The grant is dead, long live the data - migration as a pragmatic exit strategy for research data preservation [version 2; peer review: 2 approved]

RESEARCH NOTE

REVISED The grant is dead, long live the data - migration as a pragmatic exit strategy for research data preservation [version 2; peer review: 2 approved]

Tomasz Zielinski [1], Johnny Hay [2], Andrew J. Millar [1]

[1]SynthSys and School of Biological Sciences, University of Edinburgh, Edinburgh, EH9 3BF, UK
[2]EPCC, University of Edinburgh, Edinburgh, EH9 3FD, UK

v2 **First published:** 02 Jul 2019, **4**:104 (https://doi.org/10.12688/wellcomeopenres.15341.1)
Latest published: 23 Sep 2019, **4**:104 (https://doi.org/10.12688/wellcomeopenres.15341.2)

Abstract

Open research, data sharing and data re-use have become a priority for publicly- and charity-funded research. Efficient data management naturally requires computational resources that assist in data description, preservation and discovery. While it is possible to fund development of data management systems, currently it is more difficult to sustain data resources beyond the original grants. That puts the safety of the data at risk and undermines the very purpose of data gathering.

PlaSMo stands for 'Plant Systems-biology Modelling' and the PlaSMo model repository was envisioned by the plant systems biology community in 2005 with the initial funding lasting until 2010. We addressed the sustainability of the PlaSMo repository and assured preservation of these data by implementing an exit strategy. For our exit strategy we migrated data to an alternative, public repository with secured funding. We describe details of our decision process and aspects of the implementation. Our experience may serve as an example for other projects in a similar situation.

We share our reflections on the sustainability of biological data management and the future outcomes of its funding. We expect it to be a useful input for funding bodies.

Keywords

Data sharing, research data management, sustainable data infrastructure, exit strategy, research funding

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 23 Sep 2019		
version 1 02 Jul 2019	✓ report	✓ report

1 **Helen Ougham**, Aberystwyth University, Aberystwyth, UK

2 **Robert Davey**, Earlham Institute, Norwich, UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Andrew J. Millar (andrew.millar@ed.ac.uk)

Author roles: Zielinski T: Conceptualization, Software, Supervision, Writing – Original Draft Preparation; **Hay J**: Software, Writing – Original Draft Preparation; **Millar AJ**: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the Wellcome Trust through a Wellcome Institutional Strategic Support Fund (ISSF3) [204804]. This work was also supported by the Biotechnology and Biological Sciences Research Council (BBSRC) through the UK Centre for Mammalian Synthetic Biology [BB/M018040].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Zielinski T, Hay J and Millar AJ. **The grant is dead, long live the data - migration as a pragmatic exit strategy for research data preservation [version 2; peer review: 2 approved]** Wellcome Open Research 2019, **4**:104 (https://doi.org/10.12688/wellcomeopenres.15341.2)

First published: 02 Jul 2019, **4**:104 (https://doi.org/10.12688/wellcomeopenres.15341.1)

REVISED **Amendments from Version 1**

We followed the input from the reviewers to improve the paper. We have added new paragraphs to the Discussion section, including one that describes the funding alternatives, with some additional references including the example of the TAIR portal. The new Discussion text also addresses tools to ease the pain points of programmatic migrations, specifically automating the code generation and knowledge mapping, with a comparison of current and earlier tools. We updated Figure 1 to improve its layout as suggested by Reviewer 1. We also discuss in more detail the "easy" and "hard" cases of biological data management.

Any further responses from the reviewers can be found at the end of the article

Introduction

Open research, data sharing and data re-use have become a priority for publicly- and charity-funded research, as expressed for example in the UK Concordat on Open Research[1]. Data re-use depends on reliable metadata: a detailed description of the experimental conditions, materials used, handling procedures and analysis methods. Data management goes beyond the safe storage of data, because metadata acquisition and data discovery are equally important aspects for effective digital preservation[2–4]. This creates a need for computational resources that can deliver such features.

Funding bodies acknowledge that data management carries significant costs and allow budgeting for data stewardship. For larger projects this permits the development of systems suitable for a particular research domain, by supporting specific data models or streamlining metadata collection. This occasionally results in the formation of a small community resource, sometimes described as "boutique repository". Unfortunately, while it is possible to fund data infrastructure for a project, currently, there are few schemes that could support a resource beyond the timeline of the initial grant[5]. The common approach is to cover maintenance costs by "tunnelling" funds from related projects. That is not a sustainable model and puts at risk the very data that the original grant paid to preserve.

The increasing demand for data archiving induced the creation of general repositories (e.g. Figshare[6], Zenodo[7], Dryad[8]) and also institutional repositories (e.g. University of Edinburgh DataShare[9], UK Data Archive[10]). They may lack flexibility to support all the various needs of an active project, but they are valid alternatives for data preservation. We propose to address the sustainability problem and mitigate the risk to *boutique* data by implementing an exit strategy in the form of data migration to a larger, public repository with secured funding.

Problem description

PlaSMo stands for 'Plant Systems-biology Modelling' and the PlaSMo portal (plasmo.ed.ac.uk) was envisioned by the plant systems biology community during a BBSRC and GARNet workshop in July 2005. The initial 2-year development was funded as part of BBSRC's Bioinformatics and Biological Resources call in 2008 and then supported by the European Commission's FP7 Collaborative Project TiMet (2010–2015).

The PlaSMo portal (henceforth referred to simply as 'PlaSMo') became a central resource for diverse plant models: general crop models, organ-level models or complex multi-component plant system models. At the time of its creation it was a unique resource for managing and sharing plant models, many of which were refactored into common, declarative languages (SBML or SimileXML). The PlaSMo repository contained over 100 described models and nearly 400 data and model files.

The main features of PlaSMo were:

- Support for multiple XML model formats: SimileXMLv3, SBML Level 2 versions 1–4, Cytoscape XGMMLv1, SBGN-MLv1
- Validation of the model format
- Managing multiple versions of the model
- Each version could have its own assets: definition file, supporting data, graphical representation, bibliography, description and comments
- Public, private or group access
- Free text search

The PlaSMo portal was implemented as a typical Java web application of its time: Apache Struts 2 as the Model-View-Controller (MVC) framework with Java Server Pages (JSPs) deployed on Apache Tomcat. The choice of Java as the language and technology stack proved to be robust and convenient. For example, the backend database was migrated from DB2 to MySQL, new model formats were added, and the Struts framework major version upgraded, all as ad-hoc tasks without the original developer present. Such tasks benefited from Java features such as strong typing, rich exception handling, well-defined JAR dependencies and IDE support.

Nevertheless, there were costs in providing such a public service including system administration, software development for occasional updates and user support. The Struts MVC framework had to be updated in a timely manner due to security concerns. There were critical vulnerabilities discovered in Struts that could permit arbitrary code execution and we observed attempts to exploit them just 8 hours after their disclosure. After the funded interval, all this work was performed as an in-kind contribution.

We noticed that PlaSMo had not been attracting new users. Its user interface was outdated, and the researchers had gained other facilities for sharing, like wikis or general repositories. It seemed that the value of the PlaSMo project was in its data rather than in its portal, hence, we decided to migrate the PlaSMo content into an alternative repository.

Decision process

We planned to use a repository designed specifically for biological data instead of a general one like Figshare, Zenodo or University of Edinburgh DataShare. The general resources have no features relevant to biological data (e.g. model types, difference between model and data), they also tend to have a

"flat" organization structure built around a concept of datasets. We wanted to preserve the "community" aspect of PlaSMo by having its resources grouped together and available for exploratory browsing. There is a dedicated repository for biological models: BioModels[11,12] but it accepted only public (usually published) models and (at the time) was restricted to those in SBML format, whereas PlaSMo supported earlier stages of private model development and sharing among collaborators.

We chose FAIRDOMHub as the resource to host PlaSMo data[13]. FAIRDOMHub is powered by the SEEK platform for managing systems biology data. SEEK (or FAIRDOM-SEEK) software was developed as part of the SysMO project, a 6-year trans-European initiative of over 100 biological research groups[14]. We had previously evaluated SEEK from the perspective of handling plant models, so we knew that SEEK's features aligned well with PlaSMo capabilities[15]. SEEK organizes assets following the Investigation, Study, Assay (ISA) structure[16], offering user friendly navigation over the ISA tree. We could preserve the PlaSMo identity, utilizing the additional Project concept in SEEK. Below, we refer to FAIRDOMHub when we discuss the public web data repository and to SEEK when we discuss the underlying software platform and its concepts.

We represented PlaSMo records as SEEK entities in the following way:

- Each version of PlaSMo model is represented as a separate SEEK Modelling Assay
- PlaSMo model file becomes SEEK Model
- PlaSMo images and data files become SEEK DataFiles
- SEEK Model and DataFiles are linked to a corresponding Modelling Assay
- Metadata which is not easily represented in SEEK (e.g. comments) are appended to the description text of the Modelling Assay
- For each PlaSMo model a SEEK Study is created, and the Modelling Assays representing different versions of the model are linked to that Study
- For each user who deposited a model, a SEEK Investigation is created in their name, and all Studies representing their models are linked to that Investigation (see below)
- A SEEK project named "PlaSMo Model Repository" is created and all the Investigations, Studies, Assays, Models and DataFiles are linked to it
- All SEEK entities generated for public PlaSMo models are visible to anyone in SEEK
- For private PlaSMo models the descriptions of SEEK Studies and Assays are visible to anyone in SEEK but the actual content of Model and DataFiles is hidden

The main difficulty was how to handle permissions and ownership. SEEK has a very rich and flexible access control model (in our opinion, it is the best permission model we have seen so far) and SEEK assets can be linked to user profiles as their contribution. However, to benefit from these features we would need to have SEEK accounts for all the PlaSMo users.

We could not create matching FAIRDOMHub accounts for PlaSMo users: a) we were not entitled to perform such actions on behalf of the users, b) some users already had FAIRDOMHub accounts to which they would want their assets linked. To avoid contacting all the users with a request to create FAIRDOMHub accounts, we assumed a simplified approach.

Firstly, the creator of a PlaSMo model is recorded as a text label: "other contributors" in FAIRDOMHub. Secondly, for each PlaSMo user a SEEK Investigation is created with a title matching their name. The SEEK Studies representing PlaSMo models created by a user are linked to their Investigation. In that way the models of a particular PlaSMo user can be easily accessed by navigating to the SEEK Investigation named after them in FAIRDOMHub. It also solved the issue that SEEK requires a parent Investigation for all assets and we could not create a sensible convention for this based solely on PlaSMo model description.

If a person would like to claim their models, they would contact us with their FAIRDOMHub account and we would link the whole Investigation/Study/Assay tree to that account and grant the user the manager role for those assets. That way, the models' creators could later manage their records using the SEEK UI.

PlaSMo users were always encouraged to link to their models using PlaSMo's stable URLs. In order to preserve such links, we implemented a simple URL resolver that would redirect original PlaSMo references to the appropriate records in FAIRDOMHub.

Figure 1 shows the generalized route for implementing an exit strategy for data preservation.

Implementation

We based the migration project on the existing code for the PlaSMo portal, in order to re-use the Data Access Objects (DAOs) and Data Object Model (DOM), so we only needed to implement the new data transfer logic.

We developed a Java client for programmatic communication with the SEEK REST API. Firstly, we used the available JSON request payload examples from SEEK REST API v1.7.0[17] to generate a library of SEEK DOM JavaBean classes using the jsonschema2pojo v1.0.0 tool[18]. We performed this step manually as it was a one-off project and we did not plan to keep the SEEK client in sync with the API in case it changes. Potential future work could make use of the jsonschema2pojo tool to regenerate these SEEK DOM classes automatically in the event of an update to the API.

The migration code iterates over PlaSMo models, extracting information required to generate JavaBeans corresponding to SEEK's Investigations, Studies, Assays, Models and DataFiles

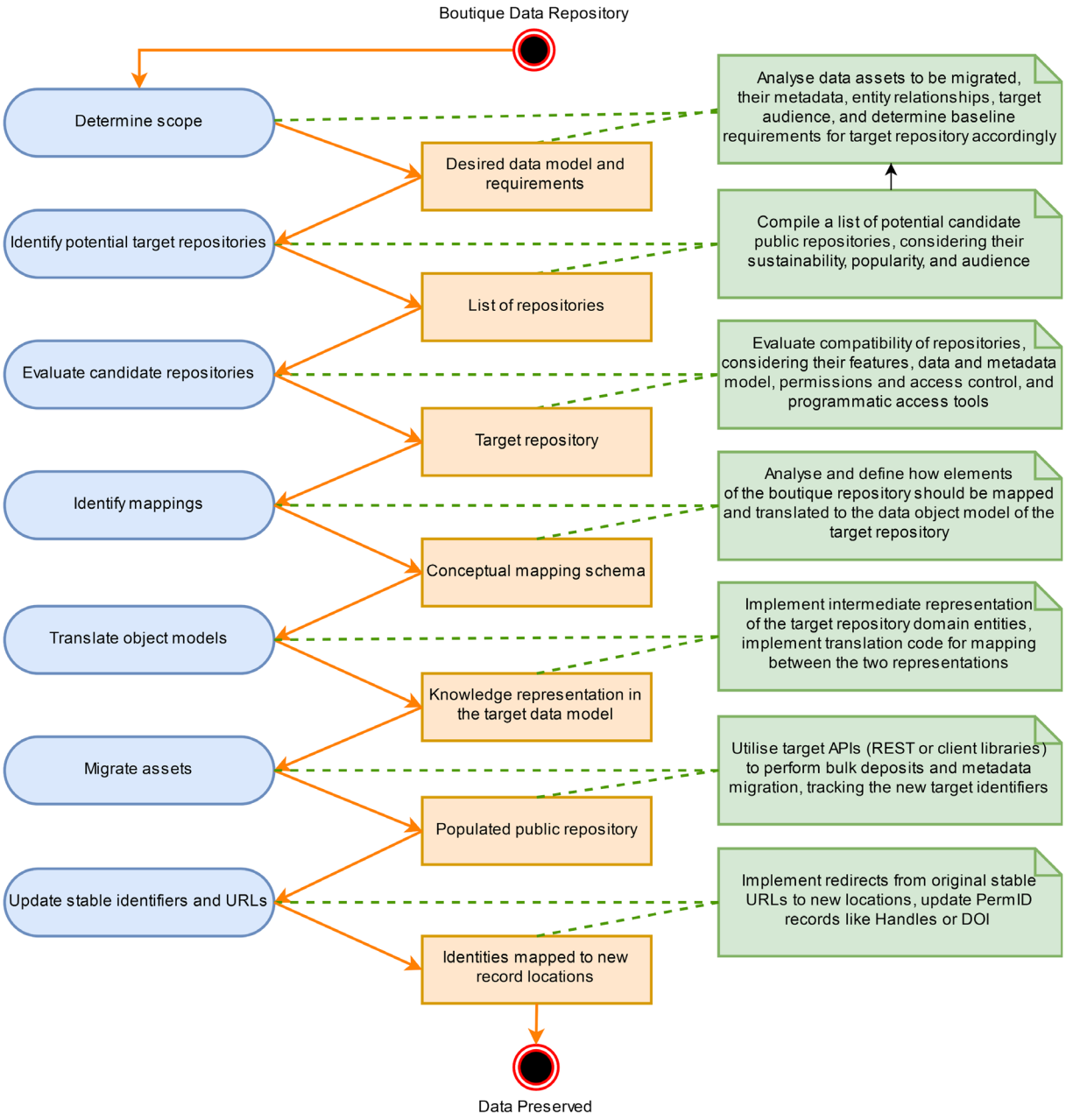

Figure 1. Implementation of an exit strategy for data preservation.

entities. It then invokes the client methods to create the entities inside the SEEK instance, which serialize the JavaBean objects into JSON and submit them to the API via authenticated HTTP POST requests. During our initial tests, not all of the REST calls were consistently successful, for example sometimes we observed HTTP status 500 errors caused on the server by SQLite3::BusyException or AbstractController::DoubleRenderError. For that reason, we decided to record the API calls in a way that would allow them to be 'replayed' if needed without a risk of creating duplicate entities, always yielding a consistent ISA tree within a SEEK instance.

We used a local SQLite database to store the SEEK API calls, which was indexed with a GUID based on PlaSMo model UID and recorded the entire JSON payload and HTTP response for each entity in the ISA tree. This request logs database was also later used to create the mapping between PlaSMo URLs and FAIRDOMHub records (see below).

The FAIRDOMHub user interface currently does not allow for setting properties (e.g. permissions) on the whole ISA tree, a feature necessary for our migration strategy. We implemented such bulk operations in a separate Java project, which

retrieves part of the ISA tree and sets the required properties – recursively through all child entities, if desired – using the Java API client.

The recorded API calls were used to generate a mapping between PlaSMo and FAIRDOMHub identifiers. The mapping was stored as a simple csv file for ease of potential future updates. This mapping is used by PlasmoMapper (a simple SpringBoot application) which resolves original PlaSMo URLs and redirects to the correct records in FAIRDOMHub.

Results

We performed the migration on 10th of January 2019. All the information from the PlaSMo portal are available under the PlaSMo project on the FAIRDOMHub. The migration process was smooth and we did not experience any problem with the API calls. It seems that the SEEK instance on the FAIRDOMHub production server is very robust and it handles all the requests flawlessly, unlike the test SEEK's Docker containers that we used during development.

Figure 2 shows the FAIRDOMHub record for version 3 of PlaSMo model 64 (Arabidopsis_clock_P2011) represented as SEEK Assay 840. The description contains the merged version specific information with the details from the main PlaSMo record (Figure 2 A, B). The other versions of that model are stored as the sibling Assays 838-841 (Figure 2 C). Each version has the list of related model and data files (Figure 2 C, D). All the

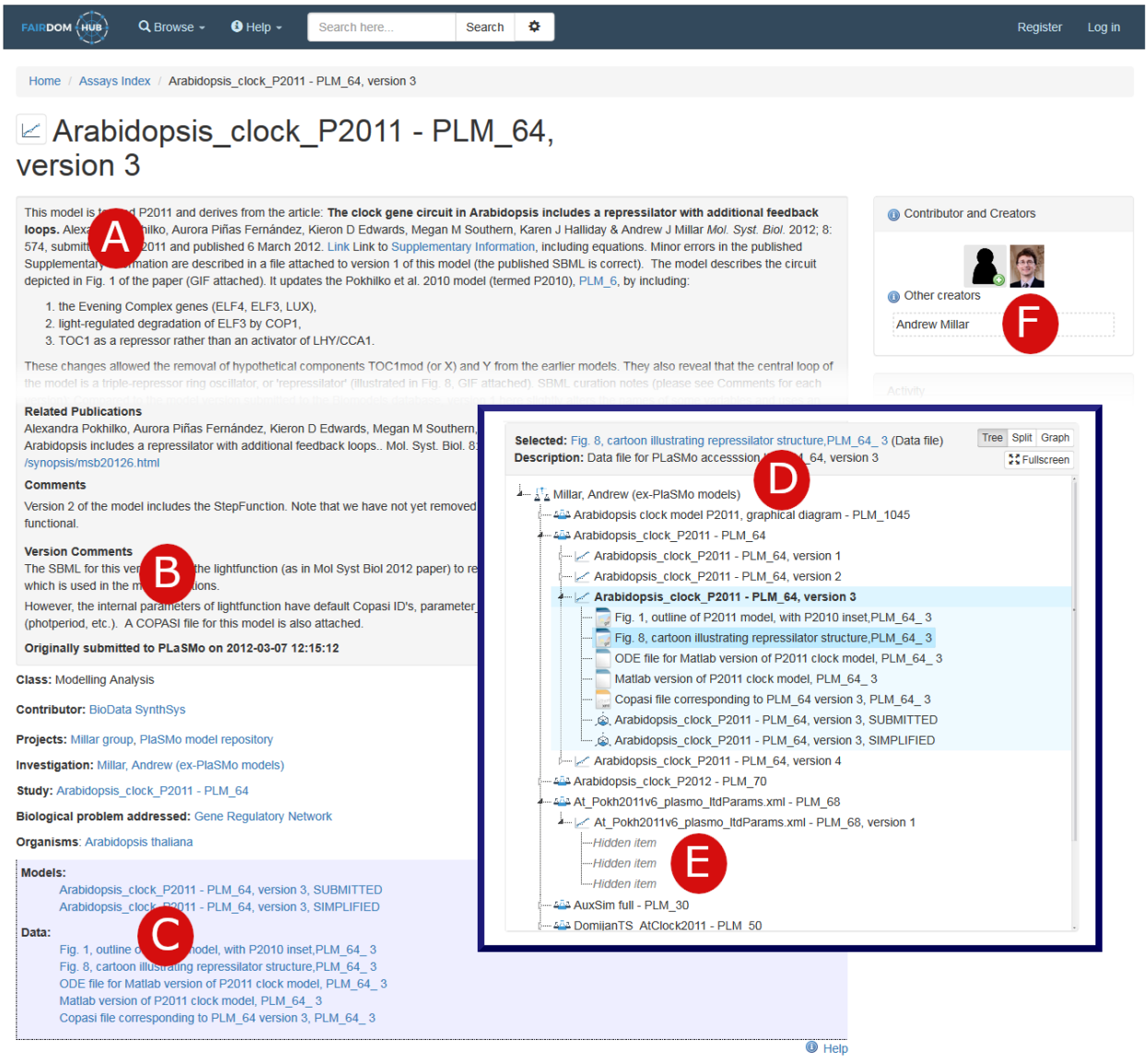

Figure 2. Screenshot (edited) presenting the FAIRDOMHub record for version 3 of PlaSMo model 64. A) Description of model 64 created from the main PlaSMo metadata; **B**) Version 3 specific details; **C**) List of linked model and data files; **D**) the navigation tree for models, versions and their data files; **E**) for private models the data and model files are hidden but the main metadata record is visible; **F**) link to the FAIRDOMHub user profile of the owner of the original PlaSMo model.

model artefacts have been linked to the model owner profile in FAIRDOMHub (Figure 2 F) by running the developed SEEK Bulk Update after the migration process.

All possible PlaSMo URLs are being redirected to the corresponding records in FAIRDOMHub, for example, the main link to the model 67: http://plasmo.ed.ac.uk/plasmo/models/model.shtml?accession=PLM_64 is redirected to Study 494; the version 3 of the model: http://plasmo.ed.ac.uk/plasmo/models/model.shtml?accession=PLM_64&version=3 to Assay 840 and the file containing Matlab version of this model: http://www.plasmo.ed.ac.uk/portal_data/data/PLM64/data/98mod_P2011.m to DataFile 2499.

We believe that the plant systems biology community will benefit from the PlaSMo models migration. The models are readily available for discovery by the larger userbase of FAIRDOMHub and models can be linked to experimental data. The potential for discovery is additionally enhanced by visibility of all the descriptions even of the private models, though for private models, the actual files are not accessible. That paves the way to potential collaborations without compromising the confidentiality of the data and is only possible due to SEEK's rich permissions model. We note that this capability fulfills the stringent data sharing guidelines of UKRI-EPSRC.

We also feel that the migration boosted the profile of FAIRDOMHub as a community resource for data management and sharing. As shown in Table 1, transfer of the PlaSMo models substantially increased the number of available modelling assets (75% increase in model files). The effect of scale is an important aspect for attracting new users and the inclusion of plant models may popularise FAIRDOMHub among modellers.

Discussion: Sustainability of Biological Data Management

We imported all the research assets from the boutique PlaSMo resource into a larger community resource: FAIRDOMHub. The migration process became feasible only after the developers of the underlying software, FAIRDOM-SEEK, released their write API in 2018. This illustrates the importance of write APIs for data management and system integration. Even with the API, the process involved laborious creation of code artefacts and mapping between concepts. We next consider how these might be avoided.

The enterprise world solved the issue of systems integration twenty years ago using SOAP[19] and WSDL[20] for web services. The enterprise ecosystem of JavaEE and .NET applications offered streamlined or even transparent generation of DOM, clients and endpoints based on the formal contract defined in XML documents. Unfortunately, the popular, new programming languages of the web, like Python, Ruby and JavaScript, lacked good support for XML and enterprise features, and as a result XML Web Services fell out of fashion. Additionally, the main driver behind the REST API is data consumption by a JavaScript UI, which has made JSON into the default exchange format. The lack of formality in the JSON REST API has been recently addressed by JSON Schema[21], OpenAPI[22] and JSON:API[23] tools, which should permit a similar level of automation as was achieved earlier by XML Web Services.

Semantic Web technologies and the Resource Description Framework (RDF)[24] specifically are meant to solve the issue of mapping between concepts. Indeed, SEEK represents its content as a knowledge graph in its backend, RDF triple store. However, the semantic mapping will only work automatically if both source and destination are using the same underlying ontologies. For example, SEEK uses the standard Dublin Core[25] for provenance terms but a custom JERM ontology[26] to document assets. Supervised translation and mapping between concepts or ontology terms seems unavoidable for the foreseeable future.

The problem of sustaining funding for research data resources is a general one, which has affected even the foundational resources of large communities such as The Arabidopsis Information Resource (TAIR)[27]. Alternative funding models have been considered, including institutional and individual subscriptions, freemium, licencing for commercial use, advertising, crowdfunding or donations[28–34]. However, these are generally applicable only to services with a large user-base, not to specialised, boutique resources such as PlaSMo. The experience of shutting down such a community repository, while preserving its data, challenges some popular views of the feasibility of Research Data Management. For example, the successful migration of all the PlaSMo data to FAIRDOMHub could suggest that there is no need to fund any new systems for data management. A future project like PlaSMo should simply use the existing FAIRDOMHub from the start.

Table 1. The total counts for each ISA entity type as they were in FAIRDOMHub before and after the PLaSMo models migration.

ISA Entity	Pre-Migration Total	Post-Migration Total	Total from PLaSMo	Percentage Increase
Investigation	197	212	15	7.6
Study	383	470	87	22.7
Assay	618	738	120	19.4
Model	255	446	191	75.0
Data File	1908	2097	189	9.9

Should funders still fund new software for data management?

In short, we believe the answer is yes, because this software can support and motivate both data sharing and research productivity.

Convincing researchers to invest the effort necessary to describe and deposit their data into a repository is the most difficult aspect of data management and a limiting factor in the wider adoption of data sharing. Data sharing can be achieved by using either "a stick" or "a carrot" approach.

The most successful "sticks" are strictly-enforced publication policies, as illustrated by the domain-specific requirements to deposit protein structural data (as in Protein Data Bank[35,36]), sequencing data (as in GeneBank[37,38]) or transcriptomics data (as in ArrayExpress[39,40]). However, these repositories handle only narrow or single data types; there is consensus within each community on the minimal information criteria; in our opinion, these are the "easy" cases. For example, pdb files are in practice self-contained with metadata principally generated by equipment or processing software and require minimal interference from a scientist. Alternatively, the deposited file might represent all the results of a large, expensive experiment (e.g. a microarray study), so the effort in preparing and describing the data for deposit is small relative to the total effort in the experiment.

The "hard" cases include those that require detailed, user-generated information, for example a description of the biological materials and the experimental treatment of a specimen before measurements began. Hard cases will have multiple, complex experimental factors that vary significantly between samples, though each sample returns a modest data volume. Data consumption can also be harder in these cases, where data volumes and the complex relationship between individual entries complicate retrieval and analysis.

The current incentives ("carrots") for data sharing are weak, considerably delayed in time and often accrue more to group leaders than to contracted researchers, hence they do not encourage widespread adoption of sharing practices[41]. An alternative approach is to incorporate data management into the daily research workflow, by providing immediate value to data producers in the form of increased productivity, specialized processing, visualisation or data aggregation. For example, the BioDare repository is widely used within the circadian community, but researchers primarily use the resource to access specialist software tools to analyse and visualise their timeseries data, so the fact that BioDare datasets are simultaneously deposited in the public domain is in reality a side effect of the researcher's core activity[42,43]. This level of tool customization and integration is project/domain specific[44,45] and not possible with general repositories. Consequently, we expect such "carrots" to be rare among repositories that cater for many research domains, such as institutional repositories.

User friendliness is the most important characteristic for successful data management. The development of user friendly solutions that facilitate research (providing the specialist "carrots" we describe above) remains a valid case for new funding.

It is worth noting, that data management solutions may not need to be built entirely from scratch. One could leverage features of existing products (like for example SEEK or OpenBIS[46]) and create plugins or integrate with them. Which approach is most cost effective and productive must be evaluated case by case, depending on the available know-how and expected user experience.

A positive example of promoting data management is the Wellcome Trust "Research Enrichment – Open Research" scheme[47], which funds small, add-on projects for existing grant holders to enable open research and data sharing. By presenting this as add-on funding, the implementation of data sharing is perceived as an additional opportunity, rather than in competition with core scientific activities for funding.

Can research data repositories be self-sustaining?

In majority of the cases, no.

The idea that domain-specific resources could often be maintained from subscription fees is unrealistic[29,33,34]:

1. There is a problem of scale. If we advocate for resources that address particular needs of scientific projects, the underlying user base or even the entire research community may be too small to sustain a public repository financially. Conversely, repositories catering to a diverse community may gain sustainability but lack user uptake.

2. Data producers already commit their time and make a substantial effort to prepare data for deposit, so we cannot expect them to be charged for deposit on top of the work they do to contribute their data.

3. Charging for access to data is against the spirit of open research and data re-use. Funding agencies generally require the public release of results, or are moving to do so, and such a model would be an infringement of their policies.

4. Micropayment models, with small fees for extra features that one might use (e.g. minted DOI or a longer embargo period) could be acceptable to the users but it is impractical in the academic world. Research groups often do not have credit cards to perform small payments automatically, but invoicing and accounting for such operations would be problematic and not cost-effective.

While it is possible to secure funding for a new project, there are currently few funding schemes to maintain existing data resources. Incremental improvements to existing resources can also be problematic, as they may not offer the novelty and impact to compete with new infrastructure. However, maintaining existing resources may be as important as funding new science as it is the only way to enable data re-use. At the same time, the data repositories should gather metrics in order to demonstrate their value, for example numbers of active users, visits and downloads.

How to deliver data longevity?

Our PlaSMo migration demonstrates that data longevity can be achieved by implementing an effective exit strategy. In our case, we found a close match for our metadata model in the FAIRDOMHub. If a sufficient match is not available, it is always possible to find a generic destination that can at least archive all the data. The implementation of a migration involves additional costs but in the long term, it is usually cheaper than maintaining a running resource.

The biggest value of data repositories lies in their data; hence, we would recommend creation of funding opportunities that could be used to "rescue the data". Data migration could constitute part of the income agreement for maintaining destination repositories. For example, a repository could receive extended funding on the condition that it would implement the adoption of data from other projects.

Currently, data migration seems to be an inevitable reality of data preservation. Permanent identifiers (like DOIs or handles) which can resolve to the actual location facilitate this process. If PlaSMo models had DOIs we would not need to deploy PlasmoMapper to handle the original URLs. Unfortunately, participation in permanent identifier schemes can incur additional financial costs, which paradoxically may accelerate the demise of a repository.

In Horizon2020, the EU funded various initiatives to provide European Research e-Infrastructure, and participating consortiums offer permanent identifiers as part of their services (EUDAT[48], ePIC[49]). Sadly, although the initiative is already centrally funded, the identifiers (handles in this case) were provided only as a paying service. More domain-specific projects have been funded to offer free IDs, such as the identifiers.org service at EMBL EBI, which is now linked to the ELIXIR project. We believe that permanent identifiers should be available free of charge not only for data projects but even for individuals as a public service, similar to street address systems.

Conclusions

We shared our experience in securing the PlaSMo project's legacy and assuring data longevity by successfully implementing an exit strategy in the form of data migration. We believe that further progress in open research and data sharing can only be achieved by integration between different resources that together can be incorporated into research workflows. We are concerned over the existing funding opportunities for data management and how they might put at risk the safety of scientific data.

Reuse potential

The Java Client for the SEEK REST API and the bulk property setter, described here, can be of value for other projects. The client can be used to integrate other Java projects with SEEK, for example to automate data deposition. The bulk property setter compensates for a currently missing feature in the SEEK UI. Running the setter is currently the easiest way to publish multiple datasets constituting a research outcome. For these reasons we made the relevant code available as two separate packages.

Data availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Java Client for SEEK API

Source code is available from: https://github.com/SynthSys/Seek-Java-RESTClient

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3250951[50]

Licence: MIT

SEEK Bulk Update

Source code is available from: https://github.com/SynthSys/Seek-Bulk-Update

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3250959[51]

Licence: MIT

PlaSMo portal

Source code is available from: https://github.com/SynthSys/PlasmoPortal

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3250855[52]

Licence: MIT

References

1. HEFCE, RCUK, UUK, *et al.*: **Concordat On Open Research Data.** 2016.
Reference Source
2. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**: 160018.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Wittig U, Rey M, Weidemann A, *et al.*: **Data management and data enrichment for systems biology projects.** *J Biotechnol.* 2017; **261**: 229–37.
PubMed Abstract | Publisher Full Text
4. Stuart D, Baynes G, Hrynaszkiewicz I, *et al.*: **Practical Challenges for Researchers in Data Sharing.** *Whitepaper.* 2018; 30.
Publisher Full Text
5. Knowledge Exchange Research Data, Expert Group and Science Europe Working

Group, on Research Data: **Funding research data management and related infrastructures.** 2016.
Reference Source

6. **Figshare.**
Reference Source

7. **Zenodo.**
Reference Source

8. **Dryad.**
Reference Source

9. **Edinburgh DataShare.**
Reference Source

10. **UK Data Archive.**
Reference Source

11. **BioModels.**
Reference Source

12. Glont M, Nguyen TVN, Graesslin M, *et al.*: **BioModels: expanding horizons to include more modelling approaches and formats.** *Nucleic Acids Res.* 2018; **46**(D1): D1248–53.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Wolstencroft K, Krebs O, Snoep JL, *et al.*: **FAIRDOMHub: a repository and collaboration environment for sharing systems biology research.** *Nucleic Acids Res.* 2017; **45**(D1): D404–D407.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Wolstencroft K, Owen S, Krebs O, *et al.*: **SEEK: a systems biology data and model management platform.** *BMC Syst Biol.* 2015; **9**(1): 33.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Troup E, Clark I, Swain P, *et al.*: **Practical evaluation of SEEK and OpenBIS for biological data management in SynthSys; first report.** University of Edinburgh. 2015.
Reference Source

16. Rocca-Serra P, Brandizi M, Maguire E, *et al.*: **ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level.** *Bioinformatics.* 2010; **26**(18): 2354–6.
PubMed Abstract | Publisher Full Text | Free Full Text

17. **SEEK REST API.**
Reference Source

18. Littlejohn J: **jsonschema2pojo [Internet].**
Reference Source

19. **https://www.w3.org/TR/soap12/**

20. **https://www.w3.org/TR/wsdl20/**

21. **http://json-schema.org/**

22. **https://swagger.io/specification/**

23. **https://jsonapi.org/**

24. **https://www.w3.org/TR/rdf11-concepts/**

25. **https://www.dublincore.org/specifications/dublin-core/**

26. **https://jermontology.org/**

27. Reiser L, Berardini TZ, Li D, *et al.*: **Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model.** *Database (Oxford).* 2016; **2016**: pii: baw018.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Gabella C, Durinx C, Appel R: **Funding knowledgebases: Towards a sustainable funding model for the UniProt use case [version 2; peer review: 3 approved].** *F1000Res.* 2017; **6**: pii: ELIXIR-2051.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Chandras C, Weaver T, Zouberakis M, *et al.*: **Models for financial sustainability of biological databases and resources.** *Database (Oxford).* 2009; **2009**: bap017.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Wilcox A, Randhawa G, Embi P, *et al.*: **Sustainability considerations for health research and analytic data infrastructures.** *EGEMS (Wash DC).* 2014; **2**(2): 1113.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Özdemir V, Badr KF, Dove ES, *et al.*: **Crowd-funded micro-grants for genomics and "big data": an actionable idea connecting small (artisan) science, infrastructure science, and citizen philanthropy.** *OMICS.* 2013; **17**(4): 161–172.
PubMed Abstract | Publisher Full Text | Free Full Text

32. Kitchin R, Collins S, Frost D: **Funding models for Open Access digital data repositories.** *Online Inform Rev.* 2015; **39**(5): 664–681.
Publisher Full Text

33. OECD: **Business Models for Sustainable Data Repositories.** OECD Science, Technology and Innovation - Policy Papers. 2017.
Reference Source

34. RDA-WDS Interest Group on Cost Recovery for Data Centres, Dillo I, Hodson S, *et al.*: **Income Streams for Data Repositories**. 2016.
Publisher Full Text

35. **Protein Data Bank.**
Reference Source

36. Berman HM, Battistuz T, Bhat TN, *et al.*: **The Protein Data Bank.** *Acta Crystallogr Sect D Biol Crystallogr.* 2002; **58**(Pt 6 No 1): 899–907.
PubMed Abstract | Publisher Full Text

37. **GeneBank.**
Reference Source

38. Benson DA, Cavanaugh M, Clark K, *et al.*: **GenBank.** *Nucleic Acids Res.* 2013; **41**(Database issue): D36–42.
PubMed Abstract | Publisher Full Text | Free Full Text

39. **ArrayExpress.**
Reference Source

40. Brazma A, Parkinson H, Sarkans U, *et al.*: **ArrayExpress--a public repository for microarray gene expression data at the EBI.** *Nucleic Acids Res.* 2003; **31**(1): 68–71.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Van den Eynden V, Knight G, Vlad A, *et al.*: **Towards Open Research: practices, experiences, barriers and opportunities [Internet].** 2016.
Publisher Full Text

42. **BioDare.**
Reference Source

43. Zielinski T, Moore AM, Troup E, *et al.*: **Strengths and limitations of period estimation methods for circadian data.** *PLoS One.* 2014; **9**(5): e96462.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Kılıç S, Sagitova DM, Wolfish S, *et al.*: **From data repositories to submission portals: rethinking the role of domain-specific databases in CollecTF.** *Database (Oxford).* 2016; **2016**: pii: baw055.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Leonelli S, Smirnoff N, Moore J, *et al.*: **Making open data work for plant scientists.** *J Exp Bot.* 2013; **64**(14): 4109–4117.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Bauch A, Adamczyk I, Buczek P, *et al.*: **openBIS: a flexible framework for managing and analyzing complex data in biology research.** *BMC Bioinformatics.* 2011; **12**: 468.
PubMed Abstract | Publisher Full Text | Free Full Text

47. **Research Enrichment – Open Research.** [Online]: Wellcome Trust;.
Reference Source

48. **https://www.eudat.eu/**

49. **https://www.pidconsortium.eu/**

50. Zielinski T, Hay J: **SynthSys/Seek-Java-RESTClient: Java RestClient for SEEK API 1.7.0 (Version v1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3250951

51. Zielinski T, Hay J: **SynthSys/Seek-Bulk-Update: Bulk Update For Seek API 1.7.0 (Version v.1.0.0).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3250959

52. Zielinski T, Tindal C: **SynthSys/PlasmoPortal: The last working version of PlaSMo portal (Version v2.1.5).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.3250855

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 22 July 2019

https://doi.org/10.21956/wellcomeopenres.16751.r35895

✔ **Robert Davey**
Earlham Institute, Norwich, UK

The authors describe a mechanism for preparing and executing a "handover" strategy to ensure long(er) term hosting and sustainability of biological datasets. This is presented through the authors' experiences in carrying this out for their own project, PlaSMo. Whilst there isn't much that is particularly novel about this article, it could be incredibly useful for those groups or consortia that are approaching a similar cliff face in funding support for their data repository or infrastructure.

My main concern is that the article doesn't really cover the breadth of information in the literature concerning other projects that have gone through a similar procedure. Araport and TAIR is one such example, and relevant to the funding sustainability argument. The references provided are generally suitable, but many of them (those that aren't referencing a public data repository) are from one of the authors. I understand this from the point of the use case - it's sensible to provide supporting evidence for their tools. However, there are other recent publications regarding plant-based information resources that could be mentioned to cover a wider scope. A more comprehensive view of issues in other plant science circles, maybe even other domains of science, would be good, given the article's quite broad title. There is a huge amount of literature on open science, reproducibility, the availability of resources and how to handle data openly to ensure longevity.

That said, the article makes salient points that are worthy of thought. This article would be of use to funders currently, who are facing a real problem on their hands as data infrastructure is increasingly needed at a time when funds for creating and sustaining these resources aren't increasing accordingly.

The article is clearly written and very easy to follow with concise points and no lengthy prose. It was nice to read, so the authors should be commended here.

I would have liked to have seen a critical view of their own exit strategy. Key parts of their strategy are often overlooked in practice when considering day-to-day actions of scientists and infrastructure developers. Are there pros/cons to this strategy, and are there other solutions available that can assist with certain "pain points"? The mapping/translation/knowledge representation process could be expensive and time-consuming. Are there tools that can assist? If not, why not?

This is particularly relevant when discussing the carrots and sticks argument. What are the "hard cases"? High-throughput plant phenotyping could be one such example, as even though there is developing consensus, there are still many issues around the longevity and usefulness of storing these datatypes. Exploring this a little more may give a helpful counterpoint to why the easy cases aren't simply about support for narrow/single datatypes. Handling this heterogeneity of data outputs, automated or manually curated, is a key issue for many scientists, and indeed feeds into the user friendliness issue.

Is the ubiquity of paid-only permanent identifiers true for all EU projects? I wasn't aware of this, so a reference to this would be useful for readers to understand this further. Some data repositories can provide DOIs for free, so is there a part of the exit strategy that could include wholesale deposition/mirroring of datasets within such a repository?

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data management, bioinformatics, crop genomics, semantic web and ontologies, e-infrastructure, high-performance/high-throughput computing, policy and strategy for data science and open science.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 17 Sep 2019

Tomasz Zielinski, University of Edinburgh, Edinburgh, UK

Thank you for your extensive and constructive review. Our detailed responses are listed below, with pointers to the revised paper text.

Reviewer: “My main concern is that the article doesn't really cover the breadth of information in the literature concerning other projects that have gone through a similar procedure. Araport and TAIR

is one such example, and relevant to the funding sustainability argument. The references provided are generally suitable, but many of them (those that aren't referencing a public data repository) are from one of the authors. I understand this from the point of the use case - it's sensible to provide supporting evidence for their tools. However, there are other recent publications regarding plant-based information resources that could be mentioned to cover a wider scope. A more comprehensive view of issues in other plant science circles, maybe even other domains of science, would be good, given the article's quite broad title. There is a huge amount of literature on open science, reproducibility, the availability of resources and how to handle data openly to ensure longevity."

To address these comments, we have added new text at the start of the Discussion. One paragraph (starting "*The problem of sustaining funding for research data resources is a general one,…*") describes the funding alternatives, with some additional references including the example of TAIR, which we agree is a relevant example.

Note, first, that this article was accepted in the Research Note format, described as a "one figure paper" in the Wellcome guidelines. It was not possible or appropriate to cover the whole landscape of open research and repository funding. Rather, we presented the particular challenges that we faced as a case study for a broadly-relevant problem, how we addressed these issues and what reflections this case provoked on the sustainability of research data sharing more generally. We ensured that the methods and conclusions are as general and reusable as widely as possible. The title reflects our use case and the content of the paper.

Second, we focus on boutique repositories that address specific needs of a particular community, in this case crop and plant systems modelling. This is a far smaller resource than Araport and TAIR, which serve all of plant science and some of the broader genomics community. Their much larger scale made a subscription model possible for TAIR, which would not work for smaller but still valuable resources, as we comment later in the discussion.

Reviewer: "The references provided are generally suitable, but many of them (those that aren't referencing a public data repository) are from one of the authors."

We had cited only two of our articles among a total of 31 references, plus 3 links to code snapshots that were requested by the Wellcome editors. This did not seem excessive.

Reviewer: "I would have liked to have seen a critical view of their own exit strategy. Key parts of their strategy are often overlooked in practice when considering day-to-day actions of scientists and infrastructure developers. Are there pros/cons to this strategy, and are there other solutions available that can assist with certain "pain points"? The mapping/translation/knowledge representation process could be expensive and time-consuming. Are there tools that can assist? If not, why not?"

We largely agree that parts can be overlooked, which is why we felt it worthwhile to document the process and the reasoning in detail, and we focussed on a specific case to do so. The new Discussion text addresses tools to ease the pain points (starting "We next consider how these might be avoided"), specifically automating the code generation and knowledge mapping, with a comparison of current and earlier tools. We had discussed alternatives for particular stages but for the process overall, the alternatives are not very helpful. They amount to either providing the service without financial support (the *status quo ante*) or closing the service and with it the data

access.

Reviewer: “This is particularly relevant when discussing the carrots and sticks argument. What are the "hard cases"? High-throughput plant phenotyping could be one such example, as even though there is developing consensus, there are still many issues around the longevity and usefulness of storing these datatypes. Exploring this a little more may give a helpful counterpoint to why the easy cases aren't simply about support for narrow/single datatypes.”

We now discuss the “hard” cases in a paragraph starting “The “hard” cases include those that require detailed, user-generated information...”. Phenotyping data are “big data” both in volume and velocity, so the main challenge is the infrastructure (e.g. storage and bandwidth). Data deposition is relatively easy as data are collected automatically and there is minimal need for human curation. Data reuse will depend on developing a consensus on the metadata for recorded values, and as the reviewer notes this consensus is emerging but not complete. We see the main challenge in this kind of data in reassembling suitable datasets for analysis from a large volume of related recordings.

Reviewer: “Is the ubiquity of paid-only permanent identifiers true for all EU projects? I wasn't aware of this, so a reference to this would be useful for readers to understand this further.

As we state in the manuscript, permanent IDs are not freely available in general. We added examples of partially funded EU projects that provide PID as a paid service. The costings in detail vary considerably, as the DOI FAQ implies: “The cost of registering new DOI names depends on the services you purchase”. Handle system has a different charging structure, see http://www.handle.net/payment.html. More domain-specific projects have been funded to offer free IDs, such as the identifiers.org service at EMBL EBI, as we note in the Discussion.

Reviewer: “Some data repositories can provide DOIs for free, so is there a part of the exit strategy that could include wholesale deposition/mirroring of datasets within such a repository?”

Having a new, free DOI minted by a destination repository does not help in preserving the previous links or ID, whereas having a PID to start with allows a transparent redirect to the new location of a migrated resource.

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 July 2019

https://doi.org/10.21956/wellcomeopenres.16751.r35896

Helen Ougham
Aberystwyth University, Aberystwyth, UK

This paper represents both a useful case study in migrating a biological research resource - in this case, the plant models repository PlaSMo - to a new public repository - and a description of the serious practical, especially financial difficulties, of maintaining specialised datasets. Data sharing, though both inherently desirable and a requirement of many funders, is not a trivial exercise in cases where no major international repository/database is available; there is often a real danger that research outputs will be either lost or made available in a form unsuitable for effective re-use.

The original PlaSMo portal, which used computational features appropriate to its time, was designed as a resource for plant and agricultural researchers to access a range of plant models, some very new, but others long-established but in danger of being lost as their originators retired and, in some cases, source code became difficult to access. It was always intended that those in the rapidly-developing systems biology community should be able to capitalise on these models to assist them in extending modelling of biological processes from the cellular to the organ, whole plant and ultimately crop level. Now, some years later, the number of plant systems biology models is much greater, and it made sense for the authors to migrate the models and datasets originally available through PlaSMo to a contemporary repository already accommodating many systems biology models and in use by current systems biologists. At the same time, this would secure the classic crop models for the plant physiologists and breeder. Carrying out this migration addressed issues of potential security threats as well as reducing the overhead inherent in maintaining the PlaSMo resource in its original form.

The paper clearly sets out the rationale for the work, the steps taken, the tools used, and the form in which the original PlaSMo models and associated files are now to be found in FAIRDOM hub; the latter has grown considerably as a result, particularly with respect to the number of models available. Although certain aspects of the original PlaSMo have been lost (ability to run web-based simulations, for example), this is an inevitable consequence of the move and is unlikely to adversely affect most current users of the models.

The paper is generally very well written; there are a few sections where the English reads a little as though it was written by a non-native speaker, but the meaning is always very clear.

A couple of minor typos: there is one instance where PlaSMo is written PlaSMO, and this should be amended for consistency; and 'GeneBank' should be 'GenBank'.

I did notice a spelling mistake on https://github.com/SynthSys/PlasmoPortal, where FAIRDOMhub is shown as FaridoHub!

All the URLs in the article worked correctly at the time of this review.

The table and the two figures are useful and appropriate. In Figure 1 the boxes on the right (in green) have 'folded over' corners which in some cases slightly obscure the text; this should be easy to address.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: I was, with one of the authors (Andrew Millar) a grantholder on the BBSRC-funded project which established the original PlaSMo repository. However I retired from Aberystwyth University in 2010 and believe that I am able to give a fair and unbiased review of this paper, which has given me an interesting and useful update on progress in the area.

Reviewer Expertise: My background is in plant science (including crop science) and bioinformatics, but not in computer science. As a grantholder on the original PlaSMo project, I am able to assess the background to the work, its curret value, and the form in the PlaSMo models and associated files have been migrated and made available, but I am not able fully to assess the computational infrastructure aspects.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 17 Sep 2019

Tomasz Zielinski, University of Edinburgh, Edinburgh, UK

Thank you for your very positive review. We have corrected the typos and the layout of Figure1 as suggested.

Competing Interests: No competing interests were disclosed.