Data sharing across the Atlantic: Gap analysis and development of a common understanding

Jean-Daniel Paris, Nadine Schneider
Laboratoire des Sciences du Climat et de l'Environnement/IPSL
CEA Saclay, Orme des Merisiers
91191 Gif sur Yvette, France
nadine.schneider@lsce.ipsl.fr

Dario Papale
University of Tuscia
Department for innovation in biological, agro-food and forest systems (DIBAF)
Largo dell'Università - Blocco D, 01100 Viterbo, Italy
darpap@unitus.it

Hank Loescher
International Program Development, CEO office
The National Ecological Observatory Network (NEON)
1685 38th Street, Suite 100 Boulder, CO 80301
+1 720.836.2404
hloescher@NEONinc.org

BACKGROUND AND OBJECTIVE

The global need to develop large, cross continental environmental datasets has been recognized[a-f]. To address this issue, COOPEUS is a joint US National Science Foundation (NSF) and European Commission FP7 (in the frame of the European Strategy Forum on Research Infrastructures, ESFRI) supported project initiated in September 2012[g]. Its main goal is creating a framework to develop interoperable e-infrastructures across several environmental and geoscience observatories in Europe and US. The National Ecological Observatory Network (NEON in the US, www.neoninc.org) and the Integrated Carbon Observatory System (ICOS in the EU, http://www.icos-infrastructure.eu/) are two of these governmental supported observatories. Here, the data products from these two observatories are centered around greenhouse gas (GHG) concentration, carbon and energy flux observations, and the surface micrometeorology surrounding these measurements.

The **objective** of COOPEUS is to coordinate the integration plans for these carbon observations between Europe and the US. Even though both ICOS and NEON have the ability to collaborate on effective issues, we fully recognize that this effort cannot be effectively accomplished without the engagement of many other partners, such as; National Oceanic and Atmospheric Administration's Global Monitoring Division (US NOAA GMD), the Group on Earth Observations (GEO, www.earthobservations.org) and the Group of Earth Observations System of Systems (GEOSS), World Meteorological Organization (WMO, www.wmo.int), the Belmont Forum (www.igfagcr.org), NSF-supported EarthCube (earthcube.ning.com)and DataOne (www.dataone.org) projects, and a wide variety of regional-based flux networks (i.e., AmeriFlux, Fluxnet). Of course as COOPEUS continues to advance, this list of partners are not exclusive and are expected to increase. COOPEUS aims to strengthen and complement these partnerships through a variety of governance mechanisms, some informal and others formalized (e.g. Memorandum of Understanding), tailored to each individual organizational governance structure.

Several of these organizations (mentioned above) have a history of collaborating and sharing of data. In this historical context, we also have recognized what has worked, exiting limitations, and what can be improved in terms of data sharing and interoperability. This COOPEUS work task is building upon these relationships and working history to strengthen these approaches and collaboration.

COMMON POSITION ON DATA MANAGEMENT ISSUES

1. INFORMATION INFRASTRUCTURE INTEROPERABILITY[1]

Research Infrastructures (RIs) form the basis for large data sets that enable future large-scale, interdisciplinary science. As a result, developing the interoperability of data among RIs helps bridge the needs to foster intra-disciplinary and trans-disciplinary science and policy. The goals of trans-disciplinary science include the ability to forecast future conditions and ecosystems states. In order to meet this goal, uncertainties in data must be known a priori for any data assimilation technique or scaling activity. This provides additional rational for the interoperability of environmental RIs/data in order to develop a prognostic capability, as well as, for basic understanding, and for future planning, policy, and societal benefit. By doing so it also enhances the value of current scientific efforts and government investments. Currently, there is no accepted approach to make large datasets interoperable. Interoperability should be focused on physical, support and information Infrastructures. **Here we focus on primarily the Information Infrastructure, defined as** the end-to-end data flows that embody why we measure a specific quantity, that the specific

[1] Adapted from NEON Interoperability framework, H. Loescher and B. Wee

quantity represented (epistemological argument), how it was measured, its associated uncertainty, and how to maintain the data's integrity (archival and retrieval processes, providence). Its interoperability is defined here in four focus areas, including:

1. Linking Science questions to requirements: Science-driven requirements can help defining main interfaces among respective datasets
2. Traceability of Measurements (Use of Recognized Standards whenever available, Traceability to Recognized Standards, or First Principles, Known and managed signal: noise, Managing QA/QC, Uncertainty budgets due to measurement)
3. Algorithms/Procedures (What is the algorithm or procedural process to create a data product?, Provides "consistent and compatible" data, managed through intercomparisons, What are their relative uncertainties? Uncertainty budget due to processing)
4. Informatics (when available and recognized by the community, standards for Data and Metadata formats, Persistent Identifiers / Open-source, Discovery tools)

We agreed that the degree to which these data structures (observatories) are truly interoperable is the degree to which these four focus areas are adopted by the collaborative facilities.

2. PERIMETER OF DATA DEFINITION AND ASSOCIATED INTELLECTUAL PROPERTY ISSUES

The perimeter of data definition centered around four activities, including

1. Open Access. All participating RIs plan to offer free and open access to all data (reproduction costs if appropriate may apply), and the supporting information about the data. While all participating RIs are mandated for free and open access, data sharing policies also have to be harmonized. The auspices that these RIs were funded under are for scientific research, and we fully expect the data users will be far more diverse in the future. In some cases, open data sharing policies may include licensing to monitor (and control?) any unwanted redistributions. However, this latter point is still being discussed and is an ongoing development topic.
2. Data Description. One identified informatic knowledge gap is the need for RIs to better describe the data through more defined semantics, ontologies and controlled vocabularies. To further enhance the use of the data and data sharing activities to the broadest research community as possible, additional attribution is also needed, such as, metadata, protocols, methodologies, algorithmic processes, documentation, raw data and uncertainties should be available either online (e.g. when needed to correctly use the data) or upon request. A better definition of the data to be shared will advance the technological options for data stewardship.
3. Intellectual Property Rights (IPR). We also recognized that IPR may exist for specific aspects of the data (e.g. algorithms, or how the information is presented in proprietary templates). IPR issues are the focus of another COOPEUS work package, WP7, we are fully engaged in working with this working group, and is an active area of development.
4. Data integrity. Activities that maintain the data integrity have to be further explored, that may include joint quality control program. It is also noted that any report, validation report, or data comparisons should be made public and as part of the respective data product's metadata.

3. TIMING OF DATA SHARING

Each of the Observatories and partner organizations have slightly different mandates for data latency (the period of time the data is collected to the time that it is may available). Yet we all recognize the need for data quickly for real-time forecasting, novel approaches, extreme events, other quality control activities, etc. We all also recognize that the data must be quality controlled. As such, quality control (QC) of data should be done as rapidly as possible to enable sharing all information in near real-time (NRT) with the maximum level of quality. For example, i) ICOS NRT data systems make measurements quickly available in non-final quality level. Yearly dataset

releases are planned with comprehensive QC data report, and ii) NEON NRT data will be made available on ~ 15 min to 1 h basis as provisional (non-QA/QC'd), and fully QA/QC'd and vetted on the 30-45 d latency period. These two approaches are very similar, and based on our knowledge of the types of data usage, we think this is quite harmonious and not an impediment.

The maximum delivery delay for fully quality checked data should be about one month (see e.g. GMES requirements, GISC report 2010[x]) although for fluxes a longer time series could be needed to correctly apply all the processing steps. To illustrate the issue, data from ICOS atmospheric stations have an automatic QC mechanism (data flagging system) and can be delivered daily (with non-validated quality level, i.e, provisional).

4. JOINT QUALITY CONTROL AND ASSURANCE PLAN

One of the interoperability focus areas is the traceability of measurement to know standards and approaches (see above). As part of this effort, we plan on developing a joint quality control and assurance plan, and determine methods to optimize our efforts in a cost effective way. We then can determine which tasks can be support 'in-house' and which tasks have to seek extra mural support. This plan shall include;

1. Intercompare standards and calibrations
2. Redundancy test in the form of round robin standards and ongoing intercomparison of in situ measurements
3. Standardized (ISO defensible methodologies to quantify uncertainty)
4. Management Strategy that can include other partners and networks

Fact sheets of all data sets should be provided to address this requirement. It is important to qualify and quantify the differences in the collection as well as processing schemes and how these have an effect on the data.

5. COMMERCIAL ACTIVITIES

It is important to foresee the potential use of these data for commercial (private industry) activities in the plan of trans-Atlantic data use and their impact. Private industry using data acquired by the RI to add-value services and products raise the profile of these RIs and their socio-economic value. It also further justifies the public in the RIs. We welcome such usage, yet we also recognize that there may be some additional data quality, informatics, interoperability needs that we cannot foresee. To help understand the parameters around using these RI data in private industry, we plan on reaching engaging with one or more industries in a prototype or partnership effort.

Respective data policies should be harmonised in this respect. Rules should be defined for data usage with for-profit motivations. Licencing principles should be defined. This will be done in collaboration with WP7, because of the need to have international legal counsel.

6. DATA OWNERSHIP, CITATION, ACKNOWLEDGEMENTS AND TRACEABILITY

As part of the 'Informatics' interoperability focus area (above), we are working towards the use of Persistent Identifiers (PID) for our data sets. One of the challenges to this effort is for datasets that our time series continually become amended. Providence, acknowledgement, citation are part of the development of Persistent Identifiers (PID). Placing this PID structure into practice, we have determined four tasks that need to be addressed;

1. Bibliometric indicators. Scientists that participate in data collection require bibliometric indicators (citations of the dataset) to assess the scientific impact of the observatory/network/program as a whole, as well as for their own scholarly career. The mechanisms to track PIDs are still nascent. We are identifying approaches that can be used to integrate bibliographic citation systems to ensure both the traceability of the data and the proper attribution is tracked. At the same time it is important to build a

system for the evaluation of the scientists for their career that takes into account also the dataset publication.

2. Co-authorship. Guidelines need to be established and jointly adopted that determine if/how/when co-authorship on academic, scholarly, and policy-related publications for those that contributed towards the dataset is appropriate.
3. Other data ingest. Principal Investigators being part of the Research infrastructures should not exert any embargo (delay before data availability to others that is not justified by technical reasons) on the data. Principle investigators outside the RIs with projects or experiments that collect (monitoring) data should be encouraged to follow our jointly adopted data sharing principles and to avoid embargo on data in any case so that it becomes part of the public archive. These data should undergo strict QA/QC rigor. The data ownership needs to be documented and managed, and should include provisions for when a legal framework or constraints placed on by the funding agency imposes otherwise.

7. METADATA HARMONISATION

Metadata harmonization is desirable to enable the provision of more systematic information on data. However the acceptance of a single standard in the community is not foreseen. As such, we will continue to work with our Observatories and Partners to achieve a common accepted format, at least for some basic common meta-information.

METHODOLOGY

Method: We have met with representatives from the Observatories (NEON, ICOS) and the partners (NOAA, AmeriFlux) to share vision, identify common practices and differences (knowledge and programmatic gaps), and propose pathways to harmonize data access and data use policy and practices.

Process:

- **Initial meeting**: 27th March 2013, Biarritz/France, 8 participants
- **Teleconference**: 29th August 2013, 9 participants
- **ICOS-NEON final meeting**: 24th September 2013, Boulder/CO, 6 participants
- **Final reading and acceptance of the document** by all participants, closed 5 december.

Initial meeting: Preparatory workshop on "data gap identification"

Participants:

- Hank Loescher (NEON)
- Alex Vermeulen (ECN – INGOS)
- Dario Papale (ICOS ETC, UNITUS)
- Christoph Gerbig (MPI)
- Timo Vesala (FMI, ICOS HO)
- Jiří Kolman (CzechGlobe)
- Jean-Daniel Paris, Nadine Schneider (ICOS, LSCE)

Outcome: Draft Position paper with following key topics

- “Perimeter” of Data definition (Data formats, metadata, etc.)
- What is needed to share?, i.e., data along with protocols, methods, metadata
- The expected data latency period, the length of time that data becomes available after collection and quality control and assurance.
- How to jointly maintain QA/QC, i.e., Inter-calibration of data from different networks

- Special needs to enable commercial use of data
- Data citation, persistence, acknowledgements and data providence

This meeting led us to **new directions** that are **complementary** and consistent to findings from the status quo questionnaire (WP7) prepared by R. Huber (University of Bremen).

Teleconference: refining with key Stakeholders the Draft position paper (29th August 2013) to further identify commonalities and differences in principles and practices among the EU and US networks

Participants:

- Hank Loescher (NEON, USA)
- Margaret Torn (Coordinator, AmeriFlux, USA)
- James Butler (NOAA Global Monitoring Div head, USA)
- Philippe Ciais (ICOS Preparatory Phase coordinator, France)
- Timo Vesala (ICOS interim director, Finland)
- Margareta Hellstrom (Carbon Portal coordination, Sweden)
- Leonard Rivier (ICOS Atmospheric ThematicCenter head, France)
- Dario Papale (ICOS Ecosystem Thematic Center head, Italy)
- Christoph Gerbig (IGAS coordinator, EU FP7, Germany)
- Jean-Daniel Paris, Nadine Schneider (ICOS, LSCE, France)

Outcome: the approach was broadly endorsed, and new topics were added: harmonization of metadata internationally, technical issues on interoperability, data ownership

ICOS-NEON final writing meeting: 24th September 2013, Boulder/CO

Participants:

- Hank Loescher, Brian Wee, Jeff Taylor, Andy Fox (NEON, USA)
- Jean-Daniel Paris, Nadine Schneider (ICOS, LSCE, France)

Outcome: developed a path forward to complete these activities and writing tasks.

References

[a]Nature 2008. Special Issue: Big Data. *Nature* v. 455, 1 (4 September 2008), doi:10.1038/455001a.

[b]Economist, The 2010. Special Issue: The Data Deluge: Businesses, governments and society are only starting to tap its vast potential

[c]Hey, T, S. Tansley, K. Tolle, 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond WA. Pp. 252 http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[d]Nielsen, M. 2009. A guide to the day of big data. *Nature* 462, 722-723, doi:10.1038/462722a

[e]Science. 2011 Special Issue: Dealing with Data. *Science*, 331, February 11, 2011.

[f]Heinz Foundation 2006. Filling the Gaps: Priority Data Needs and Key Management Challenges for National Reporting on Ecosystem Condition. http://www.heinzctr.org/ecosystem.

[g]Cooperation EU US (CoopEUS), http://www.coopeus.eu/

[h]Geo Strategy report 2010 www.globalcarbonproject.org/global/pdf/GEO_CARBONSTRATEGY_20101020.pdf

[i]United States Global Change Research Program (USGCRP) 2013. Climate Assessment report; Third Assessment.

[j]Millennium Ecosystem Assessment (MEA) 2005. Ecosystems and Human Well-Being: Synthesis. Washington, DC: Island Press.)

[k]IOM (Institute of Medicine). 2013. Environmental decisions in the face of uncertainty. The National Academies Press, Washington, DC: pp 209.

[l]National Research Council (NRC) 2007. Understanding multiple environmental stresses: Report of a workshop. National Academies Press, Washington DC. pp

[m]Presidents Council of Advisors on Science and Technology (PCSAT) 2011. Sustaining Environmental Capital: Protecting Society and the Economy. Report to the President. www.whitehouse.gov/ostp/pcast.

[x]GISC report, Inese Podgaiska, Initial stakeholder list with selection analysis and linked to in-situ data requirements, October 2010, gisc.ew.eea.europa.eu