Chapter 17
Database Maintenance, Data Sharing Policy, Collaboration

Dario Papale, Deborah A. Agarwal, Dennis Baldocchi, Robert B. Cook, Joshua B. Fisher, and Catharine van Ingen

"*If I have seen further*," Sir Isaac Newton wrote to Robert Hooke in 1676, "it is by standing on the shoulders of giants."[1] What Newton was implying was that he was able to do more, understand more, and further advance science as a whole because he was able to build on the advancements of his predecessors. If these "giants" had

[1] The phrase was, in fact, based on that of Bernard of Chartres five centuries earlier. (d. 1130): "We are like dwarfs sitting on the shoulders of giants. We see more than they do, indeed farther . . . " (*"Nous sommes des nains juchés sur des épaules de géant. Nous voyons ainsi davantage et plus loin qu'eux, non parce que notre vue est plus aigüe ou notre taille plus haute, mais parce qu'ils nous portent en l'air et nous élèvent de toute leur hauteur gigantesque.")*. Gimpel, J., 1961. The Cathedral Builders. Grove Press, New York.

D. Papale (✉)
DIBAF, University of Tuscia, Viterbo, Italy
e-mail: darpap@unitus.it

D.A. Agarwal
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
e-mail: DAAgarwal@lbl.gov

D. Baldocchi
Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA
e-mail: baldocchi@berkeley.edu

R.B. Cook
Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: cookrb@ornl.gov

J.B. Fisher
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

C. van Ingen
Microsoft Research, San Francisco, CA, USA
e-mail: vaningen@microsoft.com

M. Aubinet et al. (eds.), *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*, Springer Atmospheric Sciences, DOI 10.1007/978-94-007-2351-1_17,

not shared their "shoulders," then Newton would have been limited in his ability to "see"; in other words, if his predecessors had not shared their work – namely their science and data – then Newton would not have been able to make as many scientific contributions as he did.

Scientific questions of today are now more global than ever before. The answers to these questions are buried within multiple disciplines and across a diverse range of scientists and institutions. The expanse and complexity of data required by researchers often exceed the means of a single scientist. Data sharing in the form of its distributed collection and analysis is increasingly common. Collective research now takes place in what may be called "collaboratories" or in "centers without walls" (Clery 2006).

Creating effective artifacts, which enable scientists to collaborate on data analyses, continues to be a significant challenge for today's science activities. It is rare that providing a file system abstraction on distributed data enables acceleration of scientific discoveries. By explicitly identifying and addressing the different requirements for data contributors, data curators, and data consumers, we can create a data management architecture which enables the creation of datasets that evolve over time with growing and changing data, data annotations, participants, and use rules.

This involves also a crucial contribution by the teams and people collecting the data, that in addition to carefully acquire and process the measurements and to be ready to share their measurements within the scientific community, need to follow general rules that help to make their data well documented and safely stored and to maximize visibility to their works and sites.

In this chapter, we provide examples of the types of functions and capabilities typically provided within the data management systems, focusing in particular on databases structures and characteristics, data practices, and data user services. Finally, the importance and advantages of collective efforts like data sharing for synthesis activities and the relative data policy options are discussed and analyzed.

17.1 Data Management

The eddy covariance (EC) technique produces a vast amount of data, from the 10 Hz measurements to the aggregated 30 min fluxes and ancillary data. In addition, differently from other centralized activities like, for example, the remote sensing data acquisition centers, the eddy covariance community is heterogeneous and it requires additional efforts to get the community working together using comparable measurements. To achieve this goal it is crucial to have a structure operating in this direction and the EC fluxes community is served by several interoperating data management facilities. Each of these facilities contributes to the overall data management, coherence, and usability of the network of sites measuring EC fluxes. The first data management layer is provided by the flux tower itself and the team that manages the tower. The raw data generated from measurements at the tower are quality checked and archived by the team responsible for the tower. The other data management centers include the regional networks and the FLUXNET global network. Each of these data management centers has a role in providing the overall flux data archiving and user services.

17.1.1 Functions

A typical data management system for EC fluxes and meteorological data provides one or more of the functions of the full data management system including robust archiving of data, generation of standardized data products, authenticated access to users, additional data products, and documentation for the data. The data management system as a whole becomes the focus of collaboration among the EC fluxes measurement scientists and a means to interact with the users of the EC flux data. As a whole, the combined data management systems for shared scientific datasets should exhibit the following properties:

- *Archive:* Carbon, energy, water, and other gas flux measurements data are exceedingly valuable and careful archiving of the raw data and generated products should be an integral part of the overall data management system.
- *Quality:* Data quality indicators to identify potential problems with the data stored as parameters of the data in the database. Identified data quality problems should also be addressed and corrected when possible. In addition to the quality checks applied by the data providers the data management center should apply additional controls to identify suspicious or erroneous data that need to be flagged and communicated to both the users and the data providers, providing an additional and independent quality control tool to the latters. The correction and processing methods implemented in the database must be always updated and advanced in agreement with the new scientific and technical findings.
- *Secure:* Most scientific data require access control and accountability (e.g., to determine *post facto* who has accessed the data) for a variety of reasons. For example, even when the policy allows anyone to access the data, requiring a registration and subsequent authentication step before access is granted allows access to be tracked (e.g., so that the impact of the data might be quantified via number of unique data accesses). The overall system must meet the collective security requirements (policy and mechanism) of the data providers and the users.
- *Scalable:* The network has grown dramatically in the last 10 years (Fig. 17.1), and it is expected to continue to grow in particular in the less covered areas of the globe. The system must be ready to answer to changes along a number of dimensions: size of dataset managed, size of meta-data managed, and number of active participants (authors, curators, publishers, and consumers). In addition, the data management system should be efficient in its behavior: stability and robustness, ability to serve the data to consumers, ability for contributors to upload new data/meta-data, and ability for curators to determine what requested

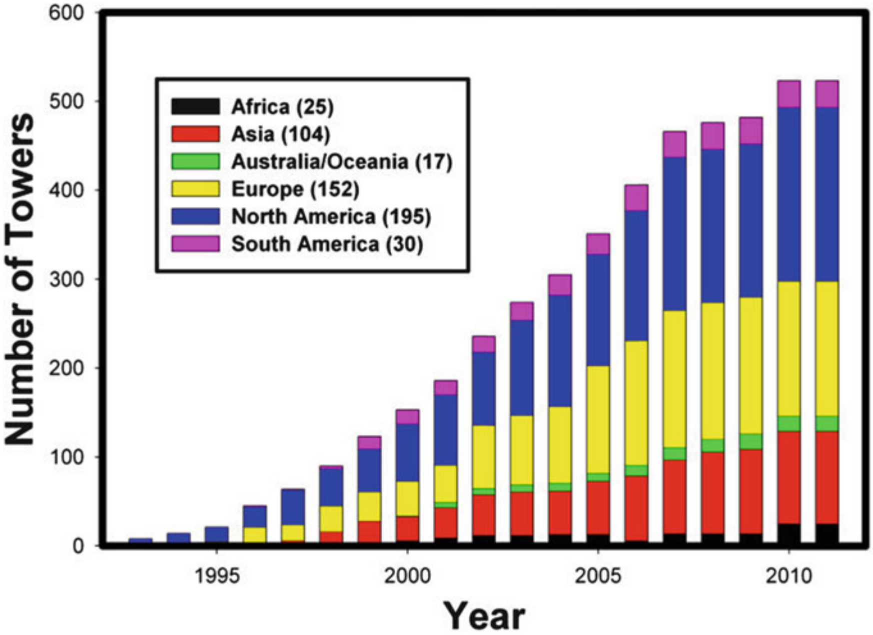

Fig. 17.1 Growth of FLUXNET by continent since 1993 (updated to March 2011). Data are from registrations to the FLUXNET ORNL database (www.fluxnet.ornl.gov, visit the website for updated information)

modifications are pending are just some examples of the services that need to be scalable and ready to answer new requests.

- *Usable:* Users must be able to easily find and obtain the data they need along with the relevant meta-data for their scientific explorations, explanations of the processes used to generate the data, and quality metrics for the data. Collectively, users need to be able to explore the data both based on keywords and on application-specific properties of the data – for example, "locate all scientific output (papers, derived datasets, etc.) directly or indirectly based on observations from SensorX in the range January 1, 1999 through June 30, 2002."
- *No special-purpose software for data access:* Ideally, the participants should not be required to learn new software packages to fully participate in the data analysis collaboration. The most attractive approach to this requirement is to ensure that users can fully participate using only a Web browser or spreadsheet program.
- *Provenance:* The data and meta-data that are held by the data management system is connected via potentially multiple set of relationships. For example, a potential consumer of a particular set of data might ask a question about the data in a particular blog, which might generate an answer that explicitly references another piece of data or meta-data. The data management system must be able to keep track of such histories and origins of data and meta-data, and such

provenance must be efficiently integrated into the rest of the data management system (such as the search capability).

- *Notifications:* The users of the system should not be expected to directly engage the data management system in order to determine what has changed since the last time they visited the system. That is, the users of the system should be able to register their interest in a variety of types of additions/modifications (e.g., data revisions/additions, meta-data revisions/additions, new users of a particular class) and be able to receive notifications via their choice of a variety of mechanisms (e.g., e-mail or SMS). In essence, the system should enable subscription to and *push* of information to the users of the system.
- *Additional processing:* The system should provide a number of derived variables calculated centrally using the raw data (in this case raw data are the calculated fluxes) ensuring the same methodology and calculation scheme. These derived products could include quality assurance and quality control flags (Chap. 4), gap-filled datasets (Chap. 6), and calculated additional variables like GPP and R_{eco} from partitioning (Chap. 9) or data-derived products such as water and radiation use efficiency, potential evapotranspiration, surface conductance, etc. In addition, it could be important to implement links with other data sources relevant for the users like meteorological networks, remote sensing products, and climate models results.
- *Track usage:* Papers and other scientific results derived from the data should be traceable, for example, creating a public list of products where each of the datasets included in the database are involved or cited such that the impact of a site’s data for a particular time period can be identified by the data owner and funding agencies.

17.1.2 Flux Tower Repositories

The long-term value and quality of the EC data depends first and foremost on the quality of the measurements and care of the data provided by the individual flux tower teams. The flux tower teams have a critical role in the overall data management. Archiving of the raw measurement data by the tower team helps to ensure that the data are not lost and available for future recalculations using new methods or corrections. The use of standardized names and units for variables and keeping the site identifier and site name consistent throughout the life of the measurement site significantly reduces confusion and processing mistakes in particular during the reprocessing of older data. This applies to all of the data from the site not just the EC and meteorological data.

It is important to organize and archive the information about the data gathered together with meta-data such as measurement setup, instruments makes, models and serial numbers, instruments locations in the tower (height) or in the footprint (location, depth in the soil), calibration dates and methods, maintenance and disturbance information for each sensor, methodologies used to calculate fluxes and correct the data. The same is valid for all the biological disturbances and management information about the site that needs to be stored together with the information about the methodologies and people that collected the measurements and information. These data are critical to the utility of the EC measurements. Ideally data should be archived in at least two locations with one being off-site (in many cases the regional network will serve as the second off-site archive of the data). Regular testing of the archiving system increases the likelihood that data can be recovered in the case of an emergency.

17.1.3 Regional Repositories

The first examples of regional databases dedicated to eddy covariance measurements were proposed at the end of the 1990s in the context of the two major regional networks AmeriFlux and EuroFlux-CarboEurope. The two databases were initially relatively simple and without many functions. But they have evolved during the last 10 years, adding new functionality and collaboration with the aim to coordinate and standardize the services.

Coordinated and interconnected regional databases are more efficient than a single global network. First, political and cultural positions are better managed by local coordinators, who are more connected and have deeper understanding of local conditions than outside coordinators. Second, the lack of global-scale funding agencies currently interested in fully supporting a global database system for EC and related data, is somewhat offset by the interest of regional and continental funding availability to support regional databases (e.g., DOE for USA or EU for Europe). Therefore, harmonization of regional databases is needed to maintain and improve the interoperability and inter-database standardization. FLUXNET is a global initiative that works with regional networks and helps to create a network of networks that share processing options, standards, and policies to enable global-scale studies (Fig. 17.2).

There are many different regional databases that coordinate networks ranging from a few tens of sites to more than 100 stations. The main EC databases are AmeriFlux (http://public.ornl.gov/ameriflux/index.html), CarboEurope and CarboAfrica (http://www.europe-fluxdata.eu), Fluxnet-Canada (http://fluxnet.ccrp.ec.gc.ca/e_about.htm), and Asiaflux (https://db.cger.nies.go.jp/asiafluxdb/).

17.1.3.1 One Example: The European Eddy Covariance Flux Database System

The European network of EC sites started in 1996 with 16 forest sites in the EuroFlux-EU project and has grown to more than 140 sites since then, thanks to several other EU-funded projects like TCOS-Siberia, GreenGrass, CarboMont, CarboEuroflux, CarboEurope-IP, IMECC, CarboExtreme, and GHG-Europe. Paral-

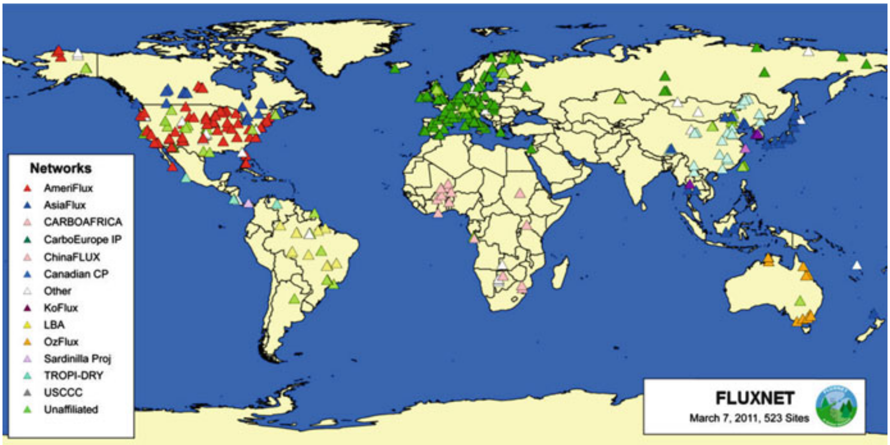

Fig. 17.2 The regional networks and FLUXNET. Continuously updated version is available at the FLUXNET ORNL webpage (www.fluxnet.ornl.gov)

lel to the network, the database system has been developed to host, quality control, standardize, and distribute the data acquired at the different sites. The database, based on SQL and .NET, is currently located in Viterbo (Italy) at the University of Tuscia and the server has a double backup system and is connected to a set of workstations for data processing. To ensure standardization and promote and help data exchange between the different projects, the different projects share the same database structure and processing so that data products are presented in the same way (same format, same quality flags meaning, same variables names and units, same processing scheme); sites involved in more than one project do not need to upload and curate the data in different interfaces (see www.europe-fluxdata.eu). At the same time, the project part of the general database maintains specific interfaces that give visibility, simplify the intra-project coordination and can host additional project-specific datasets.

In 2005, the European Eddy Covariance fluxes database introduced a number of new tools that improved the suite of services offered to data owners and data users and changed the philosophy of the database, moving for a data repository to a real data management system. A standard data processing was introduced, including the implementation of the QAQC described in Papale et al. (2006), the u_* filtering in Reichstein et al. 2005, two gap-filling alternatives (MDS – Reichstein et al. 2005 and ANN – Papale and Valentini 2003), flux partitioning (Reichstein et al. 2005), and the dataset was provided in a standard format to simplify the use and automate the reading. At the time when this chapter was prepared, the suite of processing options implemented in the database was under review to include new QAQC tools and a new partitioning method (Lasslop et al. 2010). The data processing will likely continue to evolve and improve over time and it is one of the most important characteristics of a database system to be always ready to implement new processing schemes proposed in the literature also because these are often difficult to be applied individually by each single site manager.

Two other main services introduced were the versioning and the PIs Information system (PIs are the Principal Investigators, the site responsible and data owners). The versioning system was introduced to track new versions or new processing schemes applied to already existing and used datasets. In fact, development in the calculation and processing methods or errors discovered late in the data are the main causes of new versions of datasets that have been already downloaded and used. In these cases, it is important to inform all the users that a new version is available and to track all the changes that have occurred to the data from the first version released. Only with this information is it possible to fully understand the differences between successive studies based on the same sites but using different datasets. More details about the versioning are reported later in this chapter.

The PIs Information System is a section of the database dedicated to the data providers. Here the PIs can upload the data, track the status of the data processing, define the data access and data use policies to be applied to the data, and see the list of the users that downloaded their sites measurements. In addition, an e-mail service automatically informs the PIs about who requested access to their data and analysis planned. Providing these information, the PIs are on one side more comfortable in sharing the data since they know who accessed the data and, if interested in collaborating with the study, can directly contact the data users to propose joint activities; on the other hand, they can have a quantification of how much their sites are interesting and useful for the scientific community and for this reason important to maintain active and to be funded.

17.1.4 The FLUXNET Initiative and Database

FLUXNET, which is a network of networks, is a collaboration between the regional networks and independent measurement sites which results in a global EC fluxes measurement network. This global network is brought together to enable global-scale synthesis activities. The main role of FLUXNET is to establish contacts between regional networks and EC people around the world and to maintain an updated inventory identifying which sites are active, their period of data taking, and what is measured at a site to facilitate exchanges and coordination between sites. FLUXNET also promotes synthesis activities through meetings and workshops where standardized datasets and derived products are used in scientific analyses at global scale. FLUXNET receives new data when a new dataset is gathered in preparation for a global synthesis effort (not continuously). The aim of each dataset collection effort is to produce the new FLUXNET dataset collecting preprocessed data from all the existing sites, through the regional databases or directly from independent sites. Typically, a FLUXNET Synthesis workshop is held in conjunction with the data collection effort to enable users to discuss the dataset's scientific potential, propose new cross-site synthesis analyses, and start these analyses. The first FLUXNET Synthesis workshop was held at the Marconi Conference Center in 2000. In preparation for the Marconi workshop, a dataset was gathered that contained 97 site years of data from 40 sites located primarily in the Americas and Europe. The dataset was quality checked and gap filled using a standardized methodology. The synthesis analysis efforts using this Marconi dataset resulted in 11 synthesis papers published in a special issue of Agricultural Forest and Meteorology in December 2002.

In the case of the Marconi dataset, all the synthesis teams knew the measurement site scientists personally, so communication and trust among the group members was relatively easy to establish. The data download and communication functions were handled manually (ftp, e-mail, and phone) for the Marconi dataset. The resulting dataset contained only the flux-met data from each site. The synthesis teams obtained any needed ancillary data for the sites directly from the measurement teams.

The second FLUXNET synthesis workshop was held in La Thuile, Italy, in February, 2007. As a result of the workshop and continuing efforts since the workshop, the La Thuile dataset was released to synthesis teams in September 2007. The delay in releasing the data was due to both the time needed to complete the gathering and processing of the data and the time needed to develop a portal to support sharing of the dataset. The released La Thuile dataset contains over 960 site years of data from more than 180 scientists working at over 250 measurement sites around the world. Four years later, there were more than 110 synthesis teams writing papers using these data (see http://www.fluxdata.org for the current list of published papers).

As the body of EC flux measurements increases, their value to the broader scientific community increases as well. For example, the FLUXNET LaThuile dataset was an order of magnitude bigger than any carbon flux measurement dataset that had been available before and is enabling cross-site, regional, ecosystem, and global-scale analyses. The LaThuile dataset was also more valuable than previous FLUXNET datasets because advanced standard processing methods were applied and the ancillary data were included as part of the dataset.

The amount of data and users involved in the LaThuile-2007 FLUXNET synthesis activity, together with the progress in the computing and IT structures available, led also to a radical change of the database structure with respect to the Marconi initiative. A database structure accessible through a web interface was developed, and it is currently the reference access point for the FLUXNET synthesis studies (www.fluxdata.org). This database is maintained updated and synchronized with the FLUXNET meta-database and the regional databases. A new data collection and processing is ongoing with the aim to release a new FLUXNET dataset in 2012. This new dataset will be significantly larger than the LaThuile dataset and with additional processing and uncertainty estimation tool included.

17.2 Data Practices

17.2.1 Contributing Data and Reporting Protocols

Data are generally submitted by the flux tower team to the regional databases, which are responsible for the next level of quality control and processing and also for the data transmission to the FLUXNET database for synthesis activities. This two-step process allows more precise control of the data and policies but can introduce differences between the regional networks and FLUXNET. For this reason coordination is important.

Protocols for data transmission to the regional networks are generally provided by the different databases and are often different. This does not affect the ability to share data between networks because it is one of the roles of the regional databases to import the data in the original template and export the data after quality control and quality assurance in a standard exchange format. From a centralized standpoint, one common data reporting template and protocol for all the sites is desirable. However, it is difficult to agree on protocols that can be applied in completely different ecosystems and climatic regions. Despite these difficulties, several large coordination efforts are ongoing. Some results have already been achieved; the Biological, Ancillary, Disturbances, Management (BADM) template (see www.fluxdata.org), originally developed in AmeriFlux, has been adopted by other networks (Europe, Africa). A period of suggestion of modifications and improvements to the original AmeriFlux template allowed inclusion of data types and information specific to other regions and originally missing from the template. This template is now used across the networks to report all the measurements and information that have low time resolution (daily to annual) but that are fundamental for a correct data interpretation. The BADM template is becoming an international standard in the FLUXNET community and is available as an Excel workbook to enable tower teams to fill in the information off-line and then submit the completed template. However, the data collected by the BADM template cover a wide range of data types and complexities. Some examples include plant species percents and site disturbance information which each require the entry of a text string that comes from a control vocabulary (species or disturbance type) and a value (percent or year) along with other relevant information. Although the template provides an easier mechanism for data input, the Excel format does not enable the input to be checked as it is entered. If the users do not follow the instructions for data input carefully, a great deal of manual work is required at the time of data ingest to correct the template.

Web-based database form interfaces can also provide services to directly update or submit ancillary data. This method has the advantage of fast submission and registration of the data. It also enables direct tracking of the origin of data and the ability to apply simple rules regarding the values reported to be sure that they can be imported correctly. An example is the submission of a text string where a value was expected: if entering for example the variable “disturbance_year” (the year when a disturbance happened) is reported as [two thousand], or [2000–2001], or [January 2001], or [03-01-2001], a web interface can immediately warn the submitter that it needs to be corrected. The correct value to report would be [2001] and other information about uncertainty in the year or exact day is entered as related variables or as a comment or annotation. A web interface for data entry improves the likelihood that the data can be imported correctly and will not require manual corrections.

Despite this improved control over data submission, data submitted using web-form interfaces are still affected by reporting errors. To enable these errors to be detected, a temporary table can be used to store newly submitted information until a curator for the site has verified it. The curators are experts who check the submission and ask the PI for clarification if the values are suspicious. A common example is an error in the units that makes the value "possible" but may be not "probable." Once a value has been confirmed, the curator accepts it and it is moved into the database.

17.2.2 Common Naming/Units/Reporting/Versioning

FLUXNET is a collaboration where data products are made by individual networks as well as cross-network synthesis groups. Data products include different processing methods of fluxes and meteorological time series field data, derived variables like light use efficiency, water use efficiency, or drought indices, ancillary data such as site classification, disturbance, and management history, or biomass characteristics and remotely sensing data. Today, these data products are usually produced as a collection of files. As the community grows and cross-site and cross-network synthesis studies have become more important, the need for standardization across the data has also grown.

17.2.2.1 Enabling Cross-site Analysis: Site Identifier, Variables, and Units

A prerequisite of cross-site analysis of the data is an ability to compare the data from different sites and regions. The first step in enabling analysis of the data is unique site identifiers: a site's identifier does not change unless the site changes significantly the location (i.e., different footprint, see Chap. 8) and different sites have different identifiers. This is typically accomplished through assigning codes for the identifiers for each of the sites, associated with the geographic coordinates. The use of identifiers allows the human readable name of the site to change without affecting the unique identifier (although it is not recommended that the site name change unless necessary).

A cross-site analysis also needs the data itself to be comparable across the sites. This would require in theory exactly the same instruments, setup, and methodologies to acquire and process the data implemented at all the sites. Although there are efforts ongoing in this direction with the organization of highly standardized measurements networks (see, e.g., ICOS, www.icos-infrastructure.eu), the knowledge of the sources of uncertainties related to the EC measurements processing (Chap. 7) allows the use of not fully standardized data for synthesis activities. This, however, requires the following of some basic rules such as reporting of common variables across the sites with associated meta-data to describe system and processing but also that the data be reported by all the sites with the same variable names and units. Although ideally the original data should be reported using standardized names and units that are agreed across the regional networks and FLUXNET, this is sometimes difficult due to the heterogeneity related to relevant measurements in different environments, different levels of detail, and different units adopted in different countries. A compromise is that the regional databases should harmonize variable names and units inside the network and implement conversion tools to meet the FLUXNET standards. The updated list of the standardized variable names and units are available in the regional databases and FLUXNET web pages and must be consulted during the dataset preparation before submission.

17.2.2.2 Data Releases

One of the aims for the near future is to update continuously (or frequently) the FLUXNET datasets with new and updated data. In this scenario, the flux-met data will arrive in large yearly batches and the ancillary data would be continuously updated. With a dataset in continuous evolution, the data users need to be able to indicate a version of the data that they used in performing their analysis. There are a wide number of models we can draw from in designing a data release strategy. Agencies such as USGS process their data once from collection to quality checking and release. Released data do not change. NASA uses a collection abstraction for their releases. They continually add new data to a collection as they are received and processed. NASA reprocesses the entire dataset if there are revised processing calibrations or algorithms. Reprocessed data are released as a new collection. The fluxdata.org site has adopted a strategy similar to the NASA collections and releases a frozen version of the data before each major update to the flux-met data (particularly when all the data has been reprocessed, which is currently the case). A data release spans sites and contains files with the same data types in the same format and units. For example, a data release might consist of flux-met data files with half-hourly and daily aggregations for a set of site years and quality flags indicating the results of the centralized quality controls.

All files need to be versioned, which is effectively a serial number for the file and increments monotonically at any change to the data or file format. If the same file is used to create more than one data release, the version is unchanged. For example, data files for inactive sites will remain unchanged across different successive data releases unless the data are reprocessed using a new version of the data processing.

Data releases are either classified as *frozen* or *latest.*

- A frozen data release does not change. Users of a frozen data release have the guarantee that the same analysis will give the same answer. When possible, scientists should use a frozen data release for publication, particularly for synthesis studies. Frozen data releases are made by halting all changes to an existing latest release.
- A latest data release may change at any time. New files may be added, new file versions may replace existing files, files may be removed, and processing algorithms may change.

Networks create data releases at any time. These data releases may be published to only network members, for a synthesis set such as LaThuile, or to other collaborations. A release may not contain all sites in the network either due to quality bars or due to the targeted science.

When freezing a cross-network synthesis set (e.g., a FLUXNET dataset like for the LaThuile data collection – www.fluxdata.org – or the Marconi synthesis dataset – Falge et al. 2005), all of the contributing network and unaffiliated data releases should also be frozen. In other words, when a synthesis data release x is frozen, it should be based on frozen data releases from each network as well as the unaffiliated holdings of the original data used to create that synthesis data product. This may require the networks or unaffiliated sites to freeze an existing latest data release when they would not normally do so but the advantage is that the data release is then traceable throughout the system (the network knows exactly which data went into the release as does the site).

Best practices include:

- Each network makes available at least one frozen data release and the latest data release for any data products published by that network. For example, AmeriFlux would publish both the original data used to create the cross-network synthesis version x (the last time data were frozen by that network) as well as a latest release.
- Descriptive summary documentation should be provided to explain the differences between data releases, though it is not necessary to detail the change in every single point. Knowing what data have not changed is often as interesting as knowing what data have changed. Differences in annual values or variability may also be helpful.
- Data releases should include fluxes, meteorological data, and ancillary data. The combination enables richer science.

17.2.2.3 File Naming

The name of the file downloaded by the user, exchanged between regional networks or submitted for FLUXNET synthesis datasets preparation could include important information about the content. In this paragraph one possibility will be discussed and presented to better explain and illustrate the possibilities of standard file naming. The goal is to create file names that can easily be understood by humans, clearly identify the version and source of the data, identify the type of data and processing used, and clearly identify the set this data was published into. All this information could and should also be included in the meta-data and data information files, but having a summary in the file name would help to have an overview of the data characteristics.

File names could include, in addition to a code to identify the site and a number to identify the year, information about:

- Data type, like fluxes, meteorological, biological, and ancillary data, and remote sensing cutouts.
- Data version, with a monotonically increasing number that will identify successive versions of the same data (e.g., due to changes in the processing or errors in the previous data). This number changes only when file content is different from the previous version.
- Processing applied, if the data are as submitted to the database or have been processed with additional QAQC, gap filling, partitioning.
- Time resolution of the measurements, in particular for aggregated data from daily to yearly.
- Regional networks or synthesis activity that produced the data, useful for users that are working in the context of specific continental or project activities.

For example, a file name could be structured as

AAAA_LL_PPP_T_CC-SSS_YYYY_vvv.<extension>

where AAAA identifies the network (e.g., CEUR for CarboEurope, AMFL for AmeriFlux, CAFL for Canada); LL, the processing level (e.g., L1 for raw data, L2 processed data, L3 quality controlled, L4 gap filled . . .); PPP, the data type (e.g., FLX for fluxes, MET for meteo, ANC for ancillary, SLR for soil respiration . . .); T, the time resolution (e.g., H for half-hourly, D for daily, W for weekly . . .); CC-SSS, the site code with CC indicating the country, and SSS, the site code; YYYY, the year; and vvv, the version number.

17.2.3 Ancillary Data Collection

In 2007, ancillary data reporting protocols were developed in the context of a FLUXNET synthesis activity (BADM template, see www.fluxdata.org), but the formatting diversity and heterogeneity of these data still makes consistency difficult. Reporting of data ranges, approximate values, and qualifiers in lieu of simple numeric values is a common practice, so all ancillary data should be stored as text fields to preserve accuracy information and support non-numeric entries while conversion to numbers is possible during the data quality evaluation and control.

The submission date, user, and method should also be captured when ancillary data are submitted and this provenance information stored along with the values. In addition, all past values and their provenance information are kept so that the history of the values provided for a variable can be reconstructed. This enables ancillary data views that correspond to the database state at any particular point in time to be reconstructed.

Fluxdata.org provides an example of web interfaces and protocols for reporting values. The challenge is to capture ancillary data in the portal and check the submission format of the data when they are collected by the tower and analysis teams.

17.3 Data User Services

Flux data have the potential to benefit a wide range of scientific analyses. This broad usage has the potential to significantly enhance the impact and value of the data, but it can only be achieved if the data are available and usable. The data standardization, versioning, quality assessment, and curation components mentioned earlier in this chapter are precursors to data reusability. In this section, we discuss some of the other services and capabilities and data sharing rules and motivations.

17.3.1 Data Products: The Example of fluxdata.org

Large-scale synthesis studies are becoming more common. These synthesis studies are often conducted by science teams that are geographically distributed and on datasets that are global in scale. A broad array of collaboration and data analytics tools are now available to support these science teams. However, building tools that scientists actually use requires a lot of work. In this section we will refer to an example of collaboration between the regional networks and FLUXNET communities that led to the development of the fluxdata.org portal to support data analysis by users. The fluxdata.org infrastructure provides advanced data organization, mining, and analysis features through utilization of a database to organize the data. Cross-site data reports and On-Line Analytical Processing (OLAP) data cubes enable browsing of the data. It is an example of the types of user services and products that can be provided to the user community. We discuss the types of functions and users that an EC flux portal should support.

17.3.1.1 Users and Use Cases

A first important step in the construction of a database infrastructure is the definition of the users and related needs. In fluxdata.org, there are four primary types of users and associated use cases for a carbon flux data portal. These are:

- Analysis scientists (data users) – site selection, dataset information data download, analysis support, and paper writing support
- Measurement site scientists (data providers) – information about proposed and published papers using their data, data download, and data submission/update
- Regional flux networks (data curators) – data correction, checking, and update, coordination of the regional contributions
- The public – proposed and published papers information, dataset information, and funding information

Although it is tempting to think of each of the above user groups as distinct, this is not the case. Multiple groups share many of the use cases. Measurement site scientists are typically involved in synthesis activities and regional networks.

The primary use cases are identified below:

- Synthesis site selection – evaluate criteria that will determine which sites are suited to an analysis. Typically, most of the site selection process is done using high-level aggregated data about the sites like annual sums or averages or percentages of high-quality data and ancillary and meta-information about the site.
- Dataset information – ability to quickly answer simple questions about the dataset such as which sites are included, which years of what data, where are the sites located, who is the measurement scientist in charge of a site, and what are the measurement system characteristics.
- Data download – ability to browse and download fluxes, meteorological and ancillary data for sites. Provide different levels of access to data according with the specific data access policies.
- Data update – submit updates to ancillary data and new ancillary data and track their provenance.
- Paper writing support – enable communication with measurement site scientists and gathering of citations and acknowledgments for data.
- Proposed and published paper information – access to proposed and published papers, paper progress, and paper site year usage information.
- Data curation – inform curator of submitted changes and provide an opportunity for a person familiar with the site, the *curator,* to sanity check data submissions.

Fluxes and meteorological data submission is not listed above. These data, which are the core of the database, are collected by the regional databases and transferred to fluxdata.org periodically. This organization gives the possibility to maintain a direct link with and between the regional networks. The scientists responsible for the measurements work with the regional database that imports, check and process the data and help also the fluxdata.org, giving distributed responsibilities for data gathering and processing.

In the next sections, we will discuss some of the elements needed in a user portal to support analysis usage of the carbon flux data.

17.3.1.2 The Public Access Area

The public area of the portal is accessed by all users without any restrictions or identification. The public area contains all information about the dataset and collaboration that can be made openly available. This public information is designed to be accessed by all users so the content is not replicated in other areas of the portal. The aim of this section is to present the activities to potential users, potential data contributors, and to the agencies. The public area of the portal typically contains:

- Characteristics and locations of the measurement sites with information about the science teams running the measurement sites and funding acknowledgments of the measurement sites. This information is ideally presented using interactive maps and reports. Examples include an interactive mashup of tower locations, and reports containing the average annual values of any site that can be made publicly available.
- Lists of the variables measured at sites including the explanations and availability/years of those variables along with explanations of the derived variables, gap-filling techniques, aggregation method, units, and quality markers.
- Consistent data versioning information that enables users to identify specific versions of the data for repeatability and traceability.
- Measurement site pages each listing all public information about the site and data from the site including pictures of the site if available.
- A news feed providing regular updates, announcements of changes to the dataset, and information about new functionality.
- The analysis teams' membership, papers, progress, and lists of the sites involved in each analysis, if available.
- Data fair use and publication guidelines.
- Instructions about how to participate in the activities by both sharing new data and using data for scientific analysis.

The public access area helps users get oriented. It also allows potential new users to evaluate the expected utility of the dataset before requesting access to use the data.

17.3.1.3 The Authorized User Support Area

Access to the rest of the portal should be controlled through authentication of users and by tracking user access activity. This enables the registration of data download activities, that is, important information for the sites responsible to know the number and activity of users interested in their sites and to use this information to support the existence of the site or modify the data sharing policy. In addition, having the list of users that downloaded each specific dataset gives the possibility to contact the users in case of new data versions or possible errors discovered after download. As the amount of data available increases automated data analysis and synthesis support infrastructure is helpful to users. Usually analysis teams will be less familiar with the data than the measurement site teams. Providing users with enhanced data products such as a quality-controlled and gap-filled dataset, and calculations of derived variables such as gross and net production are highly valuable. A critical element of providing this information is the accompanying methodology explanations so that the user can correctly interpret the values.

Another critical element needed by data users is the ancillary information about the site that enables interpretation and use of the data for a broad set of analyses. This information is collected using the BADM protocols mentioned earlier. Since these data tend to be collected by a wide array of individuals, they are more difficult to bring together and methods for centrally collecting, storing, updating, and presenting the ancillary site data can be an important function of the portal.

Functions ideally available in this area of a portal include:

- Download of flux-met data in standardized formats and with QA/QC data identifiers.
- Browsing and download of compilations of ancillary data for a site and across sites.
- Access for analysis teams to update the status of their analysis (allows measurement site scientists to track progress), update the list of measurement sites used in the analysis, exchange e-mail with measurement site scientists and inform them when a paper is published.
- Access reports containing annual aggregates of site flux, meteorological, and ancillary data as well as cross-site compilations of that same data and data quality indicators. Such compilations of site information enable analysis teams to narrow down their site selections without having to download and analyze the data itself.
- Notify data contributors of data usage and enable communication.
- Quick-look tools to visualize the data (e.g., simple plots) to better preselect the variables and sites of interest.

There are two typical options for notification of data usage to data contributors. The first is for an automated e-mail to be sent to the contributor of the data each time a user downloads data. A second option is to provide the analysis teams with mechanisms to identify when they are using data from a site and then provide displays which allow measurement scientists to see what analyses have indicated usage of their data. Even if the automated e-mail on download is provided, a means of indicating sites in use by an analysis can enable ongoing communication and tracking of the contributing data to a resulting publication. The scale of the datasets and the need to enable the building of trust between participants mean it was no longer possible to rely on informal and manual mechanisms to manage the communication between analysis teams and the measurement site teams.

17.3.1.4 Measurement Site Scientist Support Functions

Although typically a data portal is created to serve data users, there are many valuable capabilities it can also provide to the data contributors. Site scientists should have access to the data from their site through the portal so they can check the data processing applied in the database and the version of the data they are actually sharing. Functions available in this area of the portal should also include the following:

- Download the flux-met data for the site.
- Display all ancillary data collected for the site.
- Submit new ancillary data and update existing ancillary data.
- Search which synthesis activities have specified they are using the site's data.
- List and contact the users that have downloaded the site's data.
- Surface data releases and accompanying documentation.
- List the papers published that use their site's data.
- Specify specific papers related to the sites that should be consulted and cited if relevant.
- Specify specific acknowledgments for their sites that could be added in the publications.

17.4 Data Sharing and Policy of Uses

17.4.1 Data Sharing Motivation

The governance of shared data is an issue with practical implications. Limitations to sharing can block advancements. For example, if every part of a car is patented by a different person, then it becomes decreasingly likely that the car would ever be fully built or functional – for example, everyone might come together except for the person who owns the patents to the wheels. In fact, this "anticommons" has been a demonstrated problem in the field of biotechnology (English and Schweik 2007; Heller 1998; Heller and Eisenberg 1998). Problems may occur when cooperation or collective action ceases before a product achieves its full potential. For shared data, similar problems may be foreseeable in that a global scale scientific question that relies on multiple data sources to answer may not be answerable if data sharing is problematic.

Data sharing is not easy. An analog comes from sharing physical natural resources. Garret Hardin's "Tragedy of the Commons" (Hardin 1968) is simple to understand because at first glance it makes sense: people consume a limited resource until it is depleted because if they miss out, then someone else will not. However, there are numerous cases of shared resources governed sustainably. How is it that these shared resources do not go the way of the tragedy? Elinor Ostrom, who won the Nobel Prize in Economics in 2009, has dedicated her research career to this question. She found that certain features in a shared resource commons are consistent throughout all the sustainable cases she examined; when those features (or, as she calls them, *design principles*) are not present, then the commons is less likely to succeed sustainably (Ostrom 1990). There are eight design principles: (1) clearly defined boundaries, (2) appropriation rules related to local conditions, (3) collective-choice arrangements, (4) monitoring, (5) graduated sanctions, (6) conflict-resolution mechanisms, (7) minimal recognition of rights to organize, and (8) nested enterprises.

Nonetheless, these principles apply to natural resources that are exhaustible (e.g., can be deleted or eliminated) – data are an intellectual property resource that is not necessarily exhaustible. However, intellectual property can be misused too. Data have been considered in the literature as intellectual property, with attention to public access when data are in the public domain (i.e., publicly funded) (Drazen 2002; Hess and Ostrom 2003; Hughes 1988; Litman 1990; May 2000; Posey et al. 1995; Rai 1999; Reichman and Samuelson 1997). Data may be hoarded by data producers or taken without permission by data users (first-use rights may be compromised), and poor analyses can taint a perfectly good dataset. The finite aspect of data may not be the data themselves, but the publications and/or acknowledgments. Recent research that combined common property theory with intellectual property theory focusing in part on FLUXNET data sharing showed that cases of data sharing without conflict or irresolvable dispute incorporated more of Ostrom's design principles than did those characterized by conflict (Fisher and Fortmann 2010).

Scientifically, generally the more data, the merrier: statistical power increases and spatial representativeness can increase (though not always, and actually may bias some regions more than others) with more data. The variety and distribution of biome types, age classes, disturbance regimes, climatic controls, atmospheric coupling, among all the other myriad of ecosystem complexity components, increase with more data sharing. Perhaps more importantly, the number of eyes on the analysis increases with data sharing, which lends different perspectives, theoretical backgrounds, biases, cultural understandings, and ideas on how the world functions, or at least how each ecosystem functions. If our grand objective is to understand how the world works, then we cannot operate in isolation: we must work with the world.

Data can be made available for broader use through individual tower web sites, regional network sites, and the FLUXNET fluxdata.org site. Clear rules governing the use of the carbon-flux datasets should be laid out in a data fair use policy readily accessible to the users. Each team that wants to use the data must abide by the data fair use policy. The data fair use policy defines proper usage, coauthorship, citation, and acknowledgment behaviors. In particular, it defines required actions to be taken by the data users before publication of any results.

17.4.2 Data Policy of Use

In the business world, the central aim is to maximize monetary profit. In the academic/scientific world, the currency is publications. Publications advance careers, establish priority, and generally aim to impact scholarly communication and society (Suber 2007). Understanding the rules governing acknowledgment, citation, and authorship is of the upmost importance in academia and science.

The fundamental intellectual property rights question for scientific data is: Who has the rights to the data produced (May 2000)? For example, a *professor* who is supported by a *federal grant*, employed by and uses resources from a *university*, hires a *lab technician*, works with a post doc funded from a *private organization*, and publishes in a scientific *journal* might be required to share or give up a number of rights to the work produced. These same types of agents and agencies may surface again when a *data user* uses the generated work for further analyses. These agents and agencies grow exponentially when *international collaborators* become involved. Generally, these agents and agencies support data ownership in the public domain, though they are subject to different levels of intellectual property laws in addition to Federal statutes, contractual rights and duties, and limits within state-funded programs.

The rules that govern data ownership operate at three scales: macro, meso, and micro (Fisher and Fortmann 2010). At the macro-scale in the USA, for example, data can be copyrighted,[2] although some researchers are interested in copyrighting their work so that they can distribute their work freely (termed a "copyleft") (Heffan 1997). Meso-scale governance of data sharing and ownership occurs among institutions like universities, national academies, and organizations such as FLUXNET (Fienberg et al. 1985; Fisher and Fortmann 2010). Micro-scale rules operate at the personal level from individual understandings (or conflicts) to unwritten relationships and norms (Rai 1999).

It could be useful also to distinguish between data access policy and data use policy. They are clearly connected but not overlapping. A user could be interested to access the data without the intention to use them in publications or presentation. Examples could be just personal use or to verify published results. Giving open and direct access to the data increases transparency and visibility for the data providers and is also in line with recommendation of the Global Climate Observing System (GCOS) and the Group on Earth Observation's (GEO) guidelines and data principles.

Data use policies are instead the list of rules, steps, and requests that regulate the use of data, including the way to recognize and acknowledge the work of the data providers and database managers and could include requests of citations, acknowledgment, or coauthorship.

[2] For example, the "Digital Millennium Copyright Act" (H.R. 2281), which updated the Copyright Act (Title 17 of the US Code) to include digital data; the "Public Domain Enhancement Act" (H.R. 2601); the "Public Access to Science Act" (H.R. 2613); the "Consumer Access to Information Act of 2004" (H.R. 3872); and the "Digital Media Consumers Rights Act of 2005" (H.R. 1201).

The critical question then is "*When should you give (or be given) coauthorship versus a citation versus an acknowledgment?*" This is constantly thought about and debated for good reason: there is no consistent answer. Linking to the fact that ownership alone (see above) is as complicated as a food web, it may now be understandable why assigning the proper credit is also as sticky as a spider's web. Expectations, norms, and practices vary widely across disciplines, institutions, and countries. These inconsistencies are not necessarily without good reason. For example, in some scientific communities it might be expected that simply sharing data is insufficient for coauthorship, but some sort of "significant intellectual contribution" is needed for coauthorship. Whereas in other communities or regions, for example, in a developing country, where data collection and publication can be a much more difficult and lengthy process logistically and due to language and other barriers, simply sharing data warrants coauthorship. The concern by the latter case is that a Westerner can easily take data from a developing country, and publish the results much more quickly – and first. If that were to occur, then the scientist in the developing country would be less incentivized to share data.

Even within a country such as the USA, the rules are unclear. For example, the International Committee of Medical Journal Editors recommends that authors meet three conditions: (1) substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data; (2) drafting the article or revising it critically for important intellectual content; and (3) final approval of the version to be published. However, even within this explicit definition is undefined terminology. What constitutes "substantial," "critically," or "important"? In the Merriam-Webster Dictionary, "author" is defined as "one that originates or creates," or "initiates," or "brings into existence." While the person who shared the data may not have been the one to initiate or bring the paper into existence, the paper's existence would be changed without the data.

As scientists perhaps we seek that unifying first-principles, physics-based model to robustly and universally define authorship versus acknowledgment. Although the search for such might well be alchemic, many journals like *Nature* and *PNAS* ask authors to clearly identify their contribution to the paper in designing the research, performing the research, and contributing to analysis and writing. Designing authorship rules could follow the design principles of Elinor Ostrom (1990). We need clearly defined boundaries between who goes where, but these boundaries need to be sensitive to local conditions constructed collectively, rather than as one group imposing them on all others. Or, we could simply end the debate and hand down a set of hard-and-fast rules that go along the lines of: (1) coauthorship if "significant intellectual input" as defined by the primary author, (2) citation where the data/idea were first mentioned/published, (3) acknowledgment – make as big as possible without getting ridiculous. However, there would be people unhappy with these rules and situations where these rules poorly dictate what to do. Nonetheless, we could use this as a starting point – a lump of clay now dropped on the table, waiting to be sculpted by the community that would use and appreciate it. Working at a global level like in the FLUXNET community means that different positions, views, cultural behaviors, attitudes, and barriers need to be considered, analyzed, and possibly synthesized in a common agreed data use policy that would act as a first step to build a more connected and integrated global flux community. When the first examples of eddy covariance data synthesis activities were published more than 10 years ago, coauthorship was also used as a tool to get the people involved and ready to share their data that at that time were new and rare. Nowadays the coauthorship meaning is changing in the direction of a more direct participation in the scientific message preparation and discussion and directions to find a compromise with the data providers' acknowledgment expectations.

Examples of data policy requirements may require that the analysis team must contact the scientists at the measurement sites used in the analysis and inform them of the data usage, confirm permission to use the data, request additional information needed for the analysis, invite participation in the analysis effort, and obtain proper citations and acknowledgments for the data.

Data for a measurement site scientist represents a potential revenue stream in that it enables analyses that can be carried out using the data. For the data contributors to databases and synthesis activities like in the LaThuile initiative, the potential reward is that they get coauthorship, acknowledgment or citation of their data but the risk is that the data contribution will not be acknowledged or will be misinterpreted. The measurement site scientists' conditions for sharing their data with synthesis activities are in general to receive proper "credit" for their data contributions, have an opportunity to explain peculiarities of their data, and ensure that their own local analysis efforts not be "scooped" by the synthesis activity. Regarding the request of proper "credit" there are different opinions and positions in the community, from the request to have an opportunity to contribute to the papers and become coauthor to an open policy where the only request is to be informed about the data use and publications that include their sites.

In the case of the LaThuile FLUXNET synthesis activity in 2007, there are three defined data policies (available at www.fluxdata.org) that try to cover the different views and positions of the sites' managers participating in the activity; each data contributor has the possibility to decide under which policy he or she wants to share his or her measurements. In addition, a steering committee that includes representatives of the measurement site scientists, regional networks, and synthesis teams manages the synthesis activities, ensuring the respect of the policies and trying to solve possible conflicts. A data use policy system structured in this way may look (and in fact it is) complicated and not completely open. However, an open and free data sharing, if it is not imposed directly by the funding agencies, is possible only by creating a community where participants trust the other participants' correctness and see the advantage of sharing data and having papers published where their data are included. The objective of the LaThuile FLUXNET 2007 policies system was to start the building of such community and the increasing number of participants that decided to share more openly their data is a good indication for the near future.

17.4.3 Additional Credit Possibilities

The primary scientific end product of the flux data, besides the database itself, is the publications generated by using the data in analyses. Measurement site scientists are generally interested in seeing their data used but also want to receive "credit" for having supplied the data, ensure that the appropriate funding sources have been acknowledged, and provide appropriate references for the paper to cite. In the past it was common to add as coauthor the data providers and their staff, even if their contribution to the paper preparation was limited to providing the measurements. The requests from the measurement site scientists were handled through a personal conversation between each data provider and the paper writing team and often the simpler way was to send a draft almost finished with an open coauthors list. Now the approach to this issue is changing and it is a common and generally accepted rule to add as coauthors only the scientists that contributed intellectually and substantially to the scientific message presented in the paper. In the case of publications that analyze data from a small number of sites, this communication between lead author of the paper and data providers is still straightforward and may easily lead to coauthorship, in particular if the data are fundamental for the paper's message and the measurements responsible help their interpretation. However, there is an increasing number of analyses which rely on a large number of global carbon-flux sites and/or incorporate data from many other sources and contributors. Providing the opportunity for coauthorship and intellectual input opportunities to all data contributors in these cases is not always feasible because it would mean to receive often opposite point of view and suggestion is not easy to reconcile. In addition, there are more and more scientific analyses that make use of EC products in complex models or analysis (e.g., in data assimilation systems) where the underlying methodology and statistical analysis is often so specialized and focused that it is difficult to fully understand for many of the scientist that are working in different fields. In these cases, also offering coauthorship could be problematic since it is important for all the coauthors to fully understand methods and results of the papers where they are involved. At the same time it is however important that the data contributors' rights are preserved and full acknowledgment and credit is ensured to their work.

This same issue has been faced by many other scientific fields and in particular high-energy physics which in the case of large experiments have thousands of scientists involved in each experiment. These groups typically converge on a formal set of guidelines for authorship and create a group authorship designator which is included in all publications. In addition, they typically converge on a single acknowledgment that can be used to acknowledge funding. In the case of large-scale experiments, these mechanisms are critical since otherwise, identifying individual contributions is too difficult. Flux datasets are a bit different since each individual site year has a clear set of contributors and funders. What is needed is a hybrid approach to the problem that incorporates aspects of the authorship mechanisms used in large experiments but that also enables tracking to individual contributions.

One possibility that has been already discussed in the past is to assign a Digital Object Identifier (DOI) (see also www.doi.org) to each dataset. The DOI is a digital identifier for any object of intellectual property and provides a means of persistently identifying a dataset on a digital network and associating it with related current data like authors, owner, characteristics, location, and all the other information relevant to describe and characterize the data. The system is currently used by most of the scientific journals to identify each single paper and could be easily implemented also for the data. However, the advantage of such a system for the data contributor is also linked to a change in the evaluation schemes adopted by funding agencies but also single institutions that should start to consider identifying datasets by DOI as a high-level product. The DOI release could be for example delegated to the regional databases and linked to a minimum level of quality, documentation, and completeness to be assigned, increasing the value of such identifier. An example of a data collection identified by a DOI is the Marconi Synthesis dataset (Falge et al. 2005).

References

Clery D (2006) Can grid computing help us work together? Science 313(5786):433–434

Drazen JA (2002) Who owns the data in a clinical trial? Sci Eng Ethics 8(3):407–411

English R, Schweik CM (2007) Identifying success and tragedy of FLOSS commons: a preliminary classification of Sourceforge.net projects. Paper presented at 29th international conference on software engineering workshops, Minneapolis

Falge E, Aubinet M, Bakwin P, Baldocchi D, Berbigier P, Bernhofer C, Black A, Ceulemans R, Davis K, Dolman A, Goldstein A, Goulden M, Granier A, Hollinger D, Jarvis P, Jensen N, Pilegaard K, Katul G, Kyaw Tha Paw P, Law B, Lindroth A, Loustau D, Mahli Y, Monson R, Moncrieff P, Moors E, Munger W, Meyers T, Oechel W, Schulze E, Thorgeirsson H, Tenhunen J, Valentini R, Verma S, Vesala T, Wofsy S (2005) FLUXNET Marconi Conference Gap-Filled Flux and Meteorology Data, 1992–2000. Data set. Available on-line [http//www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/811

Fienberg SE, Martin ME, Straf ML (1985) Sharing research data. Committee on National Statistics, National Research Council, Washington, DC, 240 pp

Fisher JB, Fortmann LP (2010) Governing the data commons: policy, practice, and the advancement of science. Inf Manage 47:237–245

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248

Heffan IV (1997) Copyleft: licensing collaborative works in the digital age. Stanford Law Rev 49(6):1487–1521

Heller MA (1998) The tragedy of the anticommons: property in the transition from Marx to markets. Harv Law Rev 111(3):621–688

Heller MA, Eisenberg RS (1998) Can patents deter innovation? The anticommons in biomedical research. Science 280:698–701

Hess C, Ostrom E (2003) Ideas, artifacts, and facilities: information as a common-pool resource. Law Contemp Probl 66:111

Hughes J (1988) The philosophy of intellectual property. Georget Law J 77(287):296–314

Lasslop G, Reichstein M, Papale D, Richardson AD, Arneth A, Barr A, Stoy P, Wohlfahrt G (2010) Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation. Glob Change Biol 16:187–208. ISSN:1354–1013, doi: 10.1111/j.1365-2486.2009.02041.x

Litman J (1990) The public domain. Emory Law J 965:975

May C (2000) A global political economy of intellectual property rights: the new enclosures? Routledge, New York

Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge

Papale D, Valentini R (2003) A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network ki. Glob Change Biol 9:525–535, ISSN: 1354–1013

Papale D, Reischtein M, Aubinet M, Canfora E, Bernhoher C, Longdoz B, Kutsch W, Rambal S, Valentini R, Vesala T, Yakir D (2006) Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. Biogeosciences 3:571–583, ISSN: 1726–4170

Posey DA, Dutfield G, Plenderleith K (1995) Collaborative research and intellectual property-rights. Biodivers Conserv 4(8):892–902

Rai AK (1999) Regulating scientific research: intellectual property rights and the norms of science. Northwest Univ Law Rev 94(1):77–152

Reichman JH, Samuelson P (1997) Intellectual property rights in data? Vanderbilt Law Rev 51(1):49–166

Reichstein M, Falge E, Baldocci D, Papale D, Aubinet M, Berbigier P, Bernhofer C, Buchmann N, Gilmanov T, Granier A, Grunwald T, Havrankova K, Ilvesniemi H, Janous D, Knohl A, Laurila T, Lohila A, Loustau D, Matteucci G, Meyers T, Miglietta F, Ourcival J-M, Pumpanen J, Rambal S, Rotenberg E, Sanz M, Tenhunen J, Seufert G, Vaccari F, Vesala T, Yakir D, Valentini R (2005) On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. Glob Change Biol 11:1424–1439

Suber P (2007) Creating an Intellectual Commons through Open Access. In: Hess C, Ostrom E (eds) Understanding knowledge as a commons: from theory to practice. The MIT Press, Cambridge