# Background and Significance

The use of globally unique identifiers in biology and other scientific disciplines is increasing like never before, due to growing data production, the push to give credit for data as a unit of publication, and funding requirements to make data discoverable for reuse. Concomitantly, new systems for minting, tracking, resolving, and querying identifiers are also on the rise. As in any rapidly developing field, new and changing technologies and social practices associated with identifiers have led to growing pains and bottlenecks in workflows, as developers work to meet implementation demands and scientists adapt to new practices. As new solutions are released, new problems arise due to the increasingly distributed nature of digital research, along with the multiplicity and the unbounded nature of digital data. This causes practitioners and technologists to work on after-the-fact solutions as opposed to imagining and solving next generation challenges. Most recently, the possibilities enabled by multi-disciplinary research and big data highlight the challenges involved in cross-referencing data from diverse fields that use a variety of identifiers, workflows, and best practices. We propose a workshop that focuses on identifier solutions to big, interdisciplinary data. Using as an example the successful model of MIT Hacking Medicine [1], the proposed workshop will be designed to capitalize and catalyze the expertise, innovation, and creativity of multi-disciplinary participants that will be tasked with imagining, discussing, proposing, and hacking next generation solutions for data identifiers.

Multiple reports on the characteristics of and best practices for different identifier types have been published recently [2–6], and several recent meetings/workshops have addressed identifiers and related issues (Table 1). While some of the reports and workshops listed here provide introductory notions about identifiers, others have gone further, identifying challenges within specific domains (e.g., provenance tracking for biological specimens [2]) or addressing interoperability issues between different types of identifiers (e.g. personal identifiers and data identifiers through the use of linked data [7]). Linking and integrating data and personal identifiers is emerging as the new frontier, yet most of the focus remains on identifiers for publication (not necessarily for reuse), with topics such as citation guidelines, overcoming social issues, or challenges with legacy data. **Less has been discussed about the vision of an identifier cyber-infrastructure that spans the entire data lifecycle (including pre- and post-publication) in the context of very large, complex, multi-disciplinary datasets,** specifically, what a conglomerate of services would look like and how will data be represented. Necessarily, such infrastructure has to have as much automation as possible to be sustainable.

**Table 1. Recent meetings related to the proposed workshop**, with a focus on identifiers (I), data (D), or semantics (S). This list is inevitably biased by the experience of the organizers, but provides an indication of the relevance of the proposed workshop.

|  |  |  |  |
| --- | --- | --- | --- |
| **Meeting title** | **Focus** | **Date** | **Location** |
| Managing Natural History collections data for global discoverability [8] | I/D/S | September 2015 | Arizona State University |
| Identifiers and Linked data in the research space seminar [9] | I/S | July 2015 | Melbourne |
| Knowledge Exchange Workshop Research Identifiers - National approaches to ORCID and ISNI implementation [7] | I | June 2015 | London |
| DASISH workshop on persistent identifiers [10] | i | Dec. 2014 | Cologne |
| Community next steps for making globally unique identifiers work for biocollections data [2] | I | Nov. 2014 | Stockholm |
| Persistent Identifiers, Herbarium Workshop [11] | I/D | September 2014 | Kongsvold, Norway |
| iDIGBio - Data Modeling Workshop [12] | D/I/S | March 2014 | Honolulu |
| Workshop on Persistent identifiers @iPRES - focusing in interoperability [13] | D | September 2013 | Portugal |
| SciencePAD Persistent Identifiers Workshop [14] | I | Jan 2013 | Zurich |
| Workshop: Metadata and Persistent Identifiers for Social and Economic Data [15] | S | May 2012 | Berlin |
| NOAA Environmental Data Workshop [16] | D | May 2012 | Maryland, USA |

We suggest that the lack of published empirical data on how identifiers function pre-publication and during the transition to publication reflects 1) the intense demand on identifier system developers to produce systems and solve urgent problems, leaving less time for research and publication, and 2) a lack of opportunity for developers to interact with peers outside their organization and discipline. Thus, while individual researchers may be keenly aware of the problems faced within the systems or landscape they develop and use (Table 2), there is still no clear picture of the most pressing needs in identifier technology for scientific research across disciplines or which solutions work best for different situations. For the same reasons, it is likely that solutions to specific challenges with identifiers faced by one discipline exist in other disciplines, but that they have not yet percolated across disciplines, leading to unnecessary research limitations and redundant development.

**Table 2. Examples of identifier challenges that the organizers are aware of from their own work.** Experts in other domains are likely to share some of these challenges or have solutions to some of them, with non-perfect overlap, leading to synergy.

|  |
| --- |
| There are 4 different standards for UUIDs (universal unique identifiers), which creates a small change of ID conflict. Furthermore, some services mint IDs that they call UUIDs, but do not follow one of those standards (e.g., they incorporate date or location information). This can cause conflicts when merging data across systems. |
| Identifiers are typically minted by data aggregators and publishers, which are mostly downstream of the point of collection or data generation. However, researchers have repeatedly stated the need to have an identifier assigned as soon as the data is created, and we generally lack tools to support globally unique, permanent ID creation in the field. |
| Data aggregators (e.g., purpose driven databases) do not typically adopt pre-existing identifiers because they 1) assume they don not exist, 2) do not trust that the incoming identifiers are stable enough to adopt in their own systems, or 3) the identifiers follow a convention that is different than their system. Thus, new identifiers are typically minted for data elements at each step in the data aggregation chain (e.g., collector, museum, database) |
| Poor semantics of community-adopted vocabularies make it difficult to know the nature of ID relationships. For example, “associatedSpecimens” is a Darwin Core term typically used for storing previous identifiers but it could literally mean any possible association (e.g., a previous ID for the same thing, an ID for a host or parasite or sibling, and ID for a duplicate specimen). |
| Services are needed that can distribute and assign identifiers automatically (e.g. a university needs to distribute identifiers across different labs, departments, libraries, etc.) |
| Unclear how will the network of identifiers, domains, dates, and relationships of interest to users should be represented for big research projects with numerous derivative and related datasets. How will so much identifier information be abstracted for meaningful discovery. |
| Few tools for applying ontologies in conjunction with identifiers |
| Few tools to embed metadata about subsets/components of data as identifiers are minted |
| Unclear how related datasets should be added to existing datasets and identifiers in post publication phase |
| How to integrate identifiers into computational/experimental/observational research from data generation and processing including the important metadata attached to them (e.g. from imaging tools, data acquisition systems, data analysis tools, etc.) |
| How can repositories provide services to support authentication and validation of data related to identifiers |
| What and how much metadata should be included in the identifiers for reuse in different applications |
| Do we need domain specific identifiers? |
| Will boundaries between data pre-and post-publication data and identifiers ever disappear? |
| New economic models are needed to sustain identifiers |
| Who should be responsible for maintaining data identifiers (libraries, individuals, identifier service providers)? |
| Some existing identifier services may not be scalable/automatable. What are their limits? |

We want to give people working with identifiers the opportunity **step out of their usual sphere and think about next generation problems**, not only those problems they have at hand. Therefore, we propose a workshop in which people will be asked to not only state problems and articulate solution, but will be asked **imagine those problems for double or triple the amount of data they currently have**, and for related data in some other domain. This workshop should be a venue to pose questions related to future possibilities and more open science.

We propose that the time is right to bring together identifier practitioners and experts from across discipline to identify and elucidate the problems, desires, wishes, and technical issues associated with identifiers, and to propose solutions. Scientific identifier meetings to date that we are aware of have been primarily domain focused (e.g., biomedical, biodiversity, earth science, geography), but identifier challenges cross disciplines. It is our sense that development within domains has reached a sufficient level to solve many challenges, but has not yet reached a point that solutions are being broadly disseminated or even applied across two slightly overlapping disciplines. For example, in biology, we have a pretty good idea how to cite datasets [refs] or mint identifiers for life science data objects [17], whereas in earth science, they have a functional system in place to register and track physical specimens. Yet bio-informaticians struggle with tracking specimens, and earth scientists lack a cohesive data infrastructure. This meeting will seek synergy by inviting diverse participants from multiple domains, who might otherwise never interact with each other. Participants will have the opportunity to challenge each other with new perspectives and problems, as we will push the limits of all of their systems with truly large datasets.

Public meetings hosted by organizations such as Digital Curation [18], iPres [19], the American Geophysical Union (AGU) [20], or Research Data Alliance (RDA) [21], exist as fora for the discussion of identifier technologies, but they have much broader mandates and are unlikely to provide the focused environment that we propose here. We aim to identify the key accomplishments and next generation challenges in identifier technology that may be of use to biological and interdisciplinary researchers and identifier systems developers. The objectives of this workshop overlap significantly with the organizers’ own research, and although our research is largely in biological big data, we expect this meeting to be mutually beneficial to all participants. The outcomes of this workshop have the potential to impact the way identifiers are applied and used across science. Most interestingly, the model used for the workshop is designed to go further and articulate solutions including both technical and implementation plans as well as a sustainability one. By the end of the workshop, the teams will render projects that can become proposals, be implemented as solutions, or become entrepreneurial initiatives.

# Workshop plan

## Objectives

This workshop has three primary objectives:

**1. Document the current status of identifier technologies across scientific disciplines.** We will invite participants involved in the identifier landscape to present the newest developments in the field, and use the collective wisdom of the group assembled to discuss the applicability of the newest developments to large and complex datasets in the context of multi-disciplinary research and address the following two objectives.

**2. Identify pressing needs in identifier technologies and services.** To achieve this objective, we will carry out small group and plenary activities to address the following questions: What problems with identifier technologies have we already solved? Under what situations do existing solutions fall short? What are the most serious bottlenecks in minting and using identifiers? How truly scalable are existing systems? How would identifier services for multi-disciplinary activities look like? What are the open research questions and technical problems for identifiers and identifier systems? Where can solutions from one domain or discipline be applied to others? Which problems should we address first to have the biggest gains in science?

**3. Articulate solutions in identifier technologies and services.** The majority of the time at the workshop will be dedicated to developing solutions to the challenges identified by objective 2. The type of solution will vary, depending on the type of problem. If manual curation of a certain step is a major bottleneck, the solution may be to develop tool to automate that step. If a serious pain point is that end users consistently supply the wrong information at some point, the solution may be to create educational materials. If one group decides that a particular challenge cannot be overcome without longer, concerted effort, the solution could be to draft a proposal to fund that effort.

## Organizers

The three chairs for this meeting represent a diversity of complementary expertise (biology/bioinformatics, library and information science, and computer programming/development). We have all worked with big data in biology and identifier technologies in one capacity or another (see biosketches), but also have connections to other disciplines such as Earth Science (e.g., Walls is on the advisory board of the of the iSAMPLES RCN through EarthCube and has participated in several workshops sponsored by the Semantic Observation Network at NCEAS) and Geology (Esteva is the data curator on the Digital Rocks Portal [22]). Through our connections at UA’s Bio5 Institute and UT’s TACC, we regularly come in contact with researchers from a variety of scientific and technical disciplines, and we are used to working with groups of researchers who come from different fields and may be collaborating for the first time. All co-chairs have organized workshops, hackathons, or meetings of various sizes (e.g., [23–26]). In addition, we will hire a professional meeting facilitator.

**Ramona Walls**, Ph.D., iPlant Collaborative, Bio5 Institute, University of Arizona

**Roles:** Local meeting host, co-chair, recruitment of biologists/bioinformaticians and experts from other science domains

**Maria Esteva**, Ph.D., Texas Advanced Computing Center, University of Texas at Austin

**Roles:** co-chair, recruitment of information scientists

**John Deck**, M.L.A., University of California at Berkeley

**Roles:** co-chair, recruitment of identifier experts and informaticians across domains

**Meeting facilitator, TBD.** A professional facilitator will be engaged to help with planning and running the workshop, and assist with post-meeting reporting. Due to the rapid turnaround of this proposal, we have not yet selected a specific facilitator, but we have experience working with several people who would be good for this meeting. We feel that a professional facilitator is key to the success of this meeting, because s/he will be able to facilitate active participation by everyone at the meeting (which is especially important when the participants will come from disparate backgrounds), keep the agenda on track, and help ensure that we deliver our expected outcomes.

Additional logistical support will be provided by UA’s Biosphere 2 (to arrange lodging and catering) and the support staff at the iPlant Collaborative (help with travel arrangements and reimbursement).

## Meeting organization and activities

### Meeting organization

The meeting will incorporate elements of a workshop and a hackathon, with a mix of plenary and small group activities (see Agenda, below). The first day of the meeting will be dedicated to describing both what we do and do not know about identifier technologies. Are best practices consistent across scientific disciplines? What problems have already been solved? Where do solutions break down? Where are the major gaps in both knowledge and technology? The outcome of the first day will be a map of the existing landscape of identifier technologies and services for science, including gaps and a list of challenges/problems that need to be addressed.

The rest of meeting will be organized similar to a hackathon, but with the intention of being inclusive of many different skill sets (i.e. not just coders/developers). Small groups will be expected to develop solutions to the set of challenges identified on day 1. While those outputs may include development of software, they may also include non-technical products like a draft proposal, a survey, or educational materials. We will encourage groups to choose a solution that can be finished at least to the point of a prototype/full draft within the time constraints of the workshop. Coffee breaks, lunch, happy hour, and dinner provide important opportunities for informal discussions and new connections, and attendees will be encouraged to take part in these less formal parts of the meeting.

**NSF support will be acknowledged during the meeting and conveyed to all NSF-supported participants.**

### Proposed agenda

**Day 1**

* **Morning**
  + Introductions
    - Welcome by chairs, lay out workshop goals and expected outcomes.
    - Get to know each other activity such as “speed integration” (participants move around room and discuss a specific question in pairs for 2 minutes).
  + Lightning talks: Attendees present their current solutions to identifier issues.
* **Mid-day**
  + Breakout groups will work on two tasks:

1. Sketch out a list of identifier solutions/technologies that you find most useful.
2. Discuss and compile list of problems you are facing.
   * For this activity, breakout groups will be pre-composed to contain people from different backgrounds, rather than self-selected.

* **Afternoon**
  + Whole group: Build a map of the existing landscape of identifier technologies and services.
  + Brainstorming as whole group: What really bugs you? - Pain points in identifier technology from all perspectives. This should serve to compile and organize the list of problems identified by breakout groups.
  + Breakout groups (same groups as earlier): Prioritize the challenges/problems from the whole list compiled by all groups
  + End of day activity: one high point and one low point, from all attendees.

**Day 2**

* **Morning**
  + Whole group: Review and compile the prioritizations from the day before.
  + Whole group: Brainstorm and suggest possible solutions for breakout groups.
  + Whole group: Distill and cluster ideas from brainstorming to arrive at a final list of challenges.
  + Participants decide which project to work on and form new, self-selected breakout groups. Note that a small set of workshop participants will be designated as technical experts, and we will ask them to circulate among groups and provide expertise as needed, although they will still be allowed to choose a specific breakout group of their own.
* **Late morning/afternoon**
  + Breakout groups - Devise a plan to address the selected challenge (educational material, proposal, or technical solution), begin work on plan.
  + Reconvene mid-afternoon for quick status updates from each group, resume breakouts.
* **Mid-day:** Lunch inside Biosphere 2 and optional tour of for participants.

**Day 3**

* **Morning**
  + Present preliminary results from breakouts and get feedback from whole group.
    - People can shift among groups, merge groups, or split groups at this point.
  + Breakout groups - Finish work on projects
* **Afternoon**
  + Present final results to whole group and discussion. Decide which results should be included in the review paper or if any merit their own publication.
  + Draft meeting report and review by compiling outputs of all previous activities into a shared document.
    - Documentation will be ongoing, but we will spend some time at the end reviewing to make sure everything is there.
    - Solicit volunteers for a writing committee to finalize report and review.
  + Closing activities
    - Solicit one high point and one low point from all attendees.
    - Group discussion: What were the strengths and weakness of the workshop? What will participants carry home to implement in their own organization?

### Expected Outcomes

1. Map of the current identifier technology landscape, focusing on how pieces that are currently disconnected might work together. This should be a map both literally (in the form of a figure) and figuratively (in text).
2. List of the most pressing research questions and practical challenges in identifier technology.
3. Output of breakout groups. During days 2 and 3, 5-10 breakout groups will work to solve a selected set of challenges identified by the whole group on day 1. Each breakout group will be tasked with delivering a tangible (but not polished) product by the middle of day 3. Breakout groups will determine for themselves what that product should be, but we will encourage a range of outputs including educational materials, technical implementations, and draft proposals.
4. Review paper detailing outcomes 1-2, and 3 if appropriate.
5. Meeting report to NSF-BIO detailing how the outcomes of the meeting benefit their research communities and others.

### Broader impacts: Dissemination of results and educational opportunities

When it comes to the adoption of new technologies (for identifiers or most anything) dissemination and education go hand in hand and are as important as development. Although this workshop focuses on technical issues, we are well aware of the social issue surrounding identifiers, and they will be an underlying force in shaping any products of the workshop. We believe that some (perhaps many) of the social issues in identifiers (e.g., poor adoption, improper use, inadequate metadata) can be improved with technical solutions, such as the development of more user-friendly tools or data quality control/assurance algorithms. Esteva and Walls have an NSF EAGER proposal in review at this moment, in which we propose to address certain technical issues with identifiers, such as linking data at multiple locations and data identity. We expect the outcomes of this workshop to provide valuable input into that project and our ongoing work for the iPlant Collaborative and TACC.

The first key method of disseminating our results will be publication. We will produce both a review paper in a peer-reviewed journal and a meeting report/white paper to submit to NSF. All participants will have the opportunity to participate in the drafting and review of both publications, but it will not be required. If an appropriate post-doc or grad student can be identified at the meeting, we will offer them the opportunity to be lead author on the peer-reviewed paper and provide mentoring to do so. We aim to have about ⅓ of our participants (~10 people) be early career scientists, not only because of the fresh perspective that they can bring to identifier technology, but also because we know that early career scientists are our most powerful tool for disseminating and driving adoption of new ideas. We will also use social media to publicize the output of the workshop, before, during, and after, and produce a press release immediately following the workshop. iPlant and TACC both have communications support personnel who can assist with this.

Our second approach is that each breakout group will be tasked with including dissemination as part of their development. For example, any breakout group that develops educational materials will be asked to include a plan for sharing those materials with the scientific community. This could occur through the organizations represented at the workshop (e.g., if the output is a tutorial on preparing metadata for DOIs, it could be hosted by EZID and iPlant, who host and generate DOIs, respectively). Any technical implementation developed as part of the workshop should be targeted at existing workflows, and thus have an application in place before development begins. We will strongly encourage groups preparing proposals to identify a funding agency and program, and to follow through with submission. We are very happy to work with NSF personnel before or during the workshop to identify appropriate topics for proposals. If the outputs of any breakout group (or several combined) merit an independent publication, we will encourage the group members to publish it. Otherwise, outputs will be included in the main workshop publication. Finally, all workshop outputs will be made available through a public repository on Github (see Data Management Plan).

### Assessment plan

We will assess the workshop based on three criteria: 1) Did we achieve our objectives? 2) Did we deliver the expected outcomes? 3) Did participants feel it was time well spent?

The first two criteria can be assessed directly. The meeting report should be completed and the peer-reviewed article submitted within 3 months of the meeting. These publications will document whether or not we achieved our objectives and expected outcomes. The final question is subjective and more difficult to assess. We have built into the agenda opportunities for feedback from participants both during and at the end of the meeting. We will record and summarize these responses. In addition, we will distribute a questionnaire near the end of the last day, asking participant to evaluate the meeting for both content and format. We will make these results available to NSF for purposes of future workshop planning.

## Recruitment of participants

We have requested funds to provide meals and lodging during the meeting for ~30 participants and to provide travel support to ~10 US and 2 international participants from that pool. In addition, we expect to recruit 4-5 local participants from the University of Arizona and Arizona State University, who will be able to attend for the day if they are willing to commute. The meeting will be open for anyone to attend, but coverage of expenses will be limited by the requested funds. The bulk of support will be awarded preferentially to early career scientists and members of under-represented groups, but we reserve support for 5-6 targeted invitees with particular expertise in identifier technologies. International travel support will be used to recruit two people from developing countries. Walls will attend the Biodiversity Information Standards (TDWG) conference in Nairobi, Kenya this fall and, while there, promote the workshop to informaticians from developing countries.

To recruit participants, we will advertise the workshop through mailing lists such as EZID, RDA, EarthCube, iPlant, XSEDE, PASIG, DLF, etc., as well as through our own contacts. We expect about half the participants to come from the life sciences and will try to recruit the other have from non-biological scientific domains. We will create a webform to use as an application for attending the workshop and requesting support. The application will solicit information such as why the person wants to attend, what they expect to contribute, what they expect to get out of the meeting, what particular challenges they face with identifier technologies, and if they can supply datasets for testing applications. We will also request information on gender (self-reported, rather than categorized), home organization, and career status (e.g., graduate student, post-doc, mid-career, late-career, etc.), in order to achieve a balanced distribution of participants. There will be a place on the application to let participants know that childcare will be available for a fee during the workshop. In addition, as formal invitations are sent to those participating, we will ask them to respond and let us know if they need childcare or any other accommodations. These requests will not factor in to decisions about who to support financially.

We will assemble an independent committee of TACC and iPlant personnel to review the applications. We will device a set of criteria and a scoring system for selecting participants and have the committee scores all applications, to avoid biases. Scoring will be based on education, involvement in the theme, background or experience, and work on a project that involves metadata/publication/datasets.

Two of the meeting chairs (Walls and Esteva) are women. Esteva is Hispanic and Walls comes to science as a second career after many years in fashion design. In addition to being a developer, Deck has formal training as a landscape architect, and runs an internship program on his working farm. As such, we are all keenly aware of the need to include women and people with diverse backgrounds in technical discussions and workshops, and the challenges in doing so. To provide an environment that is inclusive to all participants, regardless of their background or experience level, we have intentionally chosen to deviate from the standard hackathon format to allow more flexibility in the types of outcomes. In recruiting participants, we will make it clear that we value a diversity of perspectives and experiences.

## Logistics

We plan to hold the workshop for three days between December 2015 and March 2016. This will allow us sufficient time to plan and recruit, but not delay to the extent that the topic grows stale. Logistical issues (arranging shuttles, reimbursing expenses) will be handled by the local host (Walls) with support from iPlant administrative staff. The local host will also arrange for on-site child or family care services with a local certified agency, if requested by participants. All three meeting organizers will be responsible for preparing workshop materials, planning the final agenda, recruiting participants, and producing the final report.

The meeting will take place at Biosphere 2, in Oracle, Arizona. It is part of the University of Arizona and is accessible to people with disabilities. We will ask participants to let us know if they have any special needs and accommodate them as appropriate. Biosphere 2 is located a little over 1 hour from both the Phoenix and Tucson airports, in a fairly remote and very scenic location. It has excellent conference and housing facilities and good wifi with sufficient bandwidth for all participants [27]. iPlant and Walls have extensive experience hosting meetings and workshops at Biosphere 2, and the venue has been very well received by attendees. Because of the relatively isolation, there is little opportunity for distractions, and participation in group dinners and social events is generally very high. Rooms at the Biosphere 2 can accommodate up to four people and are arranged in adjoining “casitas” (little houses), making it convenient for any participants who need to bring family members with them. Last, but not least, we find it rather easy to recruit people from the rest of the country to come to meetings in Arizona in the winter.