Meeting 1: 2015-04-03
Attendees: Ge Peng, Ruth Duerr, Toni Rosati, and Sophie Hou

Meeting Notes:

1. Preservability:
    a. The evaluators should help in defining what the "community" is that their dataset is part of and identifying if it is community-based or national or international standard-based in terms of archive process and metadata.
    b. If only metadata is available through the repository but the actual datasets are still in the hands of the PIs and may not be available sometimes in the future, this might be a preservability level "0" because the datasets are not potentially stored digitally anywhere. It is discoverable but not accessible from a trusted repository.
2. Accessibility:
    a. With the organizations that have more consistent applications of its stewardship practices, the datasets stewardship maturity level would likely to reflect the organizations as well.
    b. Discoverability is not the same as accessibility (see the example in Preservability).
        i. More discussions are needed to determine the pros and cons of combining or separating these two entities. If a dataset is not accessible or does not exist any more, is there any need for assessing its stewardship maturity? Probably not. However, there are many cases where the data is accessible but not readily discoverable and vice versa.
    c. The current matrix assumes that datasets are digital and publicly available online. Those assumptions may need to be relaxed to allow application to a wider range of datasets. Wording might need to be modified to accommodate datasets with confidentiality and privacy concerns. How readily and easily for users to get information on the procedures could be used.
        i. Peng's afterthought: Like the idea of defining levels of readiness and easiness for users to find information about the procedure of requesting and timeliness of receiving proper access permission to restricted datasets. Questions for discussions: Are they measurable?
            1. Ruth - definitely. In some cases, an actual proposal for use is required but access can be granted rapidly thereafter (e.g., a matter of hours). In other cases, it may take months to negotiate access.
            2. Ruth - For example, can timeliness be quantifiable? Also true…
            3. Ruth - It isn't just an issue of how easy/timely it is to find information about processes for requesting access - there may actually be a variety of access mechanisms of varying degrees of onerousness. For example, when dealing with human subject data, obtaining access to data that has been anonymized is usually quick and easy. However, a fill-out-a-form to get instant

access in a protected environment to data that hasn't been anonymized is still pretty straight forward, just a bit more technically (VPNs and other software protections of the data) and legally complicated. However, truly sensitive data may not be available unless you travel to the site, submit to a pat down (no electronics permitted) prior to entering a secure site.

    ii. Progressive? Any other attributes in addition to readiness, easiness, and timeliness?

Meeting 2: 2015-04-06
Attendees: Ge Peng, Ruth Duerr, Toni Rosati, and Sophie Hou

Meeting Notes:

1. Usability:
    a. Usability is really meant to reflect how "easy" it is to learn, understand, and use the dataset.
        i. Peng explained that one of the challenges in defining the dataset-centric stewardship maturity matrix is to distinguish what an organization does, what we as stewards do, and what has been applied to the dataset.The capability on the organization level does not always imply that capability is automatically with the dataset. For example, a data center may have a THREDDS data server available, a tech steward could potentially put the dataset on the THREDDS server. However, until that dataset is on THREDDS, the subsetting and aggregating capability will not be there for the dataset.
        ii. Peng's afterthought: It may be a good idea to provide an overall framework of the different perspectives of maturity (Organization, program, and individual dataset from top to down (tiers); product, stewardship, and use/service from origination to dissemination (left to right, life stages) in the summer session.
            1. Ruth - agreed
    b. Easy to understand is part of usability, but does not quite cover all the aspects of easy to "use".
    c. Ruth and Toni actually separated the category into sub-categories.
        i. Expertise - in the instrumentation, general domain, interdisciplinary science domain, general science, general public.
            1. Peng explained that information about the availability of expertise has been considered to be a part of use/service maturity matrix.
            2. Ruth has indicated that user expertise is a part of OAIS usability assessment.
        ii. Metadata and Documentation

            1. Documentation is more for human, and metadata is more for machine. As a result, while documentation could be complete, metadata could be incomplete and vice versa.
        iii. Access
    d. How about dataset format (in terms of file)? It is currently a part of usability.
        i. The dataset format's usability might need to be "community" based, and for now the definition of "community" could be self-defined.
        ii. Data organization could also affect the usability.
    e. A definition for of community standard should be provided for this category?
        i. However, who should be responsible for defining the community? Also, which definition should we use? There are many standards even within the same discipline.
    f. This category would be a good one to be discussed further during summer.
2. Production Sustainability:
    a. Definition for this category: how is the dataset being produced?
    b. Production Sustainability is about whether the data is still being produced and updated. Not about if the dataset is preserved in its perpetuity.
    c. What about one-time observation dataset type?
        i. In this case, this category would be not applicable.
3. Data Quality Assurance:
    a. For during production; the current matrix is intended to measure whether information about the data quality assurance (including screening) procedures/practices is captured and conveyed. However, this might be confusing since the matrix is about data stewardship, so that the data quality assurance might be construed as the assurance measures that the repository has.
    b. Ruth and Toni assessed it based on what the repository used to assure data quality.
4. Data Quality Control/Monitoring:
    a. It is information about the procedure that can be both implemented during the production or while the data is in the repository.
5. Data Quality Assessment:
    a. This category meant to capture whether the data is being evaluated scientifically.
    b. Should the data manager be assessing the "science soundness"?
        i. In general, the data manager might not be able to assess this, but we are trying to capture whether the data is being assessed.
    c. Quality metadata assessed: this is meant to document whether the quality of the data quality metadata section has been assessed and whether the result of assessment is being recorded. In other words, the metadata for the dataset should included a section recording the quality of the data based on criteria, such as accuracy, completeness, and consistency. Then, the assessment is applied to this section to ensure the quality of this section of metadata.

        i. “quality” is assumed being defined elsewhere. However, it would be important to ask the users to clarify what are the criteria used for the assessment.
        ii. Three levels of evaluation:
            1. Level 1: Data quality is assessed and available.
            2. Level 2: Data quality is documented as “quality metadata”.
            3. Level 3: The “quality metadata” is reviewed and assessed to ensure that the metadata reflects the data quality accurately.
    d. In Ruth and Ton’s case, since the datasets are mostly observations, “Data Quality Assessment” did not quite apply.
        i. Peng thinks the data quality check on the PIs’ level should be a part of the information that should be made available, and this is the part of the data quality assurance, control/monitoring, and assessment that could be recorded.
    e. In the case of instrument measurement, how do we capture the data quality assessment?
        i. However, instruments could also include “human being,” so there should be a distinction between the “type” of instrumentation.
            1. Manually measurements are not necessarily more or less than instrument measurements.
            2. The availability and the description of measurement method might be a better judgement of how mature the stewardship is in this case.

Peng’s afterthought: This is a very interesting and important point for further discussions including within the ESIP community.

For example, station measurements such as meteorological and oceanic variables are observations. They are routinely undergoing data quality assurance such as range, spatial and temporal consistency checks before made available to users in addition to regular instrumental calibration (pre- and post-employment). Data quality control/monitoring procedures are often carried out either manually or automatically. For example, quality control procedure for in situ measurements from the TAO buoy array (http://www.pmel.noaa.gov/tao/proj_over/qc.html). And data quality assessments can be carried out by comparing measurements with other source of data of the same variables - sometimes they lead to uncovering defects of instruments or calibration procedures (e.g., Dickinson et al. 2001: doi: 10.1175/1520-0426(2001)018<0799:CBTTBA>2.0.CO;2; Peng et al. 2013: doi: 10.2481/dsj.14-019).

The question is: should any types of observations be following some kind of assurance and control/monitoring procedures to ensure its quality maturity? If so, in addition to the availability and the description of measurement method, information on whether there is a procedure in place for ensuring data quality via assurance, control/monitoring, and assessment; the availability and description of those procedures, should be used to measure the dataset stewardship maturity, although the responsibility of ensuring original product quality lies with data producer(s).

Defining roles and responsibilities of stewards and stakeholders in ensuring and improving data quality will definitely help improving data quality management process. It is a subject that Peng is doing her research on and planning to give a talk on the subject in the Information Quality Summer ESIP Session proposed by Rama.

6. Transparency/Traceability:
    a. The OID does not necessarily need to be a standard format; it needs to be at least to be unique. Standard-based OID will tend to improve its interoperability.
    b. The provenance might be manifested in different ways, but whichever format is used, the provenance should be traceable and available.
    c. We might need to remove references to ATBD and OAD and provide a more generic description of the documentation that we need for achieving Level 3 and 4. References to ATBD and OAD could be used as examples for data products.
7. Data Integrity:
    a. This category is quite straightforward.

The discussion within the group has touched several times on if a consistent definition of "community" may be beneficial in terms of best practices and standards such as naming conventions or metadata standards and help users to understand what the community the assessment results is associated with. The category for community could be one of the followings: local (project, group, institution), domain/discipline (physical, biological, ecological, climate modeling, etc.), national (U.S.), or international. Discussions on the need and suitability of this classification will be useful.

Sophie has indicated that reading the Peng et al. (2015) paper was helpful for the assessment using the stewardship matrix. Peng stated that it was assumed that users of this matrix would read the paper first for the scope and rationale for each key component (or entity), and examples for some components. The template was provided for convenience of making an assessment and documenting the results and assessment history.

After the meeting, Peng added a note in the template to encourage users to read the paper prior to using the template (http://figshare.com/articles/NCDC_CICSNC_SDSMM_Template/1211954).

(Peng is also planning on adding a short description of the scope and an example of best practice(s) or standard(s) for each key component in a new version of the template, based on the recommendation by Sophie.)