0v45b8ll§3("8"ydß=tx!f)m(ß"fyz1$i4
3&30mmvn)a)=§6ywi77ßob9zm4pxsa?wif
ln8gq9r1"h9df&r&s3eo/46"bgu5sh)mp7
rtv6rxn!ys7vdqrpd2$%y)!(m7wm!w4?§s
2vutl4xlh0?q§pvy=eeh%§x2i8§96z9r73
lbr9/ss3f$350p?"byva6(l§ytxvuß6fs(
zngel2ßx7dqb!oa18hg)vwu16q1cßkohcn
?2=u5=q8wfu(bk/§349bq!ohwdt)o$u!ew
4y5m(j1swpnnr)!o46g790h8/63h?qjjq6
il0tl5q/g5n0k=)(gwsg&laxwpnib(/r(?
w9eh§kbwo6gkna0w0x=yd91o=y3/9v!§6"
x$32ßt?p1lguhgy$!ß(u%g"68y4(ig7c§6
a&jzaj2kwei!v48&"17xr7udyj)uigk%fr
he5fi07f5g9)(t)4)u7v/j=5/w46b8pdag
l§bt//0wn§7mfm7reex2bgfßx&?!41qawd
6wl/koun83%)!nm90i3er7o4qc4ßrk9r0&
3v2qv)lvyz§l1&bg%bge=rm/5n)qmgo$6l
wu&ig=s8njz8r&od2j&)vx§**The Journal**
k4xywc16jbßo)16ew9!2n/n**of Raw Data**
rgj9bfiufmr2xnir6vubyw0c9qtlaeysu3**.org**
31zze)/&6x3/vtzapz)8rnlg=nyyn7"adh
7?kb1q"4%lst9trxul!)5m$10")u=rt0"z
0mq$(3g$pcoj=cikykeqt!xc14l34%ß3d!
)&qbeqz6ßo&$%=g?b7li&!/b4§!3q)po49
0hw38qjd=!/u)$7o!3l(yvmyq%§)$azk5?
bc4?4mw34vxo(g(rj9/nmjß?qxgwa2!vqz
$tg/!tjk1qv3=/$wssf")1ahnrsa&xkfc%
lpi&y4"xa56j9lhu0j=5k3sxp298?vvvue
2§§o=sl(74h9!t0$f6%6kßa(2n3uhez9i&
6tk3&tl5j7j12eq=7wjjo4ßn?il(x9b/2m
brf07o)v$&a6s$y!zsß1l2$ecg%04!k%7%
3ys7vzqi"/d§b6k?v&=5qr1dx)!e17&)tm
av!tb?ßzl%lxok58p&27o1wdqrr%q2qg§f
of§phgs$zroie1vywkc/b5?z%zjhxf4d6i
irq61y(p5e(m=0w"a$4/w0l)bv&z3v&2p"
74/6§%vh§0w5672?gows1$e0s2ß)/b!kp1
op(i(gzqcß2i(nltw48!du!pc9%92=f64q
"w=sh)duz5jtidf$4lb=nz)fmßifs3noge
7?8%e2x$"ßw(8"x9xg0/(kv!gj9$2?p7p6
l%?a%4%h/&kuz0swzl(1xlynds5n2qoy7x
qoo/h$!%m6r/&cuhmyhdgß&7341gv9k2oi
ß§uj7hlxß0z/sp7(q))!%u?u/uhk!4&8rb
$$z262g)t§()l$g§ag!j?1zhxamkm2%!kl
9h9h?&a$!dmljyhpßxt7t2h3l5u8hi74tf
x"6v=r**Whitepaper** n**Version 1.0**a2?66u
n§§"3§sf0m&y**20.03.2018**8wu&l98(023z

Contents

Executive summary

To promote the ‘open data’ ideal in biomedical sciences we here make the case for founding a new scientific journal - *The Journal of Raw Data* - that enables scientists to make the raw data sets of their research publicly accessible. By minimizing the required effort and maximizing the scientific reward to submit datasets, researchers are incentivized to publish their raw data in our journal. In contrast to other existing data sharing alternatives, we ensure non-manipulability of the data through decentralized storage of the data and of immutable cryptographic hashes on the Arweave, a distributed internet data storage based on blockchain technology. In case of success, *The Journal of Raw Data* can be expected to have a major impact on modern biomedical research by providing reward solely for a high quality study design and conduct, taking the first step towards separating research study design and data acquisition from data analysis in order to improve scientific quality.

1 The problem

Currently, biomedical research finds itself in a 'reproducibility crisis'[1] with an estimate of only 10% of studies' results being able to be reproduced[2] and subsequently 85% of research resources wasted.[3] A variety of factors have been identified as the main causes, including inappropriate study designs, mistakes during data collection, flawed data analysis and failure to report the results without omissions.[4,5] However, all of these problems can be alleviated by making the process and content of research transparent and accessible, which is promoted as 'open science'.[6]

One crucial aspect of open science is 'open data', which refers to the sharing of raw data sets generated during the course of biomedical studies.[6] Especially 'open data' has the potential to have a major impact on modern research practice, as it enables other researchers to check the published analyses for flaws or perform additional analyses on the data.[4] Also, 'open data' encourages the publication of otherwise ignored datasets that hold no (statistically significant) results, but which are crucial to perform valid meta-analyses of studies that investigated a common research question.[6]

Because of these benefits, 'open data' has been receiving increasing interest in the biomedical research community over the last years, with an increasing number of initiatives providing open data repository services.[7] However, very few of these data repository services provide means to overcome the most important barrier to meet the 'open data' ideal, which is the missing incentive for researchers to pursue open practices.[4,6] In consequence, adherence of researchers to 'open data' standards is basically non-existent.[8]

2 Our vision

Our vision is to promote the 'open data' ideal in biomedical sciences by founding a scientific journal - *The Journal of Raw Data* - that enables researchers to obtain scientific reward for publishing their raw data in a way that prevents forgery.

Success of *The Journal of Raw Data* would vastly increase scientific quality, as misuse of analytical and statistical methods is easily detectable when the underlying datasets are publicly accessible.[4,5,9] Furthermore, the possibility to obtain scientific reward solely for a high quality study design and conduct reduces the needs to polish data and encourages the publication of complete datasets without selection bias as well as the publication of datasets with no (statistically significant) results that would otherwise be ignored.[4,5] Thus, to maximize scientific quality, our ultimate goal is to completely separate research study design and data acquisition from data analysis. Towards this goal, the foundation of *The Journal of Raw Data* is the first step.

[1] Baker M. 1500 scientists lift the lid on reproducibility. *Nature* 2016. 533: 452-454.
[2] Begley CG and Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature* 2012. 483: 531-533.
[3] Chalmers I and Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet* 2009. 374: 86-89.
[4] Munafo M et al. A manifesto for reproducible science. *Nature Human Behaviour* 2017. 1 (21): 1-9.
[5] Ioannidis JP et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet* 2014. 383: 166-175.
[6] Nosek BA et al. Scientific Standards: Promoting an open research culture. *Science* 2015. 348: 1422-1425. Web: http://cos.io
[7] Bertagnolli MM et al. Advantages of a truly open-access data-sharing model. *NEJM* 2017. 376: 1178-1181.
[8] Iqbal SA et al. Reproducible research practices and transparency across the biomedical literature. *PLOS Biology* 2016. 14: 1.
[9] Garcia-Berthou E et al. Incongruence between test statistics and p values in medical papers. *BMC Med Res Method* 2004. 4: 13.

3 The Journal of Raw Data

3.1 What is *The Journal of Raw Data* and how does it work?

The Journal of Raw Data is a newly to establish online-only open access peer-reviewed biomedical data journal and the core of our project.

The journal will provide a web-interface through which biomedical scientists can submit study protocols and raw data sets of their biomedical studies. Each submission is handled by an editor, who is an independent scientist working in the respective scientific field. If the editor decides that the quality of a submitted dataset is sufficient to consider publication, it is forwarded to reviewers, who are independent scientists that are familiar with the methods applied in the respective study (‘peer-review process’). The reviewers rate the submitted study protocol and dataset regarding quality and provide recommendations whether or not to accept them for publication. Based on these recommendations, the editor takes the final decision about the acceptance and communicates it to the author.

If accepted, the submitted study protocol and raw data set are published on the journal’s website, where they are openly available for readers free of charge (‘online-only open access journal’). Raw data sets are either stored for free in a publicly accessible database operated by the journal or against payment of a fee stored decentralized on the Arweave, a massively scalable blockchain that offers permanent data storage. In either case, cryptographic hashes of the dataset are stored on the Arweave to prevent data manipulation.

3.2 What features does *The Journal of Raw Data* offer?

3.2.1 Providing incentive for the submission of datasets

“Publish or perish” - Publications in scientific journals are what ultimately decides upon success or failure of a scientific career. Thus, any possibility to obtain an additional publication provides a strong incentive for biomedical scientists. By submitting raw data sets to *The Journal of Raw Data,* scientists obtain additional scientific journal publications, while preserving the possibility to publish their analyses in any other scientific journal as usual.

3.2.2 Preventing data manipulation

Once published in *The Journal of Raw Data*, datasets are protected from data manipulation by storing cryptographic hashes of the data on the Arweave (→ section 4.3). As this enables anyone to verify the integrity of any of the published data, neither the scientists, nor us or any other third party is able to undetectably manipulate published datasets.

3.2.3 Minimizing the effort required for the submission of datasets

To minimize effort and time required to submit datasets, we provide submission interfaces with optimized usability that allow to submit datasets in virtually any kind of data structure and format, thus relieving scientists from the task of extensively preparing and arranging the data (→ section 4.1).

3.2.4 Facilitated data sharing

To lower the boundaries for other researchers to perform additional analyses or meta-analyses of datasets published by *The Journal of Raw Data,* all datasets are published in intuitively comprehensible and machine readable structures in easily and openly accessible file formats (→ section 4.1). Furthermore, *The Journal of Raw Data* offers a variety of tools to search across all published datasets for specific types of data and to obtain datasets composed from the data of multiple datasets (→ section 4.2).

3.2.5 Decentralized data storage

The Journal of Raw Data offers to store published datasets either in a database operated by the journal or decentralized on the Arweave, a massively scalable blockchain (→ section 4.3). Storage of datasets in the journal's database is offered free of charge and guarantees data availability for at least 10 years with public data access limited by data traffic capabilities of the database. In contrast, storage of datasets on the Arweave requires payment of a fee, but offers permanent storage and faster data access due to the decentralized storage concept (→ section 4.3).

3.3 Comparison to other options to handle raw data sets

	Additional scientific reward for the dataset	Protection against data manipulation	Low effort	Easy data sharing	Decentralized data storage
The Journal of Raw Data	++	++	++	++	++
Keeping data private	-	-	++	-	-
Other data repositories	-	+	-	+	-
Other data journals	++	+	-	+	-

Table 1: Positive and negative aspects of options for biomedical scientist to handle their raw data sets

3.3.1 Keeping datasets private

The currently by far most popular option for scientists to handle their raw data sets is to keep them private. For newly obtained datasets, the main reason is that scientists do not want to give up their primary right to exploit the data and to surrender their scientific edge on competitors within their field. However, with growing age of datasets these initial reasons dissolve and the sole main reason to keep datasets private is that it requires less effort in comparison to publishing them.[16] However, even keeping datasets private does not come completely without effort, as virtually all major scientific organisations, institutes and journals require scientists to store all raw data for at least 10 years.[10]

The major drawback of keeping datasets private is that it does not offer any scientific reward. Furthermore, private datasets can be quite difficult to share and hold only a low credibility, as they are potentially subjected to data manipulation.

3.3.2 Open access data repositories

The second most common option to handle raw data sets is to submit them to open access data repositories, which are web-services storing datasets in a publicly accessible way. The main reason for the growing popularity of openly accessible data repositories is that an increasing number of scientific journals require scientists to openly publish their datasets.[6]

Major drawbacks of data repositories are that they do not offer any scientific reward, while storing a dataset requires a substantial amount of effort and time.[16] Also, while data once published in an open access data repository is protected against manipulation performed by the scientist, data can still be manipulated by the data repository itself and subsequently by third parties that attain administrative access to the data repository through fraud.

3.3.3 Other data journals

A third option to handle raw data sets is to publish them in one of the few already existing data journals.[11] Similar to *The Journal of Raw Data*, these journals publish the study protocol including a data description and either publish the associated dataset themselves or require scientists to submit the datasets to open access data repositories. In either case the published items count as scientific journal publications by which the scientists are rewarded for submitting their datasets.

However, the amount of datasets that are currently published through the available data journals is evidently extremely low,[12] due to the high amount of effort and time that existing data journals require from scientists to format and describe their datasets.[16] Furthermore, identically to data repositories, while published data is protected against manipulation by the scientist, data can still be manipulated by the data journal and subsequently by third parties that attain administrative access to the data journal through fraud.

[10] Recommendations of Deutsche Forschungsgemeinschaft (DFG) for “Safeguarding Good Scientific Practice” (2013).
[11] Web: http://www.forschungsdaten.org/index.php/Data_Journals.
[12] Only 8 of the data journals listed in citation 17 are included in the 2016 Thomson Reuters InCites Journal Citation Report with a median of only 31 published items in 2016 (range: 3 [Geoscience Data Journal] - 490 [Journal of Chemical & Engineering Data]).

4 Technical Key Aspects

4.1 Data submission interfaces with automated data formatting

4.1.1 Optimizing usability

As the motivation of scientists to submit their datasets is strongly dependent on the required effort and time to do so, we provide data submission interfaces that are optimized to facilitate the submission process through a high level of usability. We offer a web-interface to provide the lowest possible effort to upload single datasets, and we offer a “dropbox-like” interface, through which a virtual folder can be mounted in the local file system, to provide the highest convenience for frequent users and datasets of high complexity.

4.1.2 Automated data formatting

Key component of both interfaces is an adaptive dialog system, which allows scientists to submit data in virtually any format and structure, thus alleviating the effort and time required for submission. As the adaptive dialog system collects all necessary information about the datasets, it understands their content, format, and structure and is therefore able to re-format and re-structure the data into open and standardized formats to allow access to the data independent of specific software programs and to ensure accessibility in the far future. To maximize reusability of the data, data storage will adhere to the FAIR principles (Findable, Accessible, Interoperable, Reusable).[13]

4.1.3 Privacy protection

As clinical studies on humans are required to protect the anonymity of the participants, the adaptive dialog system prevents the submission of protected personal data like names or dates of birth and anonymizes medical imaging data by clearing personal information from image metadata and scrambling facial features in images. Furthermore, all datasets are cross checked by the journal’s staff for deanonymizing information before publication.

4.2 Data access interfaces facilitating reuse of datasets

4.2.1 Licensing of published items

To promote reuse of published datasets, all items published by *The Journal of Raw Data* will be openly available under Creative Commons licensing (CC-BY).[14] This type of licensing allows scientists to retain their original rights in the published items, but allows the publisher to distribute the work, and all readers to unrestrictedly reuse the work, as long as the source work is appropriately cited.

[13] Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016. 3: 160018.
[14] Web: https://wiki.creativecommons.org/wiki/BIS_committee_UK_OA_Policy.

4.2.2 Data access interfaces

All study protocols and datasets published by *The Journal of Raw Data* are accessible through the journal's web-interface or the journal's "dropbox-like" file sharing interface, regardless whether the datasets are stored in the journal's database or on the Arweave. Both interfaces provide functions to access single published items, but also to search across all published datasets for specific types of data and to obtain datasets composed from the data of multiple datasets to facilitate additional analyses and meta-analyses of published data. Furthermore, both interfaces allow easy handling of encrypted data by offering intuitive encryption key management and thus offering the possibility to use all available functions also on datasets that are stored on the Arweave, but that are not openly published (→ section 5.2.2).

4.3 Data storage on the Arweave

4.3.1 What is the Arweave and how can data be stored there?

The Arweave is an '*open, irrevocable, unforgeable and uncensorable archive for the internet*'[15] based on blockchain technology. As a 'blockweave' the Arweave is a data structure in which blocks of data are linearly chained together as in an ordinary blockchain, and additionally cross-linked to random previous blocks. This data structure is mirrored evenly across a distributed network by incentivizing voluntary network participants (i) to store as many blocks as possible and (ii) to collect and store rare blocks that are not widely mirrored by others.

Storing data on the Arweave is associated with costs, which depend on the amount of data to store and the current capacity and occupancy of the Arweave network. However, costs occur only once during the storage process and as soon as the network confirms the transaction, the data becomes an irrevocable part of the Arweave and is thus protected from manipulation.

4.3.2 Which data is stored on the Arweave by The Journal of Raw Data?

For every dataset that is published by *The Journal of Raw Data*, a minimum set of information including descriptors of the dataset along with cryptographic hashes of the contents of every file in the dataset will be stored on the Arweave. These cryptographic hashes allow to prove that the raw data stored on the Arweave or in the journal's database is complete and not manipulated, as it is practically impossible to manipulate the raw data in a way by which the manipulated data produces the same hash as the original raw data. Apart from this minimum set of information that is required to be stored on the Arweave, the submitting scientist may choose to store any additional amount of their data up to the complete dataset on the Arweave by bearing the additional expenses (→ section 5.2.1).

[15] Web: https://www.arweave.org/whitepaper.pdf.

5 Business Aspects

5.1 The biomedical publication market

The size of the global scientific journal publishing market is estimated at about US$ 10 billion for the year 2013, with a compound annual growth rate of approximately 4.5%.[16] As biomedical journals account for about 30% of all scientific journals, but for up to 60% of the revenue, the biomedical scientific publishing market can be approximated at up to US$ 6 billion per year and growing.[16]

This market revenue is generated either by charging readers of scientific journals for access to the articles (the “traditional” revenue model) or by charging authors for publishing their articles (the “open access” revenue model, sometimes also referred to as the article publication charge (APC) business model).

The dominance of the traditional revenue model reflects in the numbers that in the year 2013 only 12% of scientific articles were published as open access,[16] representing only 4% of the total journal market’s revenue.[17] Thus, the open access market share in 2013 equated to about US$ 250 million per year, generated by approximately 250.000 publications at publication charges between US$ 400 - US$ 5000 per item.[16]

However, even though the current market share of open access publications is relatively small, it shows a far above average growth rate of 32% per year between 2012-2014 as an expression of the increasing importance of open science principles for the scientific community.[17] Furthermore, a recent raise in publication charges between 2010-2015 assumingly reflects an increasing number of large-scale publishers shifting revenue models of established journals towards open access.[16]

An important factor encouraging this growth is the increasing willingness of scientific institutions to provide the costs for open access publication.[18] This increasing willingness can be expected to further increase after recent disagreements between major scientific institutions and non-open access publishers.[19] In the short term, providing open access publication costs is an additional expense for scientific institutions, but in the long term these additional costs are mitigated by the reduced costs for journal subscriptions in the traditional revenue model. In either case publishing costs are paid by the scientific institutions.

However, even though publication charges are ultimately paid by the institutions, the final decision about whether to publish in traditional or open access scientific journals is made by the scientist who authored the publication and who requests funding of the charges from his funding institution. Therefore, particular attention has to be given to the factors that influence the scientists’ decisions whether or not to pursue open access publications.

[16] Ware M, Mabe M. The STM Report: An overview of scientific and scholarly journal publishing (2015).
[17] Outsell Inc. STM 2015 market size, share, forecast & trend report.
[18] Web: http://oad.simmons.edu/oadwiki/OA_publication_funds.
[19] Schiermeier Q. Hundreds of German universities set to lose access to Elsevier journals. *Nature* 2017. 552: 17-18.

5.2 Monetization

Primary objective of *The Journal of Raw Data* is to improve scientific quality in biomedical research by promoting open data (→ section 2). However, revenues are required to fund the journal's expenses and sustainable growth. Therefore, *The Journal of Raw Data* generates revenue (1) directly from scientists for submitting raw data sets (B2C revenue) or (2) from institutions for storing their datasets (B2B revenue):

5.2.1 Revenue from scientists for submitting raw data sets

To generate revenue we focus on the authors of the 250.000 biomedical journal publications per year that are published in open access journals requiring submission fees (→ section 5.1). For this group it is evident that open access funds are available and the scientists are willing or even required by their funding body to adhere to open science principles. Accordingly, we expect these scientists to be equally willing to obtain additional funds to pay for openly publishing their raw data sets, if the required effort and time are low and the incentive is sufficiently strong.

However, the strength of the incentive equals the value of the scientific reward obtained through the additional publication in *The Journal of Raw Data.* Therefore, it is crucial to maximize the number of submitted datasets, since a higher number of submitted datasets allows to reject low quality submissions and to maximize the quality of published datasets. Higher quality of published datasets along with tools to facilitate additional analyses lead to a higher number of citations, which increase the journal's visibility, reputation and value in scientific indicators, thus increasing the incentive for scientists to submit their datasets.

Therefore, we initially follow a 'freemium' pricing strategy, which maximizes the number of submissions, while at the same time allows to obtain revenue from scientists that are willing to pay. Accordingly, we offer a free option that allows scientists to publish raw data sets in the journal's database, including storage of hashes on the Arweave to guarantee dataset integrity. In contrast, the paid option additionally stores the raw data itself on the Arweave, with the main differences to the free option in terms of accessibility and duration of storage (→ section 3.2.5). After an initial phase of a few years this pricing strategy can be adjusted in order to obtain the required revenue. Adjustment can be limited to modifying the differences between the free and the payment option or include to completely abandon the free option.

5.2.2 Revenue from institutions for storing raw data sets

As a secondary approach to generate revenue *The Journal of Raw Data* offers scientific institutions encrypted storage of datasets on the Arweave against payment. Using these services relieves institutions from the burden of operating their own data repositories and offers a variety of advantages, including use of our automated data formatting and data access interfaces that reduce effort and time to upload, access and share datasets, protection from data manipulation and ubiquitous availability of the data through decentralized data storage (→ section 3.2). While providing these advantages, data access control is maintained completelyin the hands of the institution and/or the scientists.

5.3 Stakeholders

As the main focus in the initial phase is to maximize the number of dataset submissions, we follow a multi-tier strategy that focuses primarily on (1) the scientists that control the datasets, but also on cooperations with (2) scientific institutions and (3) other scientific journals. Furthermore, we support (4) open science initiatives in their work to influence politicians, regulatory authorities and the general public towards supporting open data principles.

5.3.1 Biomedical scientists

Primary target of our strategy are scientists that possess raw data sets from biomedical research. As the large majority of scientists keep their raw data sets private, our primary strategy is to convince this group to submit their datasets for publication by providing scientific reward as incentive while requiring low amounts of effort and time (→ section 3.2).

5.3.2 Scientific institutions

A secondary target of our strategy are major scientific institutions, as most of these operate their own data repositories to which they oblige all of their scientists to submit their datasets to ensure safe data storage for at least 10 years as required in most cases by regulatory authorities or funding institutions.[10] Thus, we offer scientific institutions encrypted storage of their datasets on the Arweave, which relieves them from the burden of operating their own data repositories and allows them to use our automated data formatting and data access interfaces (→ section 5.2). As is requires only minimal effort to publish data that is already stored through our services, it is expected that an above average proportion of scientists from institutions who store their data through our services will submit their datasets for publication in *The Journal of Raw Data*.

5.3.3 Other scientific journals

Another secondary target of our strategy are scientific journals that do not publish raw data themselves, but only analyses results. Especially cooperations with journals that require scientists to publish their raw data with public access would motivate a large number of scientists to submit their datasets to *The Journal of Raw Data*. In return, the cooperating journals could benefit by obtaining a guarantee that datasets submitted within the cooperation are handled with priority and are thoroughly checked to ensure validity.

5.3.4 Open Science Initiatives

Another component of our strategy is to support open science initiatives in their work to influence politics, regulatory authorities and the public towards supporting open data principles. It can be expected that increases of the market segment of open science directly translates into an increasing market share of *The Journal of Raw Data*.

5.4 Marketing

The here described marketing efforts aim at our primary customers, which are publishing biomedical scientist, with a special focus on those that publish in open access journals (→ section 5.2). Potential cooperators and business customers like other scientific journals or scientific institutions are exclusively contacted via direct-marketing.

5.4.1 Publishing about *The Journal of Raw Data*

One approach to initially inform the potentially most interested biomedical scientists is to publish about our vision and concept in scientific open access journals. As we expect that every scientist who is convinced by our concept will help to spread the word within the scientific community, we invest every effort to provide good arguments and an attractive product.

5.4.2 Online advertisement

To inform biomedical scientists about the possibility to obtain additional scientific reward by publishing their datasets, we utilize online advertisement that allows to focus on our target group. Especially online advertisement via Google AdWords and advertisement on websites frequented predominantly by biomedical scientists allow a very accurate targeting.

5.4.3 Advertisement at conferences and events

Besides online advertisement, sponsoring of scientific events and conferences allows to target scientists of a specific area of biomedical sciences, either to optimize revenue if it becomes evident that scientists from a specific area are more ready to use our paidoption or to reach scientists from an area that is otherwise not reached.

5.4.4 Direct marketing

In addition to the described group-targeting advertisement we use direct marketing to attract influential scientists, which can be expected to have a multiplying marketing effect when they are convinced to support our journal. Direct marketing will tie influential scientists to our journal by recruiting them as reviewers for the submitted datasets or as section editors to handle the submissions. Both reviewers and editors are incentivized for their support by receiving free storage of their own datasets on the Arweave.

5.4.5 Cooperations with other journals, institutions and social networks

Another approach to achieve a multiplying marketing effect is to establish cooperations with scientific journals that publish analyses and results of studies, but not the raw data sets, or cooperations with scientific institutions (→ section 5.3). Furthermore, cooperations with online social networks that focus on scientists can be a valuable way to reach our primary customers and add to usability by offering a way to sign into our customer system without separate authentication.

5.5 Expenses

Not included in this version

5.6 Timeline

Not included in this version

6 Risks

6.1 Risks in behaviour of biomedical scientists

6.1.1 Data publications are not considered independent publications

Risk: The scientific community might consider publications of raw data sets and publications of the analyses results obtained from that same data not as two independent publications, but as a double publication of the same results. Thus a publication of the raw data set of a study would prohibit to publish analyses results.

Countermeasures: Convincing the scientific community of the concept of separate data publications can be achieved best by emphasizing the benefits for scientific quality, as data publications help to promote open data principles. Furthermore, contrasting the contents of data publications against the contents of analyses publications illustrates that the two types of publications hold complementary, but different information. To make sure that these arguments reach the scientific community, we apply a multichannel and multilevel marketing approach, especially targeting influential scientists (→ section 5.4).

6.1.2 Acceptance or rejection of raw data sets is perceived arbitrary

Risk: The scientific community might perceive it arbitrary which raw data sets are accepted for publication and which are rejected.

Countermeasures: To make the decision criteria for acceptance or rejection of datasets transparent and to give reviewers guidance, we publish explicit quality criteria for all types of datasets. While we directly adopt dataset quality criteria that already exist for a variety of areas of biomedical sciences, we define and publish dataset quality criteria for other areas in collaboration with influential representatives from the respective areas.

6.1.3 Scientists keep raw data sets private

Risk: The majority of scientists might still keep their raw data sets private, even though incentive to publish them is offered through scientific reward and the effort required for submission is minimized.

Countermeasures: We actually expect the majority of scientists to still keep their datasets private, despite all of our efforts. Thus, this does not affect our strategy, which is primarily focused on the fraction of scientists that already publish in open access journals. Influencing the general majority of scientists towards adhering to open data principles will be a slow and weary process that can only be successful after we have attracted the group of scientists who already support open science. Transformation of scientists from the unconcerned general majority to the group of open science supporters is encouraged by supporting open science initiatives in their work to influence politics, regulatory authorities and the public.

6.1.4 Scientists prefer other data journals

Risk: Scientists might prefer to submit their datasets to data journals other than The Journal of Raw Data.

Countermeasures: Currently, the number of available data journals and the number of datasets published via these journals is evidently very low.[12] However, it is expected that more and more data journals are founded in the near future, thus increasing competition for biomedical raw data sets. In case of increasing competition, we expect scientists to decide for a journal primarily based on the value of the scientific reward offered by the respective journal and the required effort to submit the dataset. In terms of required effort, we expect our journal to hold an advantage over other journals (→ section 6.3.1). However, in terms of scientific reward, other journals might be able to provide higher value, especially in the early phase. As the value of scientific journals is determined by indicators that are mostly based on the number of citations of publications by the respective journal in relation to the number of published items, journals that publish only very few items hold an advantage as they are able to select those high quality datasets with a high probability of being cited from all submissions. To be able to provide a similar high scientific reward for high quality datasets *The Journal of Raw Data* will offer several sub-journals that are ranked in regard to dataset quality. Thus, we provide the possibility to achieve a maximum scientific reward by publishing in our highest ranked journal, or to achieve a medium value scientific reward for datasets of average quality.

6.2 Risks in behavior of funding institutions

6.2.1 Funding institutions refuse to pay for data publications

Risk: The institutions that provide the funds for biomedical scientists to pay for open access publications of the analyses results might refuse to pay for additional data publications.

Countermeasures: Identically to convincing the general scientific community, convincing funding institutions of the concept of separate data publications can be achieved best by emphasizing the benefits for scientific quality, as data publications help to promote open data principles. To make sure that our arguments reach the funding institutions, we recruit scientists that hold influential positions within the institutions as supporters via direct marketing and also approach the institutions *ab initio*.

6.2.2 Funding institutions refuse to pay if there is a free option

Risk: The funding bodies might refuse to pay for the storage of the researchers' raw data on the Arweave as long as there is a free alternative provided by the journal, which is storing the data on the journal's own database.

Countermeasures: Our revenue model is subject to change as required to obtain the necessary revenue. Possible changes in the revenue model include to discontinue the free publication option and charge a fee for all types of publications if necessary (→ section 5.2.1).

6.3 Risks in technical development

6.3.1 Submission of raw data sets with low effort is not possible

Risk: Since no other yet existent data repository or data journal provides input interfaces that allow to submit datasets with low effort, it might be the case that the task of submitting raw data sets of biomedical studies is necessarily associated with high effort.

Countermeasures: Based on a comprehensive analysis of the dataset submission process, we identified the step of formatting and structuring datasets according to open standards, which is required for further utilization of datasets after publication, as the paramount component in the submission process that determines how the required effort for the submission is perceived by users. Accordingly, we focus our development efforts on offering an automated data formatting system that allows users to provide the necessary information about the dataset within the submission process. This system reduces the perceived effort through a continuous stepwise process that does not require the user to interrupt the submission to perform tasks in other programs, that only requests the minimal required information about the dataset and that is clear and specific in its requests, thus avoiding confusion and effortful side-processes.

6.3.2 Automated data formatting is too difficult to be implemented

Risk: The key technical component that decides about success or failure of The Journal of Raw Data is the automated data formatting system that relieves biomedical scientists from the effort to structure and format datasets before their submission. Thus, the system must be able to understand the content, format, and structure of the submitted datasets in order to re-format and re-structure the data into open and standardized formats. Due to the large variety of available data formats and possible data structures, this task might be difficult to be reliably implemented.

Countermeasures: The problem of automated data processing might be complex but every part of it is technically feasible, therefore the question is not so much whether it is possible to solve it, but whether is it possible to be solved using the available resources. However, the complexity of the task is primarily a problem of high input variability and not a problem of finding the correct output for each input. Thus, the problem can be solved best by a scalable continuous development process, in which backend developers enhance functionality of the system step-by-step in close co-operation with experienced data scientists that are part of our team. In case of dataset formats and structures that the automated system can not handle, data specialists which uninterruptedly monitor the dataset submission process take over interaction with the submitting scientist and create queries for the system's backend developers, who implement the required functionality as soon as the available resources allow. Thus, in dependence of the available resources the system is extended in functionality step-by-step and every type of dataset is, once submitted, included in the input range of the automated data formatting system.

6.4 Risks concerning the Arweave

6.4.1 The Arweave network might not be fast or durable enough

Risk: The Arweave network might not establish itself with a sufficient performance or durability as necessary to provide a secure long-term storage of datasets.

Countermeasures: The performance and persistence of the Arweave network depends on the network participants, which are incentivized for participation in the network through Arweave tokens. As the value of Arweave tokens depend on the use of the Arweave, the best countermeasure against a too weak network is using the Arweave, which *The Journal of Raw Data* does by storing datasets. However, to avoid data loss and provide a necessary minimum performance for the datasets stored on the Arweave, *The Journal of Raw Data* mirrors all datasets stored on the Arweave in its own database and participates as a node in the Arweave network.

6.4.2 Arweave tokens might become too expensive

Risk: In contrast to the previous point, which would be a consequence of too little use of the Arweave, a too high demand for storage on the Arweave might result in Arweave tokens to become too expensive to store the raw data sets on the Arweave.

Countermeasures:
As *The Journal of Raw Data* provides an Arweave node, a raise in demand that increases the incentive for the nodes to participate in the Arweave network might be counteracted by increasing the capacity of our node and thus gaining more incentive through the node that counteracts the increasing costs for storage. Therefore, balancing the strength of our node in a cost-benefit-ratio allows to ensure stable storage costs within a certain range. However, in case that storage prices were driven by other forces than storage demand, e.g. through speculative investments, storage prices might leave the balancable range. In case such adverse conditions are not resolved within a reasonable amount of time, we will allow direct access to the datasets on our own servers even when the data is stored on the Arweave, as we will in any case mirror all datasets in our own database.

7 The Team

Priv.-Doz. Dr. Falk von Dincklage

Falk is an anesthesiologist and biomedical scientist at Charité - Universitätsmedizin Berlin, one of the largest university hospitals in Europe. As a clinician, he is part of the management of the Department of Anesthesiology and Intensive Care Medicine. As a scientist, he oversees multiple biomedical and medical technology research projects and is a member of the doctoral committee at Charité - Universitätsmedizin Berlin.

Gregor Lichtner

Gregor is a neuroscientist, computer scientist and PhD candidate, currently working at the Department of Anesthesiology and Intensive Care Medicine of Charité – Universitätsmedizin Berlin at the intersection of medicine and computer science. Before his employment at Charité - Universitätsmedizin Berlin, Gregor has worked several years in the health technology industry as a lead developer for user interfaces of medical applications, specialized on usability design.

Georgi Tadeus

Georgi is a management consultant, currently working in several projects for multinational companies mostly within the chemical and process industry. His main focus is developing business and organizational design as well as management systems. As a trained biochemist with a strong analytical background, he has a profound interest in improving data availability in the biomedical research.

Claudia Friedrich

Claudia is an anesthesiologist and biomedical scientist at Charité - Universitätsmedizin Berlin, who has spent parts of her beginning research career in tissue engineering and organ fabrication at the Harvard Medical School. She is involved in organizing biomedical programs and conferences like the European Students Conference, the transatlantic Biomedical Sciences Exchange Program or the German-American Conference at Harvard University.

Priv.-Doz. Dr. Dr. Felix Balzer

Felix is an anesthesiologist and head of a data science research group at Charité – Universitätsmedizin Berlin, who is appointed as the professor for “E-Health and Shared Decision Allocation” at Berlin’s Einstein Center Digital Future. His research at the intersection of medicine and informatics has a focus on complex data analyses to identify process indicators for anaesthesiology and critical care.

Akira Poncette

Akira is an anesthesiologist and data scientist at Charité – Universitätsmedizin Berlin, leading various digital health projects with industry partners. Akira is also a lead organiser of the global non-profit organisation Hacking Health implementing novel techniques such as the hackathon into hospitals to foster innovation in healthcare.