they have been moved to another GitHub project corresponding to a parent project. Finally, those repositories on the very right correspond to repositories with more than one year of creation. Although the impact of the BioHackathon might not be big on them, it is still interesting to observe how there is room at the BioHackathon for those projects to evolve as they commonly count with more resources and even with a dedicated team. This broad spectrum shows us how the BioHackathon supports a variety of projects from inceptions to well-established ones, a mix promoting the exchange of ideas across different level of expertise and fully aligned to the benefits of the BioHackathon Europe: supporting new and existing developments and resources, and promoting personal and community building.

Figure 3: Percentage of commits per repository before, during and after the event. Repositories created for the event are surrounded by a green rectangle, those with less than four months of creation by a purple rectangle and the rest (with more than one year of creation) by a black rectangle.

Discussion and future work

Despite its relatively short history, it is clear that the BioHackathon Europe is bringing about positive outcomes. Undeniably, post-event surveys can only capture a partial picture of these outcomes, yet they have proved to be a useful approach which can be completed by others.

In relation to personal and professional outcomes, it is true that happy participants are more likely to give positive feedback, or give feedback at all. Yet, given the suitably high level of response obtained for 2018-2020 (range 44-75%), the feedback likely aligns with reality. The set of questions provided useful insights into what participants gained from dedicating their time to attending the events. This is especially interesting with regards to softer outcomes such as teamwork and networking, which are notoriously difficult to measure given their intangible nature. The surveys also contained free text boxes where participants were invited to enter narrative-based feedback - these could in the future be looked at in more detail, for instance by coding comments to tease out recurrent themes, and perhaps use this to improve the more standard survey questions. With regards to estimating the acceleration effect brought about by the event (between 3 and 5.6 times faster towards outcomes, compared to without the event), the most important aspect of this exercise is that the result should be meaningful. Some might see it as a back-of-the-envelope calculation, yet the sensitivity analysis (that considers significant upstream preparation effort) suggests that the effect is not too far off what one would reasonably expect from such an event.

The 2020 edition of the event was special on many levels, given that it was held virtually in the midst of a global pandemic. Yet, survey results remain similar to that of previous years, especially overall satisfaction levels, but also personal and professional outcomes (2019 results are included in (Martin et al., 2021)). Beyond undeniably proving that successful virtual BioHackathons are possible (despite expectations from most), the virtual format is likely to have made the event more inclusive, not least because engaged participant numbers almost doubled (297 participants) compared to 2018-2019 (147 and 149, respectively). This has been acknowledged by participants recognizing that a virtual mode makes it easier for people to

partially join at their own time and pace which, due to time restrictions, might be more difficult in a face to face mode (Labra-Gayo et al., 2021), building on numerous narrative comments in the free text boxes of the post-event survey of 2020. Of note, female participants increased in 2020 (31%) than in 2018-2019 (19 and 18% respectively).

With regards to published outcomes, although the set of eight publications can feel limited to some, it ought to be remembered that publishing hackathon outcomes is not mainstream. It is also possible that some publications have been missed, and automation of the process via text-mining may help increase detection rate. BioHackathon organizers should continue to encourage participants to publish outcomes of their participation at the event and acknowledge any relevant support, so that a scientific legacy is visible to others (who may wish to build on and cite the publications), as well as those funding the event.

Regarding data collected from GitHub, despite being a very basic approach, gathering a minimum of data, i.e., creation date, number of commits, license, we observed some interesting facts, the most relevant being the lack of license for some projects. No license means no reusability, licensing is one of the most important elements of data and software metadata. It is in fact that important, that it is explicitly mentioned by the FAIR principles for data (Wilkinson et al., 2016) and software (Lamprecht et al., 2020) and the recommendations for research software (Jimenez et al., 2017), among others. We hope that all projects in future editions of the BioHackathon Europe and NBDC/DBCLS will include a license file that can be parsed by machines. As part of the future work, we need to include more metadata and cover branches other than the master/main one. We should also include support for GitLab. In addition, we would want to propose a way to produce a less speculative analysis as to what happens with repositories after the BioHackathon and why some of them stop or decline the activity after the event. Some of the possibilities include integration into other branches or repositories, or abandonment of the project; however, so far we can only figure this out by manual inspection combined with a series of searches on the Internet, and this is not enough nor sustainable. Further analysis beyond software repositories are also within the scope of the outcomes, for example, BioHackathons are a great opportunity to meet new people and start new collaborations that could lead to new co-authorships.

Efforts similar to the one presented in this manuscript, which particularly targets the BioHackathon Europe, should also be carried out for the NBDC/DCBLS BioHackathon in Japan, which has been running for more than twelve years now. Although measuring the impact has not been formally addressed, informal ways such as direct feedback have been taken into account. Furthermore, the BioHackrXiv preprint server is a direct response from the NBDC/DCBLS BioHackathon organizers to increase the exposure of hacking projects and is a direct way to document and share outcomes from the participant projects. Since 2021 and as a result of a suggestion from BioHackathon Europe 2020 participants, BioHackrXiv is now indexed in EuropePMC, thereby furthering the reach and visibility of outcomes published in this server.

An important aspect in both BioHackathons, Europe and NBDC/DBCLS, is the creation or extension of collaborative networks that can be evidenced by, for instance, new co-authorship or new joint projects. This is an additional aspect regarding outcomes and impact that should be addressed in the future.

GitHub repositories

[Project repository zbmed-semtec/BioHackOutcomes](#)

Note: Data corresponding to the tables is also available in the GitHub repository in TSV format.

Acknowledgements

Post-event perception surveys were developed as part of two projects funded by the European Union under Horizon 2020: RI-PATHS (777563), and ELIXIR-CONVERGE (871075). Elina Griniece (EFIS) provided useful advice on the design and interpretation of the surveys and their results. BioHackathon Europe receives funding and in-kind support from ELIXIR, the research infrastructure for life science data. The GitHub analysis was done as a BioHackathon Europe 2020 project. The authors acknowledge all the support provided by organizers and the participants in the BioHackathon Europe 2020.

References

- BioHackrXiv Preprints. (n.d.). Retrieved from <https://biohackrxiv.org/>
- Bono, H., & Kasukawa, T. (2020, July). TogoEx: The integration of gene expression data. BioHackrXiv. doi:[10.37044/osf.io/esrc9](https://doi.org/10.37044/osf.io/esrc9)
- Bresso, E., Monnin, P., Bousquet, C., Calvier, F.-E., Ndiaye, N.-C., Petitpain, N., Smail-Tabbone, M., et al. (2021). Investigating ADR mechanisms with Explainable AI: A feasibility study with knowledge graph mining. *BMC Medical Informatics and Decision Making*, 21(1), 171. doi:[10.1186/s12911-021-01518-6](https://doi.org/10.1186/s12911-021-01518-6)
- Build software better, together. (n.d.). *GitHub*. Retrieved from <https://github.com>
- Community, B. (2021, April). BioSchemas/bioschemas.github.io. Bioschemas. Retrieved from <https://github.com/BioSchemas/bioschemas.github.io>
- Community, B. (n.d.). Bioschemas. Retrieved from <https://bioschemas.org/>
- DBCLS - database center for life science. (n.d.). Retrieved from <https://dbcls.rois.ac.jp/index-en.html>
- ELIXIR A distributed infrastructure for life-science information. (2019, December). Retrieved from <https://elixir-europe.org/>
- Europe, B. (2021, March). Elixir-europe/BioHackathon-projects-2020. ELIXIR Europe. Retrieved from <https://github.com/elixir-europe/BioHackathon-projects-2020>
- Europe, E. (n.d.-a). Biohackathon Europe 2019. Retrieved from <https://2019.biohackathon-europe.org/>
- Europe, E. (n.d.-b). BioHackathon Europe 2020. Retrieved from <https://2020.biohackathon-europe.org/>
- Garcia, B. A. B., Leyla AND Batut. (2020). Ten simple rules for making training materials fair. *PLOS Computational Biology*, 16(5), 1–9. doi:[10.1371/journal.pcbi.1007854](https://doi.org/10.1371/journal.pcbi.1007854)
- Garcia, L., Antezana, E., Garcia, A., Bolton, E., Jimenez, R., Prins, P., Banda, J. M., et al. (2020). Ten simple rules to run a successful BioHackathon. *PLOS Computational Biology*, 16(5), e1007808. doi:[10.1371/journal.pcbi.1007808](https://doi.org/10.1371/journal.pcbi.1007808)
- Garcia Leyla, G. A., Giraldo Olga. (2019). Biotea-2-bioschemas, facilitating structured markup for semantically annotated scholarly publications. *Genomics Inform*, 17(2), e14. doi:[10.5808/GI.2019.17.2.e14](https://doi.org/10.5808/GI.2019.17.2.e14)
- Gawron, P., Hoksza, D., Pinero, J., Chilet, M. P., Esteban, M., Fernandez, J. L., Colonna, V., et al. (2020, July). Disease and pathway maps for rare diseases. BioHackrXiv. doi:[10.37044/osf.io/gmbjv](https://doi.org/10.37044/osf.io/gmbjv)
- GitHub. (n.d.). GitHub REST API - GitHub Docs. Retrieved from <https://docs.github.com/en/rest>

- Gray, A. J. G., Goble, C., & Jimenez, R. C. (2017). From Potato Salad to Protein Annotation. In *ISWC Posters and Demo session* (p. 4). Vienna, Austria. Retrieved from <http://ceur-ws.org/Vol-1963/paper579.pdf>
- Gray, A. J. G., Papadopoulos, P., Micetic, I., & Hatos, A. (2021, June). Exploiting bioschemas markup to populate idpcentral. BioHackrXiv. doi:[10.37044/osf.io/v3jct](https://doi.org/10.37044/osf.io/v3jct)
- Groom, Q. J., Dillen, M., Huybrechts, P., Johaadien, R., Kyriakopoulou, N., Fernandez, F. J. Q., Trekels, M., et al. (2021, March). Connecting molecular sequences to their voucher specimens. BioHackrXiv. doi:[10.37044/osf.io/93qf4](https://doi.org/10.37044/osf.io/93qf4)
- Harrow, J., Hancock, J., ELIXIR-EXCELERATE Community, & Blomberg, N. (2021). ELIXIR-EXCELERATE: Establishing Europe's data infrastructure for the life science research of the future. *The EMBO Journal*, 40(6), e107409. doi:[10.15252/embj.2020107409](https://doi.org/10.15252/embj.2020107409)
- Jimenez, R. C., Kuzak, M., Alhamdoosh, M., Barker, M., Batut, B., Borg, M., Capella-Gutierrez, S., et al. (2017). Four simple recommendations to encourage best practices in research software. *F1000Research*, 6, 876. doi:[10.12688/f1000research.11407.1](https://doi.org/10.12688/f1000research.11407.1)
- Kim, J.-D., Wang, Y., Fujiwara, T., Okuda, S., Callahan, T. J., & Cohen, K. B. (2019). Open Agile text mining for bioinformatics: the PubAnnotation ecosystem. *Bioinformatics*, 35(21), 4372–4380. doi:[10.1093/bioinformatics/btz227](https://doi.org/10.1093/bioinformatics/btz227)
- Labra-Gayo, J. E., Hevia, A. G., Alvarez, D. F., Ammar, A., Brickley, D., Gray, A. J. G., Prud'hommeaux, E., et al. (2021). *Knowledge graphs and wikidata subsetting* (Preprint). BioHackrXiv. doi:[10.37044/osf.io/wu9et](https://doi.org/10.37044/osf.io/wu9et)
- Lamprecht, A.-L., Garcia, L., Kuzak, M., Martinez, C., Arcila, R., Martin Del Pico, E., Dominguez Del Angel, V., et al. (2020). Towards FAIR principles for research software. *Data Science*, 3(1), 37–59. doi:[10.3233/DS-190026](https://doi.org/10.3233/DS-190026)
- Lazarov, G., & Castro, L. J. (n.d.). BioHackathon Outcomes GitHub Metadata. Retrieved from <https://github.com/zbmed-semtec/BioHackOutcomes>
- Martin, C. S., Repo, S., Marquez, J. A., Blomberg, N., Lauer, K. B., Sitja, X. P., Velek, P., et al. (2021). Demonstrating public value to funders and other stakeholders-the journey of ELIXIR, a virtual and distributed research infrastructure for life science data. *Annals of Public and Cooperative Economics*, n/a(n/a). doi:<https://doi.org/10.1111/apce.12328>
- National Bioscience Database Center - NBDC. (n.d.). Retrieved from <https://biosciencedbc.jp/en/>
- Pergl, R., Hooft, R., Suchanek, M., Knaisl, V., & Slifka, J. (2019). Data Stewardship Wizard: A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. *Data Science Journal*, 18(1), 59. doi:[10.5334/dsj-2019-059](https://doi.org/10.5334/dsj-2019-059)
- SAVOJARDO, C., Martelli, P. L., Babbi, G., Anteghini, M., Manfredi, M., Madeo, G., Capriotti, E., et al. (2021, March). SB4ER: An elixir service bundle for epidemic response. BioHackrXiv. doi:[10.37044/osf.io/b34vm](https://doi.org/10.37044/osf.io/b34vm)
- Suchanek, M., Hooft, R. W., & Bourhy, K. (2020, December). Progress on data stewardship wizard during biohackathon europe 2020. BioHackrXiv. doi:[10.37044/osf.io/9mnkb](https://doi.org/10.37044/osf.io/9mnkb)
- Virtual-biohackathons/covid-19-bh20. (2021, April). virtual-biohackathons. Retrieved from <https://github.com/virtual-biohackathons/covid-19-bh20>
- Wages and labour costs. (n.d.). Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Wages_and_labour_costs
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. doi:[10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18)

Zobolas, J., Kim, J.-D., Kuiper, M., & Vercruysse, S. (2020, December). Linking pubdictionaries with unibiodicts to support community curation. BioHackrXiv. doi:[10.37044/osf.io/gzfa8](https://doi.org/10.37044/osf.io/gzfa8)