August 17, 2016

Memo

To: Eric Young and Jackie Burns
NRSP Review Committee

From: Jim Jones, Gerrit Hoogenboom, and Cheryl Porter (On Behalf of NARDN NRSP Proposal PIs)

Re: Response to review comments for NRSP_Temp 11 Proposal

Executive Summary

The overall goal of this proposal is to develop **infrastructure** that allows databases of any Land-Grant University, ARS lab, and the NAL to interconnect **AND** for users to access data from any location with links into the infrastructure. These bullets are summary responses to the concerns raised by the NRSP Review Committee.

- The proposal does not use ICASA formats for storing data, but instead uses a **flexible, efficient data format** developed by the international AgMIP modeling community over the past four years that is based on the published ICASA data dictionary.
- **We do include other (non-crop) systems**. We have an activity that is actively developing a data dictionary for dairy systems (Wisconsin and Kansas) for implementation using our approach,
- Yes, the business model was weak; we will strengthen it, including a plan to make it sustainable
    - Development of a national multi-state Experiment Station Project,
    - Strengthen involvement of both ARS and the NAL in the implementation of the national infrastructure for interconnecting databases, and
    - Work with the new AgMIP Open Data working group.
- We will be inclusive and cooperate with various partners, including all Land-Grant Universities, private companies, and others. We will facilitate this through one or more NIFA-funded workshops.
- We will work with professional societies and others to develop and publish quality control protocols.
- The national multi-state Experiment Station Project and Extension will be key to targeting all states for Outreach and Education on the concept and use of the Open Data system.
- We request that ES Directors appoint an advisory team to work with us to revise the proposal to ensure that it meets the regional and national goals of the Land-Grant Universities.

Ultimately the proposed NARDN-HD system will allow for any data collected by Land-Grant University faculty to be permanently stored beyond the lifespan of a project and to be locatable, accessible, machine readable, and usable and reusable by the science community at large to the benefit of society. Based on the general acceptance of the proposed concept by the Experiment Station Directors, we are waiting for guidance on how to move forward.

Narrative

We are very appreciative for the review of the NRSP_Temp 11 proposal that was developed by a larger group of PIs representing more than 15 different Land-Grant Universities. Although we were very disappointed by the decision of the NRSP Review Committee, we realize that the proposed goals are challenging and have not yet been achieved in the agricultural community in the US. There are technical, institutional preferences, and sustainability issues that have to be addressed. Those involved in the proposal have worked very hard to address these during the past four years, including acceptance of the proposed network by the broadest possible community within Land-Grant Universities and beyond. Broad community acceptance has been confirmed by commitments from:

- individual researchers from many states (more than 15),
- a large multi-state Experiment station project on dairy (25 states involved),
- the National Agricultural Library (NAL)

- USDA-ARS (see ARS Admin Shafer and Liu letter; contributions by 2 ARS labs in IA, AZ)
- the AgMIP global community of agricultural modelers,
- a major international agricultural research organization (the CGIAR),
- the International Life Science Institute (ILSI), and
- more recently (after proposal submission), strong interest by the national association of public yield trial leaders to harmonize yield trial data for access in quantitative formats.

Obtaining these agreements and commitments listed above has been challenging; it required considerable time working with groups to develop the approach over a period of about 4 years (with input from agronomists, IT experts, modelers, and collaborating initiatives), and to demonstrate that this solution is simple, agnostic relative to what contributing states use locally, and extensible to new variables and data domains. The approach was used to harmonize quantitative data for use in multiple crop models in five AgMIP data development workshops that have been held during the last 4 years. It was also demonstrated at a workshop at the NAL in May 2015, in which data in different formats from different states and ARS labs were harmonized and accessed through the Ag Data Commons at the NAL. Researchers from 14 Land-Grant Universities participated in this workshop and agreed to the approach because it does not require additional investments for expensive computer infrastructure, and it allows states that have databases in place to participate without changing their existing procedures. We are convinced that our proposed approach will lead to a widely-accepted solution that complements NAL plans for archiving national agricultural data.

Below we provide a summary of our responses to the review committee comments and a request for moving forward.

Key response summary (numbers refer to those from the NRSP Review Committee, Attachment 1):

1. ***ICASA Formats:*** The proposal does not use ICASA formats for storing data, but instead it uses a flexible, efficient data format developed by the AgMIP modeling community based on the published ICASA data dictionary (White et al., 2013; Porter et al., 2014; Boote et al., 2015; Ginaldi et al., 2016). A number of alternative approaches were considered during the last 5 years in consultation with IT experts, agronomists, and modelers before arriving at this consensus-driven solution. The proposed distributed data network design leverages NAL's Ag Data Commons which can provide access to a wide variety of disparate data, including harmonized data that allows users to access and use data from multiple sources.
    a. A general, extensible data dictionary for dairy research will be developed by University of Wisconsin and Kansas State University. This is important to the multi-state project (see letter from Ron Lacewell).
    b. We will pursue a planning grant with NIFA as suggested, and use it to hold a national workshop to allow all states to participate in the further planning and implementation of the proposed system.

2. ***Business model:*** A three-pronged approach will be used. First, we will develop a national multi-state Experiment Station Project to allow states to support their own faculty in participation in ways that contribute to their own research and extension goals. Secondly, we will work with ARS and the NAL to strengthen their involvement in the implementation of the national infrastructure for interconnecting databases, continuing with those institutions already involved (IA and AZ), but broadening to include LTAR sites. And, third, we will work with the new AgMIP Open Data working group, co-led by Cheryl Porter of University of Florida and Medha Devare of the CGIAR Consortium, to obtain funding for developing infrastructure components for interconnecting agricultural research data globally. Our business model will aim for the following characteristics: i) eliminate unrecovered indirect costs and other non-allowable items as requested, ii) modify the budget such that the system will be sustainable

without NRSP contributions after 5 years, iii) increase participation and contributions by ARS, ensuring that the system addresses the Open Data requirements for quantitative analysis and modeling by ARS labs as well as experiment stations, and iv) collaborate with NIFA and industry partners to create funding mechanisms for sustaining the system. We believe that this business model can be developed in time to submit a revision of our proposal in 2017. In addition, we will work with Experiment Station Directors to help guide our development of the business plan to ensure that it meets the technical and financial realities that they face.

3. ***Additional partners:*** We will hold open the effort to all states with a workshop with funding requested from NIFA to complement support by experiment stations in creating the new national multi-state project. We will also work with research leaders in different ARS locations and in Washington to develop stronger ARS roles in NARDN's sustainability, including development, implementation, and operation, and assist NIFA to include funding mechanisms for faculty, bring in additional partners (private sector, foundations, IT firms, international collaborators) to ensure that this harmonized agricultural research data initiative is effective and will also support other US efforts on data harmonization (e.g., GODAN, GEOGLAM, AgGateway, etc.).

4. ***Quality Control Process:*** We will work with professional societies, including Agronomists, Soil Scientists, Animal Scientists, Crop Scientists, and others, to develop and publish quality control protocols. This approach is planned for LTAR data in cooperation with ARS, but not discussed in the proposal. This is very important and a good suggestion.

5. ***Outreach and Communication:*** The proposed national Multi-State Experiment Station Project will be key to targeting all interested states. We expect that the individual state representatives will be the conduit to the scientists at their respective Land-Grant institution and introduce them to the concept of Open Data and data storage. We are planning workshops at the key societies mentioned previously to expose the broader community of agricultural scientists to the proposed NARDN Open Data system. We will expand our existing global outreach and acceptance process, working with IT specialists from the US, Europe, Australia, and China for a global harmonization effort through AgMIP and other networks.

Suggested Pathway Forward:
Given the fact that the proposed concept is well-supported among the Experiment Station (ES) Directors, we are requesting that the ES Directors appoint an advisory team to work with us as we revise the NRSP proposal to ensure that it meets the regional and national goals of Land Grant Universities to achieve the level of interconnectedness needed to address the key issues. In our attempts to better understand and model performances of various agricultural systems, we are hitting more walls due largely to the lack of data with sufficient environmental, management, and genetic variation. Although plant breeders routinely include large numbers of genotypes in their studies to understand and select varieties suited for target environments, data with wide variations in environments and management needed to understand and model GxExM interactions are generally not reusable in their existing formats (Gustafson et al., 2014). Our approach would allow these walls to be overcome and lead to more thorough understanding and more robust models and solutions. What we are proposing would not only provide a solution to the federal mandate of Open Data, but it would also lead to advances that are just not possible without having data with the variations in G, E, and M.

During the next year, we will accomplish the following, working with this ES Advisory Group:
1. Revise the proposal as suggested by the NRSP Review Committee, working with the ES Advisors,
2. Develop a new business plan with characteristics requested by the NRSP Review Team, working with the Land-Grant Universities, ARS, and the NAL,

3. Incorporate a first level of dairy data into the harmonization infrastructure, with the Dairy CAP in Wisconsin developing the data dictionary needed and UF incorporating it into the infrastructure for harmonization. This will also include working with the existing dairy multi-state project,
4. Work with the Wheat and Corn CAPs to ensure that their databases are harmonized and can be connected to the NAL central hub,
5. Work with ARS to design the interfaces for LTAR data, and with a few key states to incorporate the data from state-wide variety trials,
6. Develop a NIFA proposal for a workshop to help support the further development of the core infrastructure and with planning by different states,
7. Explore the development of a nationwide Multi-State Regional project on Open Data to enable all experiment stations and states to participate in the development and evolution of the system,
8. Continue to connect with broader public and private national and international efforts (GODAN, GEOGLAM, CGIAR, AgGateway, industries, etc.) to build toward the type of infrastructure needed for broader understanding, modeling, and analytics of agricultural systems.

We will request a small amount of support from different sources to continue with the development of the infrastructure for this system, primarily at UF, and by the Dairy, Corn, and Wheat CAPs in Wisconsin, Iowa, and Idaho, respectively. We will also work closely with the NAL to help increase their funding from ARS as noted in the Shafer and Liu letter. For example, during this year, Wisconsin will use funding from their Dairy CAP to complete an initial version of a data dictionary for dairy to show extendibility of the approach. We want to develop a proposal for the duration of 5 years that meets the requirements that you laid out in your comments, with guidance from the ES advisory team. Our intent is to develop an infrastructure that will enable Land-Grant University faculty to contribute their valuable data to an Open Data system that includes features for extension to other types of data. Ultimately the proposed NARDN-HD system will allow for all data collected by Land-Grant University faculty to be permanently stored beyond the lifespan of a project and to be locatable, accessible, machine readable, and usable and reusable by the science community at large to the benefit of society.

We look forward to hearing from you and request that you work with us and support us in this effort. This will ensure that we do not lose time with the development of the infrastructure and at the same time we can develop a proposal that more fully addresses your goals and concerns for the NRSP_Temp 11.

Key References:

Boote, K. J., C. Porter, J. W. Jones, P. J. Thorburn, K.C. Kersebaum, G. Hoogenboom, J.W. White, and J.L. Hatfield. 2015. Sentinel Site Data for Crop Model Improvement—Definition and Characterization. In: Improving Modeling Tools to Assess Climate Change Effects on Crop Response, Adv. Agric. Syst. Model. 07. ASA, CSSA, and SSSA, Madison, WI. doi:10.2134/advagricsystmodel7.2014.0019.

Ginaldi, F., M. Bindi, A. Dalla Marta, R. Ferrise,S. Orlandini, F. Danuso. 2016. Interoperability of agronomic long term experiment databases and crop model intercomparison: the Italian experience. Europ. J. Agronomy 77:209-222.

Gustafson et al. 2014. Climate adaptation imperatives: 1. Untapped global maize yield opportunities. Int. J. Agric. Sustainability. DOI:10.1080/14735903.2013.867694.

Porter, C.H., C. Villalobos, D. Holzworth, R. Nelson, J. W. White, I. N. Athanasiadis, S. Janssen, D. Ripoche, J. Cufi, D. Raes, M. Zhang, R. Knapen, R. Sahajpal, K. Boote, and J. W. Jones. 2014. Harmonization and translation of crop modeling data to ensure interoperability. Environ. Mod. & Software 62:495-508.

White, J. W., L.A. Hunt , K. J. Boote, J. W. Jones, J. Koo, S. Kim, C. H. Porter, P. W. Wilkens, and G. Hoogenboom. 2013. Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. Computers and Electronics in Agriculture 96:1-12.

Attachment 1. NRSP Review Committee Recommendations (in Eric Young email, 6/7/16):
The NRSP-RC recommends rejecting the proposal as presented. The proposal may be resubmitted for consideration next year, or future years, provided the following concerns are addressed.

1. There was a lot of concern with ICASA as the core standard, its focus is on crop simulation and may not be appropriate for other types of data sets. Alternative data formats should be considered, the scope of data types proposed may be too broad for a single data format. A revised proposal should resolve the issue of a data format that is not applicable to many potential uses of data.
    a. The proposal may need to consider different formats for plant and animal or other subsets of data types or limit the project to data sets where a single format is appropriate.
    b. The writing committee might consider applying for a NIFA planning grant to bring diverse data format expertise together to settle on the best format(s).
2. The business model needs to be better articulated, more realistic, better leveraged, and show sustainability beyond 5 years. A revised proposal must address the short-term commitment of NAL, keeping in mind that a new budget line in USDA NIFA is unrealistic. Also, leveraged support should not be primarily in-kind salaries and unrecovered IDC.
3. Consider bringing in additional partners, particularly private entities, for expertise and financial support; ex. data analysis firms, consultants, private industry, other federal funding agencies, foundations, etc.
4. Develop a quality control process for data sets being received to ensure harmonized data is reliable.
5. Develop a more definitive outreach and communication plan that explains the target audience and outcomes desired for workshops or other activities; for the harmonized data sets; and for the ultimate end user of results. Define how Extension and education fit into a continuing outreach and communication effort.

The committee would like you to know that the concept is well supported, timely, appropriate, and created lot of positive interest; but implementation details as presented have too many problems and barriers, and do not appear to be sustainable. There is clearly significant power in having big data sets available for further use and the proposal to bring ARS, NAL, and Land Grant Universities together on this issue is very good.