Virtual Herbarium - Dora Canhos

I only had time to read your report this week, and found it interesting to see how similar different fields of interest are in this discussion of sharing data openly and on-line. I would like to emphasize some points that are very common to our 15 years' experience with the *species*Link network in Brazil.

The first point I would like to mention is that sharing data openly on-line requires a cultural change. You mention that the projects indicated a desire to control the process of sharing, to know who is using the data and for what purpose as a common theme. Unless you close a system, controlling who has access to the data by requiring a username and password, it is impossible to know who is using it and for what purpose. Even with all this control, it is impossible to stop these "controlled users" from sharing the data with other people. The fact that it is digitized and on-line makes it easy to use, change, add value, etc. etc. So people that are obliged to share THEIR data feel very insecure.

Excluding "sensitive" data from "the rest" (and I will get back to "sensitive" data), and when making "the rest" available on-line, in my opinion what is required is not a closed/controlled system, but a disclaimer. If there is any miss use, or if the data is not accurate and used inadequately, whoever uses it is responsible and not the data provider. This helps people feel more secure in releasing their data.

As to sensitive data, this concept also changes, as people feel more comfortable in sharing their data. One of the studies I am carrying out in our OCSDNet project is in finding out what data is being blocked and why. Reasons vary, such as not publicizing geographic coordinates of species in red lists or of species of commercial value, or blocking data that has not been published. At the same time we have data providers that want to publicize geographic coordinates of endangered species so that there can be social control at those sites. There is no consensus, but there is freedom in following one's own convictions. We even have a case of a curator who did not know the data were blocked. Some curator in the past blocked the data for whatever reason and no one unblocked it. Therefore, in my opinion, one must acknowledge that sensitive data exist, but by default, all other data of public interest must be open.

Therefore, another concept is "data of public interest". One of the purposes to define a data model for data from biological collections (such as herbaria, museums ...) was to determine what data should be shared. A biological collection normally has much more data than that that is being shared. Perhaps an example is the physical location of where the specimen is stored - a specific cabinet and/or shelf. This information is important locally, but of no interest in a global system. So one has to separate important in house information, normally used for local management, from information/data of general interest.

Another point you mentioned was that a substantial proportion of projects do not have the internal capacity (technical expertise, resources) to provide and maintain an e-infrastructure to receive, store, organize, and disseminate data over time. This is a point I often mention when funding agencies require that each and every project must be responsible for coming up with a system that will openly share the data that the project produces. I imagine that most projects will have a spreadsheet with their data that will probably be part of the final report. However, if a project does succeed in developing and maintaining an e-infrastructure, the system will probably close down as soon as the project (funding) ends. Therefore, in my opinion, an e-infrastructure with adequate long-term funding must be in place to receive and store this data, making it possible to retrieve it when necessary.

Once again, one must always qualify this data. It may be a one-time experiment that someday, someone may want to reproduce and the paper describing it may only present an analysis of the data and not the full data. In this case, this data should be available in some repository, not necessarily integrated to other data systems, but with a unique identifier so it can be recognized and retrieved. Stable university data depositories could probably be used. And then there are data that have a time series that must be compared (locally, nationally, globally), and data that are of global interest, among others. In this case, a permanent e-infrastructure must be in place to receive this data and to make it accessible to target users and probably to other users and systems for uses that were not previously identified. This cannot, in my opinion, be an attribution of an isolated project. There must be an e-infrastructure in place or developed for the purpose and maintained over time. Remembering that "maintained" also means dynamically and continuously developed.

Against all this, your study indicated that each project "expressed a strong desire to retain control over access to project outputs" and ... "rejected third party repositories as an acceptable path to sharing". A way to overcome this resistance is to maintain the control as to what is sensitive or of public interest to each data provider.

Maintaining an e-infrastructure requires specific expertise, infrastructure, and long-time planning and funding and, in my opinion, cannot be the mandate of an isolated project. An e-infrastructure is not a mere deposit of data. There must be interaction with the community of interest. It must be part of the social network that is collecting, organizing, and using the data. It must be permanent.

All this said, a data management plan (DMP) at the project level continues to be essential. If the data is to be indexed by an existing e-infrastructure or deposited in an institutional repository it probably must use accepted standards and protocols. A DMP is also necessary to ascertain that project data needs and outputs are attended.