FUNDING, SUSTAINABILITY, AND GOVERNANCE

Matti Heikkurinen

Director, Emergence Tech Ltd, Kent, UK
Email: matti@emergence-tech.com

1 STATE OF THE ART

Discussions about funding, sustainability, and governance of any e-infrastructure service tend to be culmination points of the tensions between bottom-up and top-down, static and dynamic, and prescriptive and descriptive approaches. Data management brings in additional challenges due to the time aspect. While a computing centre or network technology might be considered obsolete after a few years, data management requirements often extend decades or even centuries into the future. Funding commitments need to be made for long periods of time, and decisions about sustainability and governance models could be seen as setting a precedent that ties down parties for years into the future. At the same time, it is anticipated that the amount of data being produced will continue to exceed the capacity of the managed (curated and otherwise maintained) data services, adding the need to discuss prioritisation of the different datasets.

For the purposes of this report, it is useful to present a quick summary of the current state of play separately in the three domains under discussion, before looking into their interactions and dependencies. The treatment is fairly EC-centric; however, the following basic patterns are likely to be applicable on the global scale as well.

- **Funding** of the data services is seen as important, and there is a near-consensus about the need to secure funding for storing data. However, what the source of this funding should be and the criteria on which the funding should be distributed are less clearly defined. For example, it has been suggested that as part of the research grant process, a research group would need to present a plan for long-term data management. However, whether this plan would require explicitly allocating a part of the research grant for this purpose or just indicating a credible "permanent" repository will already provide challenges in reaching consensus. In Europe, the first "infrastructure"- level investments (e.g., in the form of the EUDAT project (http://www.eudat.eu)) have been announced. This should increase the policy-level awareness of data management as a cross-cutting issue that should be studied and managed as an entity that is not subservient to other established e-Infrastructure services or to the research agendas of individual research disciplines.
- **Sustainability** of the data management is harder to evaluate. Sustainability is usually mentioned as something that is an inherent aspect of the funding commitments. However, it is often hard to see whether this assumes mere sustainability of "storing of old data" (which is likely to become considerably less expensive in the future) or also long-term curation of the data (which will not benefit from improvements in the price/capacity ratio of the underlying technologies to the same degree). Thus the large-scale initiatives that provide data services will likely have to engage in major awareness-raising actions with regard to cost structures, investments needed, and benefits (ones that can be anticipated as well as ones that appear through fortuitous discovery) of sustainable data services.
- **Governance** issues – beyond discipline-based solutions – are in an early stage of development. This area will be influenced by the surrounding regulatory framework (e.g., through regulations related to privacy or EC rules related to the reuse of the public sector information or implementation of the *Digital agenda* flagship initiative (http://ec.europa.eu/information_society/digital-agenda/index_en.htm)), as well as by the discussions about the value of the data for new user communities. It should be noted that even in the corporate domain, data governance is a young and evolving discipline, with some sources saying that communication efforts make up 80-95 percent of a successful data governance project (http://www.information-management.com/issues/2007_48/10001356-1.html). Identifying and engaging with the necessary stakeholders and building a common framework for decision making in the more diverse environment of research e-Infrastructures is likely to require similar emphasis on communication.

Looking at the interplay of these three issues, there are certain inherent differences regarding what the most natural approach would be. Initiating discussions related to long-term funding and sustainability necessitates some top-down decisions and agreements that influence the available choices for governance models. Securing long-term commitments from funding agencies – especially commitments that exceed the lengths of their typical funding cycles – is feasible only if many aspects of the overall sustainability framework can be described as constant. At the same time, many of the requirements for the data infrastructures described in the other reports (such as security, privacy, long-term curation, and interoperability) have been the topics of active discussion quite recently, with different approaches being considered. The results of these (and several other) discussions need to be taken into account in the governance models, too. Even in the 2020 timeframe, it is possible that neither a consensus opinion about the best paradigms will emerge, nor an agreement on whether a common model is even possible for all kinds of data and applications. Thus preparedness to accommodate new bottom-up approaches as tests for alternative approaches is needed. The importance of involving users in the governance model is not limited to data infrastructures, as reflected in the e-IRG White Paper 2011, for example. The governance section of this white paper states that "active participation of users with leading edge service requirements in strategic governance decisions concerning e-infrastructures is essential" (e-IRG White Paper 2011, p. 9).

The overall goal of the data infrastructure – increasing the advantages and benefits derived from data, for example through encouraging, combining, and deriving them into datasets that support new kinds of use cases – poses an additional challenge that is related to the dynamics of the value generation. New use cases and user communities mean that the value of stored research data could increase in an unanticipated manner in the future more often than today. As mentioned in the report discussing open data policies, research data gain value with use, and improvements in the interoperability and metadata descriptions of the data may lead to a situation in which much of the increased use emerges from sources outside formal collaborations and stakeholder groups that were known when data were gathered. However, the resulting increase in value can also increase the operational costs of the archive holding the data, sometimes considerably. Thus an additional challenge in finding the most appropriate funding and sustainability models lies in being able to take into account the changes in the values of the dataset being managed and in being able to react to these changes – up to and including re-allocating resources to support the rapidly growing use of certain datasets.

When resources are re-allocated, some data might need to be moved to secondary archival solutions earlier than originally anticipated. This could be especially challenging if it is difficult to identify a legal entity or a project that is behind the increased need for resources. Thus there needs to be a mechanism that ensures that – as much as possible – all relevant stakeholders are consulted before any major decisions are made. Through this process, it is also possible to identify individuals and research groups who have common interests in specific datasets but who have not organised themselves as interest groups or formal collaborations and to encourage them to seek the benefits of organising themselves. After all, the goal is to create a flexible data infrastructure that offers maximal flexibility to accommodate situations that are not anticipated today.

Identifying all of the concerned stakeholders in this process might become more and more challenging. Trans-disciplinary, trans-organisational and trans-national research is going to give rise to new research groups and disciplines that do not exist today, but whose voices should be heard when decisions about the priorities and plans for the future GRDI are made. Virtualisation and Cloud-like solutions providing the underlying infrastructure as a service will also challenge the categorisation to clearly separate groups of data "providers" and "consumers". Many of the groups are simultaneously users of the data and producers of combined and/or derived datasets, acting in the role of a "prosumer". This "prosumer" role could become the norm, as reusing and cross-linking of data as the basis for products and services become more and more commonplace.

Considering the dynamic behaviour of this system, one could anticipate similarities with distributed software development projects, especially open source ones. In the case of open source software, the developers are almost without exception also the users of other open source software suites. They might modify these other suites for their own purposes and then publish their modifications (or specific configuration files) as patches, contributions to the original source repositories, or as derived works ("forks"). The amounts of raw data are quite different, but exploring whether the models that have been used to solve sustainability, funding, and governance issues in the software development domain could also be applied to data might lead to interesting insights. The best practices and experiences of several decades that culminate in success stories (e.g., the Linux (http://www.linuxfoundation.org/) and Apache (http://apache.org/) foundations and the WebKit project (http://www.webkit.org/ and http://en.wikipedia.org/wiki/Webkit)) that are supported by a wide variety of backers who are often in competition with each other, also provide information regarding the history of the project that is relevant to the discussion and could be applied to data issues.

It is also important to acknowledge the limitations of this approach. The governance models of the open source software projects vary and are hard to compare objectively because it can be difficult to separate the success of the organisation from the skills, personalities, and management styles of the project leaders. And it is not hard to find examples of planned open source governance models that have become irrelevant either through a lack of interest or through the active rejection of the developers. Furthermore, the level of liability assumed by an organisation that publishes open source software components is seen as easier to manage and limit than the level of liability in the case of provision of data management services. At the minimum, the host organisation needs to review the degree to which data that are managed by the new service can potentially be seen as personal data. Thus, several regulations related to privacy and security would come into play, separating this from software that can be provided "as is", with the IPR owned by the contributors (where employees are individuals rather than representatives of their organisations). For this reason, it is likely that a formal governance structure needs to be set up earlier in the development process than with software. In the case of data, this setup also needs to include identifying or creating a legal entity that will bear the formal responsibility for the service.

As with the open source software, the prosumer model makes it harder to anticipate and measure the benefits of a specific data management service. Storing and curating the data from the original research work might be responsible for most of the direct costs, but the actual value could emerge through multiple stages of filtering, combining, and producing derived datasets. Thus simplistic, discipline-based "return on investment" (ROI) calculations are probably not the best approach when evaluating the efficiency of investments to specific data services. This is also analogous to open source solutions that do not employ dual-licensing strategies.

The funding, sustainability, and governance challenge could be condensed into the following problem statement.

> ***Create a sustainability model that is stable enough to encourage long-term (top-down) commitments while maintaining enough flexibility in governance and funding structures to allow full participation of the emerging (bottom-up) activities.***

The awareness of the importance of data management is constantly growing, which helps generate the necessary political will for committing resources to the provision of services. The challenge is to channel this political will into a system that provides visible short-term benefits and success stories while making sure that the emerging activities that use the data services feel that they have a way to participate in the governance of the services.

2 TEN-YEAR VISION

A realistic high-level goal for the funding, sustainability, and governance models of 2020 would be the emergence of a common conceptual model capturing most of the value network supported by the GRDIs. A model that is accepted by the users, creators of the data, GRDI service providers, and funding agencies would also form a basis for effective incentive structures to support sustainability of this ecosystem. Even with limited incentives, it should be possible to support a virtuous cycle for all of the stakeholders: the more communities that join in and the more the service is used, the easier it will be to justify future investments and sustainability-related guarantees.

More detailed visions for funding, sustainability, and governance depend to a high degree on the evolution of the GRDI ecosystem and its uses. It is likely that (especially with regard to governance) the research data infrastructures need to take into account more general regulations related to the storing of data (through, for example, regulations related to privacy (Personal data protection in the EU, new approach: http://europa.eu/legislation_summaries/information_society/data_protection/si0020_en.htm) or reuse of public sector information (Public sector information, EC homepage: http://ec.europa.eu/information_society/policy/psi/index_en.htm)).

Nevertheless, by 2020 it is likely that a hybrid distributed ecosystem of disciplinary data infrastructures (e.g., HEP, Bio) and multidisciplinary ones (e.g., new areas, such as Arts and Humanities) will still coexist. Efforts to make this ecosystem interoperable will be continued in order to reuse best practices and promote effective interdisciplinary collaboration. The goals of these efforts are promoting several primarily operational improvements and enhancements, such as interoperability, data exchange, data preservation, and distributed access. Therefore, separate governance for each of the ecosystem building blocks will still exist as well as possible collaboration mechanisms among disciplines (e.g., observers of other disciplines in disciplinary bodies) and global interdisciplinary governance bodies that increasingly contribute to collaboration and interoperation strategies. Collaboration or integration with other e-Infrastructure components (such as the computing that will process the data) might appear, likely through disciplinary or multidisciplinary initiatives.

At the same time, some initial steps towards monetised metrics that estimate the value created by the use of GRDIs as a basis for the incentive schemes will be made, perhaps in the form of discipline-based pilots. These kinds of metrics will probably complement project-style funding and possibly provide tools for a rapid transition towards a sustainable service provided by a clearly identified legal entity instead of a project consortium. A possible life-cycle model would fund the new data service initially as a project ("producer push"). This would then go through a hybrid stage, in which part of the costs are covered by project-like funding and part are based on usage-based metrics, before moving to a system that is tied to metrics based on usage and creation of value.

3 CURRENT CHALLENGES

In addition to the fundamental tension between top-down and bottom-up approaches (identified also in the e-IRG White Paper 2011, p 31), there are a few concrete issues that could add to the difficulty of reaching a comprehensive funding, sustainability, and governance model.

Potential competition from emergent collaborative data management solutions that grow out of data-sharing activities of small-scale initiatives might force GRDIs to maintain a high level of awareness-raising activities regarding costs and the value of the service provided. Developers of these solutions might not start out to create a data management solution, but data management solutions might emerge from solutions that were intended to solve relatively simple problems, such as backing up or transferring large amounts of data (e.g., through sharing and transferring of physical hard drives). However, after these solutions are deployed and become part of the culture of the collaborations using them, it will be hard to compete with them, at least in their original niches. Thus, GRDIs should provide easy-to-use (with access based on standard mounted drive, database, or version management system interfaces) and inexpensive services that solve very basic needs of emerging user communities. The challenge of this approach is that GRDIs need to compete with projects that only require a few hard drives (costing 100 euros each) and manpower (which is usually financially unaccounted for) and offer services to a user community that is quite large and diverse.

Another sustainability challenge is related to simple Cloud-like IaaS or PaaS services, which offer (at least on the surface) very attractive price/capacity ratio and lower manpower requirements for the organisation deploying them. GRDIs probably need to co-opt this approach and aim at producing interfaces that work with dedicated data solutions as well as Cloud-based approaches. This way GRDIs can position themselves as providers of tools and services that minimise the risk of a vendor lock-in.

The third challenge, mainly related to governance of the GRDIs, is likely to emerge from the regulatory framework surrounding the GRDIs. At the moment, external developments could have unanticipated consequences. For example, the "right to be forgotten" related to the new personal data protection strategy could have a major impact on GRDIs, depending on the development of the related jurisprudence. Rather than being specific to any particular issue, the challenge is that at the moment the GRDI community does not have a collaboration model to identify regulatory issues that concern the sector, nor do they have established channels for consolidating and expressing their opinions efficiently.

4 RESEARCH DIRECTIONS PROPOSED

In order to address funding and sustainability issues, it will be necessary to create more comprehensive models for both the costs and the direct and indirect socioeconomic benefits of GRDIs. The results of these studies would also contribute to the discussions related to governance models because groups that create socioeconomic benefits based on GRDIs (and thus help justify their sustainability) should also have a voice in the governance of the services.

There should also be an effort to analyse whether or not successful models that have addressed sustainability, funding, and governance issues in the software development domain (especially in the open source sector) could be applied to data. Studying the motivations of the sponsors of these organisations could also provide alternative ways for justifying long-term investments in the funding and sustainability of GRDIs.

Finally, complementing the current approaches with large-scale pilots for innovative and interoperable solutions funded by the CIP programme (mentioned in the EC Digital Agenda as one of the tools) could be used to test and showcase some of the anticipated socioeconomic benefits of the GRDIs. These large-scale pilots from the CIP programme might also create a clear link between GRDIs and a strategic flagship project of the EU. These activities should naturally be linked with similar activities on the global scale.

5 RECOMMENDATIONS

Most of the challenges identified above rely fundamentally on building a human network that reaches most of the GRDI stakeholders and has high enough visibility and credibility that it can influence the regulatory processes surrounding the management of research data. The following are some examples of the several initiatives and organisations that could act as building blocks for this network.

- e-Infrastructure Reflection Group (e-IRG)
- EUDAT project
- International Council for Science - World Data System (ICSU WDS: http://icsu-wds.org/)
- Open access pilot initiatives supported by the EC

When this network starts to gain momentum and external recognition from the governments and funding agencies, it becomes much easier to motivate various stakeholders to join in the discussion on such issues as metrics, flexible governance structures, etc. This critical mass will also reduce the likelihood that new data management solutions will be developed in isolation for too long.

GRDI users and service providers should develop a common strategy for approaching the political decision makers. In Europe, maximising the visibility of the GRDI community in the Digital Agenda Assembly would be a natural goal that could also be used to catalyse into action those initiatives and organisations mentioned above.

6 ACKNOWLEDGEMENTS

The author would like to express his gratitude towards the participants and organisers of the 2010 GRDI2020 event in Cape Town. Several of the ideas presented in this paper have their origins in the stimulating exchange of ideas and experiences during the event.

7 REFERENCE

e-IRG White Paper 2011 Retrieved from the World Wide Web, June 12, 2013:
http://www.e-irg.eu/images/stories/e_irg_whitepaper_and_comments_2011.zip

(Article history: Available online 1 July 2013)