Project Description

1. Overview and Objectives

Just as awareness and use of globally unique persistent identifiers (GUPIs) for digital data are gaining momentum, shortcomings in their implementation also are becoming apparent. In the field of Biology, there is a disconnect between the data management practices of researchers who collect and analyze data, and the way in which identifier solutions are applied at the point of data publication. Repository systems for curated data that offer access to GUPIs, require significant manual data preparation at the tail end of a research project. More often than not, this approach renders enough information for purposes of data citation but is insufficient for data reuse, because it fails to capture information collected as part of the research process, preserve identifiers created prior to publication, and record post publication events. Furthermore, the notion of a single identifier type and location for all data in a project clashes with the complex, distributed nature of many biological datasets, whose parts or copies may reside in different repositories and across diverse storage platforms. Many existing data publication models ignore the evolving nature of research data and the relationships it establishes throughout the continuum [1] of data management and archiving. This leads to the misleading notion that the curation process stops at the publication stage, when in fact it needs to be ongoing. Moreover, many digital repositories cannot truly commit to long-term archiving of data to which they assign identifiers, and post-archiving events such as transfer to other repositories and registration and verification of copies and derivatives of an identified dataset, and assessment of integrity may not be managed and accounted for over time. For the researcher interested in reusing data and finding it through a GUPI, getting to the complete representation of a dataset in a distributed environment such as the current one may constitute a challenge.

As data grows in size and complexity, the challenges mentioned above will only increase. There is a growing need for services to verify, track, and report events (i.e. provenance) in relation to identified datasets over time. Such services should start as early as possible in the life of a research project and be as much as possible automated. We thus propose to conduct research into "**identifier services**" with a focus on biological data. We will develop a set of proof of concepts/prototypes to achieve the following objectives: **1)** model identifiers to the lifecycle management of bio data including their transition into GUPIs, **2)** conduct automated verification of the data linked to GUPIs to track presence at registered locations and integrity and identity over time, and **3)** assess how collection creators that participate in the prototype development use identifiers and respond to identifier services.

We are submitting this project as an **EAGER** proposal because it entails an exploratory proof of concept and thus it does not fit other NSF calls. It contains service approaches for GUPIs that to the best of our knowledge have not been addressed. This research has the potential to introduce new modes of managing identifiers and corresponding data in relation to: the research lifecycle, the methods employed and their efficiency, and the social consequences for researchers and data users.

To fulfill these objectives, we will leverage and expand data modeling work done as part of the iPlant Data Commons (iDC), and use **real world biology datasets** from iPlant, the Texas Advanced Computing Center (TACC), and the National Ecological Observatory Network (NEON), focusing particularly on DNA/RNA sequence data. We will leverage technologies that have been developed for iPlant such as the Agave API [2] and bio data analysis methods for data verification and validation. We will expand upon the widely-used concept of a **landing page**, to build an interface where individual users and repositories can manage and track identifiers pointing to and from datasets, their metadata and provenance, to bind distributed components [3], and where automated data verification services can be recorded.

2. Research Significance

Much of the current research and development around digital identifiers focuses on facilitating data citation and discovery post-publication. Projects such as ODIN [4] facilitated the integration of persistent identifiers for data, researchers, and publications to improve visibility of curated data and research as well as to ease metadata input [5]. Building upon ODIN, project THOR [6] will continue to pursue a network of identifiers through data integration, interoperability and linked data. While the latter research focuses on data at the point of publication, different voices in the community have expressed the need for more empirical research on the use of multiple identifiers [7] to support bounded [3], sustainable and verifiable collections [8]. Amongst those, FORCE11 [9], an open community that aims to transform scientific information and scholarly communication into part of a global knowledge network through findable, accessible, interoperable and reusable (FAIR) data practices, calls for next steps to assess the quality of research outputs from inception and into the future [10]. In turn, the bio community, has expressed the need to better understand the use of identifiers across the research process for purposes of tracking provenance and to assemble large datasets [11].

In this proposal we address problems arising for large, dispersed, biology datasets and changing events. Instead of assigning identifiers only at the last stage for curated datasets, we will explore usage of different identifiers throughout the continuum of data management, publication, archiving, and reuse. For this we will prototype an "identifier infrastructure" for identifiers management, permutation, and data validation/authentication across time. We are interested in understanding the technical and social implications of a more decentralized model for identifier services that embed and automate data curation best practices required to assign, maintain and verify identifiers including post-archival events. The project, based on a variety of test-bed collections, will provide insights into what constitutes useful global identifiers to represent data as tested by users in the field**.** This project will deliver prototypes that will be fodder for future development proposals and to post-custodial theories in application to data [12].

3. The Current Landscape for Biological Data

Contemporary bio data are created and used in the midst of complex, distributed social and computing environments. This is particularly true for big data, which by its very nature cannot be generated, analyzed, and housed long term at a single location. For example, most genomic sequence data must move at least once, from the sequencing center to the investigating lab, and they are likely to move multiple times as part of the analysis, publication, and reuses processes, including pre-publication sharing among multiple collaborators, which is now the norm. Other forms of big data in biology – sensor data, automated field- or image-based phenotyping data – undergo similar processes and therefore face a similar challenge: how to manage and track a dataset across the continuum from generation to re-use to ensure its identity and integrity.

Increasingly, publishers are requiring a GUPI for datasets that accompany scientific papers. While this is laudable, obtaining an identifier often implies depositing data in a single place, which in biology is not realistic, as many established repositories accept only certain file types or have size limits. For example, the Sequence Read Archive (SRA) only accepts certain types of sequence files, leaving the other components of a dataset without an identifier and the collection disassociated. Moreover, researchers in the life sciences domain have expressed the need to assign identifiers as soon as data is created and specimens are available [11] and to account for and track different data components (copies or subsets) which may be stored in different repositories, may have other identifiers, or may never be published. Even when these identifiers can be minted, challenges in maintaining and linking to them remain.

Although funding agencies strongly advocate for data sharing and reuse [13, 14], most data repositories follow practices that strongly support citation and discovery, and only a few are starting to support some services for evolving datasets and post archival events. In biology, the state of knowledge for data management/curation can be exemplified by repositories or networks such as the International Nucleotide Sequence Database Collaboration (INSDC, e.g., SRA or Genbank), DataOne, and Dryad [15–17]. These repositories focus on the use of metadata to indicate data quality and rely on submitters to provide extensive metadata at the time of publication. DataOne notably emphasizes the importance of collecting metadata and exploiting its utility throughout the data life cycle [18], yet there are limited resources to support this practice, and researchers often must rely on laboratory information managements systems.

Biological data repositories rely largely on internally generated identifiers (e.g, Bioproject, Biosample, or sequence IDs from INSDC or URIs from databases like UniProt) or DOIs (e.g., at DataOne or Data Dryad), both of which, like all identifier types, have different strengths and weaknesses [7, 19]. For example, INSDC or other database identifiers can be issued before publication, but do not come with a guarantee of long-term persistence, whereas DOIs are intended to persist, but are only issued at the time of publication and so far, the persistence is based on social contract that is difficult to reinforce. Although there are mechanisms to link pre-publication, internal identifiers to INSDC identifiers or DOIs, we have heard numerous stories from scientists describing difficulty in identifying the provenance of published data, such as sequence data in national repositories (e.g., [20, 21]). Furthermore, as new types of standard identifiers are coming online (e.g., RRIDs, ORCIDs [22, 23]), additional research is needed that examines the functionality of these different identifier types across data workflows and data continuum.

Within this landscape, iPlant cyberinfrastructure (CI) provides end-to-end services for life sciences computational research, reducing the need to move data repeatedly and facilitating provenance tracking. As part of iPlant, the iPlant Data Commons (iDC), currently in progress, aims to provide long-term storage of valuable biological datasets, reduce the burden on researchers in submitting such data for publication and archiving, and to serve data in a way that facilitates their reuse in new analyses, particularly within the iPlant CI. The iDC will host public data and serve as a transitional point to prepare and submit data to other repositories, including the assignment of GUPIs (e.g., currently testing submission pipeline to SRA). iDC services are conceived as modules to be used across the iPlant computation and data storage environments either as part of end-to-end research or at a given stage of a data lifecycle, such as to prepare data for publication. The initial implementation of the iDC focuses on genomic data, which is iPlant's larges source of data and may become one of the largest sources of such data in the world [24]. Despite iDC services, not all analyses can be done on a single CI and there is still need for iPlant datasets to change locations, to bring new data into a dataset, to import information about the specimens that are sequenced, and to bind dispersed objects. The research proposed herein will address that gap and contribute to further developments in iDC and elsewhere by prototyping identification services to represent the research process, wherever it happens, and link data through identifiers as they evolve, improving long-term access to sustainable, verified data.

4. Research Plan

To accomplish the objectives stated in Section 1, we will undertake three interrelated proof of concepts. Each of them involves tasks and their evaluation. Tasks 1 and 2 prepare the stage for the proof of concept studies, which will involve real world datasets, be built over the technical infrastructure Agave, and use the landing page as the interactive interface to services, data representations and event recording.

4.1 Data selection

We have identified multiple possible datasets to use for this research. The collections, either forming or completed, are hosted in iPlant, TACC, or NEON and have or will have copies or subsets in other repositories such as SRA. We will select datasets that contain genomic sequence data and accompanying metadata of varying completeness, as well as other types of associated data, such as genomic variants, environmental data, or phenotypic data.

Task 1. *Work with collection creators to map their datasets to our existing genomic data model and to specific tasks in a manner that is appropriate to their workflows and lifecycle stages.* This will take place during the first three months of the project. We have working relationships with data creators at iPlant, NEON, and TACC, who can participate in these activities.

4.2 Technical Platform: Agave and Landing Page

The Agave platform [25–27] will be the underlying technology to implement the identifier services and to manage the landing pages here proposed. Using the Agave Systems APIs, we will enable users to register system(s) where data is stored, connecting the data with the identifier metadata. With these connections we can then take various actions to validate data/metadata and verify data using automated systems. In addition, Agave-based services can be leveraged to assist researchers with additional data-lifecycle needs, such as movement between storage locales, and automatic update of metadata. The landing page will be the face of the manual and the automated functions. It will be dissociated from diverse locations and from different identifiers but will track, verify, and report independently from and on behalf of individuals or repositories in custody of data.

Task 2: *Implement the project's backbone: Agave and landing page.* The landing page will serve as a web application where users will create and enter the information for their projects. It will be powered by the Agave platform to provide the data management APIs for the subsequent tasks (i.e. the functionality that we describe below). During the research, the prototype service will be hosted at iPlant as the interface where clients can talk to a distributed infrastructure including where the data and metadata are stored. For future development, the services could live anywhere, and the application can easily scale horizontally by utilizing services such as Amazon Cloud or other Content Delivery Networks.

4.3 Objectives, Associated Tasks, and Evaluation

4.3.1 Modeling identifiers across the research lifecycle

During the course of research, when data is created and transformed and many decisions are made ad hoc by the researcher, is when data dispersion and vulnerability happen. Waiting to deposit data until the research project is finalized with all different subsets in a stable form, in a single repository, and using one global identifier may not be realistic nor representative of most research. Within a project, pieces of data get treated differently by different researchers, and evolve through experiments and results at an uneven pace. Moreover, data are not the only element that needs to be recorded and described in a project. Processes, provenance, specimens and publications are also key to bio data interpretation and reuse. A project may reuse data or subsets of data from other research which already has global identifiers, or data with local identifiers that need to be tracked. Once experiments are completed, each set of resultant data may need its own GUPI for purposes of linking with specific publications. In addition, some experiments could be published before the entire project is completed, requiring an identifier to be related to others at a later date. In the context of complex datasets, diverse identifiers required to bind data and the rest of the entities during the research phases are difficult to manage and to transition into GUPIs. Such a complex scenario leads to the questions over which we build this proof of concept: How do we track large, complex data across time, and how do we assign identifiers that pull these together to obtain a complete representation of the research? How can an identifier morph from internal to public? Which services can support this and provide ease to the user?

The co-PIs in this proposal, Esteva and Walls, designed a data model for sequencing/genomic data that is being implemented in iPlant's Data Commons. A different version of the data model is already successfully implemented for experimental engineering data in the Digital Rocks repository [28]. The genomic data model (Fig. 1), assumes metadata and identifiers for the following research entities: project, specimen, specimen context, sequencing,

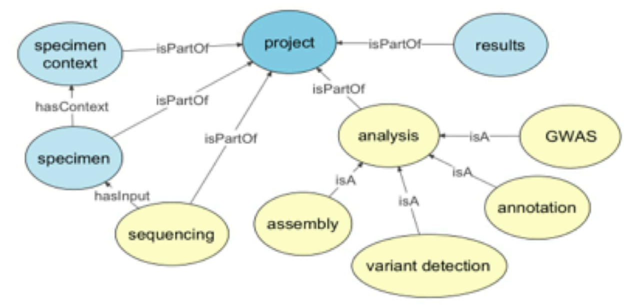
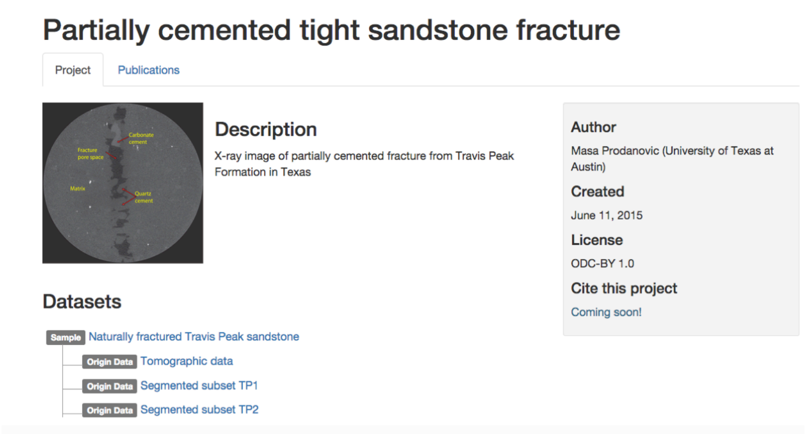


Fig. 1. Simplified version of iPlant's genomic data model. Blue: physical or information entities, yellow: processes.

analysis, and publication of results. These entities embed semantics about the corresponding data and expect the use of identifiers (internal or external, global or private, e.g., ARKS, legacy data reuse, specimen IDs. etc.). Within the model, when a dataset is linked to an identifier, it inherits the semantics and available metadata that correspond to that entity automatically. A particular research project may not include all the entities, or it may need to repeat some of them. Using the data model we will accomplish the following task:

Task 3: *Use the genomic data model as a framework to explore the use of identifiers across the research process.*

Using the infrastructure from Task 2 and the model described above, the service will bind corresponding objects and render a graphical representation of the dataset and related information as users register and tag their data, wherever the data is located, similar to Fig. 2. Currently, through iDC, we will be able to provide ARKS, DOIs, and internally generated UUIDs and provide support external identifiers such as ORCIDs, BioSample IDs, or local IDs.



Partially cemented tight sandstone fracture

Project | Publications

Description

X-ray image of partially cemented fracture from Travis Peak Formation in Texas

Author

Masa Prodanovic (University of Texas at Austin)

Created

June 11, 2015

License

ODC-BY 1.0

Cite this project

Coming soon!

Datasets

Sample Naturally fractured Travis Peak sandstone
- Origin Data Tomographic data
- Origin Data Segmented subset TP1
- Origin Data Segmented subset TP2

Fig. 2. Example of the Digital Rocks Portal developed by Esteva and Hanlon. Tree structure in lower right shows project data linked to data

We will explore the implementation of identifier workflows in the following collection type cases:

- An evolving collection in pre-archival stage as it transitions through computational analysis stages (i.e. as data are being produced).
- A distributed collection whose subsets are transitioning to more than one long-term repository but who will leave a complete copy on another repository (i.e. final hosting at multiple locations). E.g., users may get DOIs at different locations for the distributed parts of a dataset, but we aim to provide the single location and semantics to establish that these are all pieces of the same work.
- A finalized collection in transition to a long-term repository that does not provide global identifiers, when creator wants to request GUPIs through the identification services. This is a pressing use case for researchers who are submitting sequence data to INSDC but want to keep a copy of the data in iPlant for reuse within the CI.
- A finalized collection with a landing page to which identifiers from derived/versioned/related datasets have to be integrated.

Evaluation: Throughout this task we will evaluate the following developments:

- Degree to which processes can be automated.
- Semantics and metadata integration as identifiers are minted or captured in relation to the model.
- Graph of the project based on the data model as users register, fork, and use identifiers.
- How well the data model fits in relation to what a creator considers to be the next steps in the research process.
- Permutation of local/internal identifiers into GUPIs when datasets are ready to be published.
- Comparison the functionality of different identifier types (ARK, DOI, URI, UUID, and local identifiers including file-naming conventions) at different points in the workflow.

4.3.2 Identifiers and data verification services

Data sustainability and persistent identifiers are inextricably tied. As repositories and institutions accept bigger and more complex data, making sure that the datasets attached to persistent identifiers are viable and authentic throughout time is an as yet unresolved challenge. Organization such as EZID [29] rely on the repositories/institutions that contract their services to maintain identifiers via a social contract that cannot be enforced. In turn, institutions, which may also distribute identifiers among departments and researchers, may not have the resources to verify them. This panorama requires new, automated services to verify data linked to identifiers across time. Such services could be decentralized from repositories and institutions, and used by individual researchers or centers independently from where the data is hosted. We identified three characteristics to verify a dataset: a) availability from a persistent location with a record of provenance, b) verification of data integrity, c) confirmation of identity, understood "whether an object x and an object y are the same object" [30], over time and in relation to copies distributed in other locations. The following three tasks will address these characteristics. Each task will depend on data locations being registered via Agave. Having specified the access mechanisms to a repository through Agave, we then can point to the identifier and location for a dataset to implement the following tasks:

Task 4: *Implement a service that automatically checks that there are data at the diverse locations pointed to by a registered GUPI.* Through Agave, a spider process will go to the data location, validate that it is there and/or verify that the server did not get shut down. To facilitate recording the movement of data, we will provide tools through the landing page to download and move data (Agave currently accomplishes these functions). This will allow recording the new data locations automatically. In the case that data are not in the specified location, the service will contact the user to request a new location. If there is no response from the creator in a specified time, the service will interact with the identifier provider service to indicate that the data is not at the location and the identifier is not maintained. As data moves from host to host, all verification events will be recorded to trace provenance.

Task 5: *Implement a service to check dataset integrity.* Upon verifying location, the service will recalculate the checksum of the data and assess it against the recorded checksum (calculated when the data is registered in the identifier service). This service is especially important for datasets that have identifiers but are not managed within a trusted repository that verifies data integrity on a regular bases. In the case of data accessible through ssh, we will also verify mime type, file size, and file name (file system metadata). In turn, considering the platforms in which the test-bed collections are hosted, we will investigate what sort of system inquiries we can achieve once we obtain access to that storage platform. For example, if the data is on iRODS, or DSpace, could we develop tiny acts that run against these systems through Agave or Docker services to verify the data metadata?

Task 6: *Implement a service to systematically check putative multiple copies of the same dataset for identity.* Data identity refers to the characteristics that make a specific dataset definable, recognizable, and distinguishable from other objects [3]. When a user requests a persistent identifier to be mapped to multiple copies of the same dataset, each copy of the data has to be verified as being the same. In the simple case in which two repositories maintain an exact copy of a dataset as binary files, verification can be done by comparing file metadata (e.g., size, creation date, checksum, etc.). However, identity can be complicated by the different storage and distribution formats used by different data management systems, which is common for many biological data object types. Likewise, additional identification information added by different repositories may lead to slightly different versions of records/files for the same biological information. On the other hand, two sequence files may differ by one nucleotide due to a Single Nucleotide Polymorphism (SNP), but have the same file size and creation date, or contain multiple sequences with small variations, leading to false positives in identity scoring. This task will assess if copies of data accessible at the locations specified by the user can be identified as the same object.

Building off work in [31], we will develop a model to systematically access the multiple data objects registered with our service and conduct content-based comparisons and verification. The model will support accessing content of a data record through multiple means such as via database query, web service API, and file parsers of common structured biological data types, and then make comparisons among all copies of the data. This proof of concept will be implemented only for sequence data, but many of the elements will be applicable to other data types.

Evaluation: In tasks 4 through 6, we will consider the accuracy of the integrity and identity results, tested according to computer science methods and qualitative methods, as assessed by creators and advisory board members in relation to conceptualizations of data integrity and identity. The evaluation will also assess the efficiency and scalability of the methods and the degree to which automation is achieved.

4.3.3 Users responses to identifier types and to identification services

This proof of concept entails observations of the key components of tasks 3 to 6 detailed above. From this evaluation, we will derive information about data identification practices by data creators and their response to the new developments prototyped in this project. We recognize the limitations of this evaluation, both in terms of the time available to conduct observations and to the limited scope of the project. Because these are prototypes, usability studies do not apply. Nonetheless, the results will serve to understand usage of identifiers and to provide feedback for further development of the identification services.

Task 7: *Evaluate the adoption of different identifier types and identifier services by collection creators.* We will observe users as they interact with the identifier services through logs of identifier requests and registration, recorded events, the representation of their collections, and the issues that they report to us. We will create a structured observation log focusing on the following themes:

- Types of identifiers used and or proposed by data creators for managing data prior to publication.
- Data creators' decisions as to which and how many GUPIs to utilize in relation to the data model proposed.
- Data creators' decisions as to how to utilize GUPIs in relation to the structure of their datasets, their parts and locations.
- Use of one GUPI versus usage of sub-identifiers in cases of collections with multiple sub-components.
- Understandability of the landing page interface including the recorded events.
- Accuracy of the research representation graphed through the use of identifiers as evaluated by collection creators.
- Usage of and value attributed to the services.
-

. Timeline

Project goals and tasks	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
Requirements gathering/advisory committee specs	X							
T1. Select and map datasets	X							
T2. Implement AGAVE services for basic landing page	X	X						
T3. Use data model as a framework to explore multiple use cases		X	X	X	X	X	X	

T4. Implement automated data availability check		X	X	X	X			
T5. Implement a service to check dataset authenticity				X	X	X	X	
T6. Implement a service to check and compare dataset identity				X	X	X	X	X
T7. Evaluate identifier services from data creators perspective			X	X	X	X	X	X