Internet of Samples:
Toward an Interdisciplinary Cyberinfrastructure for Material Samples

Grant proposal submitted to the National Science Foundation (NSF) on November 1st 2019. Funded by NSF August 2020.

Grant #2004839 - Columbia University*
Kerstin Lehnert (PI)*, Columbia University 0000-0001-7036-1977
Sarah Ramdeen (CoPI), Columbia University, 0000-0003-1135-5942

Grant #2004562 - University of Arizona
Ramona Walls (PI), University of Arizona 0000-0001-8815-0078
Sarah Kansa, Open Context, The Alexandria Archive Institute, 0000-0001-7920-5321
Eric Kansa, Open Context, The Alexandria Archive Institute, 0000-0001-5620-4764
Raymond Yee, Open Context, The Alexandria Archive Institute, 0000-0002-6241-2559
John Kunze, Technology Consultant, California Digital Library, 0000-0001-7604-8041

Grant #2004642 - University of California, Berkeley
Neil Davies (PI), University of California, Berkeley, 0000-0001-8085-5014
John Deck, University of California, Berkeley, 0000-0002-5905-1617
Christopher Meyer (CoPI), Smithsonian Institution, 0000-0003-2501-7952
Thomas Orrell, Smithsonian Institution, 0000-0003-1038-3028
Rebecca Synder, Smithsonian Institution, 0000-0002-0028-6139

Grant #2004815 University of Kansas Biodiversity
David Vieglais (PI), University of Kansas Biodiversity Institute, 0000-0002-6513-4996

** Lead Institution / Lead PI*

DOI: 10.5281/zenodo.4480196

Collaborative Research: Frameworks: Internet of Samples: Toward an Interdisciplinary Cyberinfrastructure for Material Samples

1. INTRODUCTION

NSF's Big Idea of "Harnessing the Data Revolution" [U1] envisions a cohesive approach to research data infrastructure that will enable data-driven discovery, better data mining, machine learning and more. We propose to develop an innovative and transformative cyberinfrastructure that will integrate material samples collected and used for scientific research into the digital research data ecosystem: iSamples, the 'Internet of Samples'. Our goal is to establish a core research infrastructure that provides consistent services for unique and persistent sample identification and sample metadata registration across all disciplines. iSamples will enable inter- and cross-disciplinary use of samples, making them FAIR research resources and research outputs (Findable, Accessible, Interoperable, Reusable; Wilkenson et al. 2016). iSamples infrastructure has the potential to transform scientific research by making it possible to easily discover, reuse, and fuse sample-based data, paving the road towards advanced mining of sample-based data within and across domains.

We propose a multi-disciplinary and multi-institutional project to design, develop, and promote iSamples as a national-level service infrastructure to uniquely, consistently, and conveniently identify material samples, record metadata about them, and link them persistently to other samples and any digital content derived from them, including images, data, and publications. We propose a flexible and scalable architecture, which will broaden adoption and implementation by diverse stakeholders, thus contributing to the sustainability of the iSamples system. iSamples will capitalize on existing identifier infrastructure such as IGSNs (International GeoSample Numbers; [U2]) and ARKs (Archival Resource Keys; [U3]), but is agnostic to identifier type. Likewise, iSamples will encourage a high-level metadata standard for natural history samples (across biosciences, geosciences, and archaeology), and support community-developed metadata standards in specialist domains, but does not require adherence to a particular standard. By integrating established discipline-specific infrastructure at SESAR (geoscience) [U4], CyVerse (bioscience) [U5] and Open Context (archaeology) [U6], iSamples will build on existing capabilities, make these consistent, and expand their reach serving science and society much more broadly.

2. INTELLECTUAL MERIT

Material samples are a basic element for reference, study, and experimentation in many scientific disciplines, especially in the natural and environmental sciences, material sciences, agriculture, physical anthropology, archaeology, and biomedicine. Observations made on samples collected in the field and in the laboratory constitute a critical data resource for research that addresses grand challenges of our planet's future sustainability, from environmental change; to food, energy, and water resources; to natural hazards and their mitigation; to public health (e.g., IWGSC 2009). The large investments of public funds being made to curate huge volumes of samples that have been acquired over decades or even centuries and to collect and analyze new samples demand these samples to be openly accessible, easily discoverable online, and documented with sufficient information to make them reusable (McNutt et al. 2016). They need to be linked to the data derived from them (interoperable) and to the interpretations of these data published in the literature, which is also essential to making sample-based data reproducible. In order to achieve this, material samples need globally unique, persistent identifiers (PID), which link to persistent landing pages with metadata that describe the sample and its provenance and which allow unambiguously linking samples with data and publications (Lehnert & Klump 2012).

The current ecosystem of sample and collection management in the U.S. and globally is highly fragmented across disciplinary communities and stakeholders including museums, federal agencies, academic institutions, and individual researchers. A multitude of institutional catalogs, diverse practices for sample identification, and discipline-specific data and metadata standards exist. Also, many, if not the majority of

DOI: 10.5281/zenodo.4480196

samples remain hidden in labs, offices, and basements, as researchers and institutions do not have the resources and often lack expertise to properly curate their samples and make them FAIR (McNutt et al. 2016).

> *"I was having a specimen-tracking crisis over the weekend, with someone wanting to do carbon 14 analysis on some bones I identified 20 years ago that are stored at UCLA and have all sorts of numbers on the tags, none of which seemed to match my original specimen IDs. So frustrating!"* (S. Kansa, personal communication to R. Walls)

> *"The key measurement was the one backarc basalt called "PPTUW"... Subsequent efforts to confirm the observation ran into problems. The apparently-same sample was variously called PPTU, PPTUW/5, PPTUW-1, and TVZ19 in four other papers. None of those papers gave its latitude and longitude…!"* (J. Gill, personal communication to K. Lehnert)

Increasingly, samples need to be studied by diverse research teams for cross- and inter- disciplinary research areas with critical societal relevance such as sustaining natural resources, controlling the spread of infectious diseases, and coping with environmental change [U7]. For example, geochemists, archaeologists, zooarchaeologists, ecologists, agriculture researchers, soil scientists, epidemiologists, and others all sample soil and derive data from those samples. They curate these samples following practices and protocols in their domain. The data for a soil sample collected to study microbial diversity will likely be stored in the BioSamples database at NCBI [U8], and will thus be difficult for an archaeologist or geologist to find and reuse, while geochemical data for the soil are likely to be stored in the EarthChem Library [U9], difficult to find for the biologist. Alignment of core requirements and discoverability of samples and derived data can provide tremendous benefits for cross-disciplinary use. Geochemical data from ocean water samples collected by a researcher studying global circulation patterns could be used to inform the work of a microbiologist examining viral diversity in the oceans. Pottery samples from an archaeological excavation could be used by a volcanologist studying the impacts of an ancient eruption. An ecologist studying nutrient cycling could use data from rock samples collected by geologists.

Recently, the paradigm of the 'Extended Specimen' (Webster 2017), which includes the physical voucher specimen linked to all digital data derived from it, has gained momentum. The report of a recent NSF workshop (Biodiversity Collections Network 2019) argues for an Extended Specimen Network (ESN), emphasizing that *"Rapid advances in data generation and analysis have transformed understanding of biodiversity collections from singular physical specimens, to dynamic suites of interconnected resources enriched through study over time."* Any science that relies on data derived from material samples would benefit from Extended Specimens. In the Geosciences, the fusion of sample-based observations into databases such as the EarthChem Portal with over 30 million analytical measurements on over 1 million samples has made it possible for the first time to use Machine Learning techniques to explore spatial, temporal, and compositional patterns in large volumes of geochemical and petrological data (e.g., Ueki et al. 2018; Keller & Schoene 2018; Hasterok et al. 2019; Hazen et al. 2019; Liu et al. 2019). The first criterion for establishing an ESN is: *"Develop a robust, comprehensive specimen identifier system in collaboration with other international data aggregators and providers"* (Biodiversity Collections Network 2019).

A widely-adopted, interdisciplinary, flexible system for creating globally unique and persistent identifiers for material samples and indexing metadata about those samples is urgently needed to enable and advance discovery, access, and reuse of the vast volumes of material samples generated and studied in increasingly cross-disciplinary basic and applied scientific research. iSamples will be this system.

iSamples will provide universal services for creating and assigning persistent, unique, and resolvable identifiers to material samples in a consistent manner across disciplines and for registering and indexing metadata using modern and emerging semantic web technologies. The result will be a searchable global index of material samples linked to appropriate metadata and derived data products that will make it possible for researchers to more efficiently and effectively incorporate samples and sample-based data from their own field and from domains outside their field into their research. iSamples will help maximize the return on investment that the U.S. government is making into the collection and curation of samples. It will enable

previously impossible connections between diverse and disparate sample-based observations; support existing research programs and facilities that collect and manage diverse (e.g., biological, geological, hydrological) sample types such as environmental observatories on land and in the oceans; facilitate new cross-disciplinary collaborations; and provide an efficient solution for FAIR samples, avoiding duplicate efforts in different domains.

3. Innovation Through Integration

3.1 Technical Integration: Building on Existing Capabilities.

For a cyberinfrastructure to be successful and adopted broadly, its design needs to build on, embrace, and expand existing, recognized and trusted infrastructure, practices, and cultures. iSamples will provide an innovative, universal solution for FAIR samples across disciplines by taking advantage of existing capabilities for persistent sample identification, identifier and metadata registration, and specimen cataloguing through a multi-disciplinary collaboration of key stakeholders in this field and integrating these to create a new, innovative CI:

- **International Geo Sample Number (IGSN)** [U2] is a persistent unique identifier for Geoscience samples governed by an international non-profit organization with 24 members in 4 continents (Lehnert et al. 2011; Klump et al. 2018; Golodonuic et al. 2017; Conze et al. 2017; Car et al. 2017; Hobern et al. 2018)
- **System for Earth Sample Registration (SESAR)** [U10] provides services for investigators and sample repositories to register samples, manage sample metadata, and obtain IGSNs. It is operated as part of the NSF/GEO-funded data facility IEDA (Interdisciplinary Earth Data Alliance) [U11].
- **CyVerse** cyberinfrastructure for the Life Sciences [U5] is an easy to use platform for complex data analysis, sharing, and publication, especially of big data. It provides support for tool developers and diverse scientific communities.
- **Open Context** [U6] publishes archaeological data from survey, excavation, and collections and laboratory studies. Open Context mints stable URIs for each observed artifact, ecofact, sample, and context. It now totals roughly 1.6 million indexed records of archaeological specimen (Kansa & Kansa 2010, Kansa 2010).
- **EZID and N2T.net** [U12, U13] are services of the California Digital Library (CDL) for issuing, managing, and resolving over 600 kinds of persistent identifiers (e.g., ARK, DOI, IGSN, URN, Taxon, GO, Pubmed, PDB) and related metadata, including mappings across standards. The developer of EZID, now employed at UC Santa Barbara Libraries, has offered to serve as an unpaid advisor on this project (LOC Janee).

Currently, a multitude of identifier systems are applied to material samples, often specific to a single domain, including URNs, HTTP URIs, DOIs, and IGSNs (e.g., Guralnick et al. 2015). Life Science Identifiers (LSID) [U14] have been largely abandoned in favor of a simpler URI based identification. Usage of the IGSN on the other hand has been growing and extending beyond the geosciences, an indication that it offers features and services that fulfill essential requirements for a trustworthy identifier such as persistence, global uniqueness, and international governance. The IGSN is implemented as a distributed architecture with 'Allocating Agents', who act on behalf of a national agency (e.g., US Geological Survey, British Geological Survey, Geoscience Australia), organization (e.g., Australian Research Data Commons, German Geoscience Research Center GFZ), or community (e.g., SESAR), under the overall coordination of the IGSN e.V. - a non-profit organization with currently 24 members in 4 continents [U15]. Similar to agents of the widely-recognized Digital Object Identifiers (DOIs) [U16], the primary requirement of an allocating agent is to maintain the mapping from an IGSN to a web landing page corresponding to each sample. A minimal schema for describing samples registered with IGSN has been developed [U17], but individual allocating agents can choose to supplement the base metadata with additional information. An important part of the IGSN system is the engagement with scholarly publishers, with a goal of making each mention

of an IGSN within a report or paper a hyperlink and creating machine-readable links between data, samples, and publications via inclusion of IGSNs as 'relatedIdentifier' in the metadata of a DOI, a critical feature for creating Extended Specimens. IGSN became an allowable relatedIdentifier in DataCite's metadata kernel v4 [U18] in 2016.

While promoting the use of the IGSN as a trustworthy persistent identifier specifically designed for material samples, iSamples will support other identifiers such as ARK and DOI that are in use for samples, taking an integrative approach to broaden adoption.

3.2 Knowledge Integration: Collaboration of Leading Experts

The proposed project will build on the achievements and experiences of its PIs, who have been leaders in the development and promotion of Cyberinfrastructure (CI) components pertaining to material samples in different science disciplines (biology, geosciences, archaeology), including sample metadata registries and persistent identifiers for specimens, observation ontologies and data models related to the collection and analysis of material samples, and infrastructure for interoperability of sample-based data:

- Lead PI **Lehnert** has led the development, implementation, operation, and evolution of the IGSN and SESAR since 2004, as well as the NSF-funded EarthChem data facility [U19] and the new NASA-funded Astromaterials Data System [U20]. She has served as the elected president of the IGSN e.V. since 2011 and is PI of the IGSN2040 project funded by the Sloan Foundation to develop a strategy for future sustainability and scalability of the IGSN (Lehnert et al. 2018, 2019). She led EarthCube's iSamples Research Coordination Network and co-chairs the Interest Group for "Physical Samples & Collections in the Research Data Ecosystem" in the Research Data Alliance (RDA) [U21].
- PI **Walls** has spent the past four years leading the Data Commons efforts of the CyVerse CI, which provides full life-cycle data management services for big data such as bulk data upload and metadata application, fine-grained data sharing capabilities, and data publication indexed using schema.org [U22]. She is the lead developer of the Biological Collections Ontology (Walls et al. 2018). She has been a board member and chair of the Compliance and Interoperability Group of the Genomics Standards Consortium (GSC) [U23] and convener of several working/interest groups in Biodiversity Information Standards (TDWG) [U28].
- PI **Davies** directs the University of California's Gump South Pacific Research Station, which hosts the Moorea Coral Reef LTER site [U25] and supports field sample collection across the physical, biological, and social sciences in marine and terrestrial environments. As a Senior Fellow at the Berkeley Institute for Data Science, Davies works with Senior Personnel **Deck** and **Meyer** to develop CI in support of biodiversity genomics and whole ecosystem modeling, such as the GEOME platform (Deck et al. 2017) [U26].
- PI **Vieglais** is Director of Development and Operations for DataONE and responsible for the design, development, and operation of DataONE [U27], which currently provides access to over 800,000 datasets. As Senior Scientist at the University of Kansas he has contributed broadly to biodiversity informatics, including development of the Darwin Core [U28], initial operations of GBIF [U29], and several efforts building distributed systems for sharing biodiversity data including Fishnet, Herpnet, ORNIS, MANIS, and VertNet (Stein and Wiezcorek 2004, Constable et al. 2010) [U30-U33].
- Senior Personnel **Kunze** is an Identifier Systems Architect at the California Digital Library, where he invented the decentralized, non-paywalled ARK persistent identifier scheme, the BagIt file packaging format [U34], the Pairtree storage convention [U35], and the N2T resolver [U13]. N2T.net, which is a scheme-agnostic resolver for ARKs, DOIs, URNs, and over 600 other identifier schemes, is a core part of the ARKsInTheOpen.org initiative, which he now leads.
- Consultants **S. Kansa** and **E. Kansa** lead Open Context [U6] and are co-PIs on the Digital Index of North American Archaeology (DINAA) project (Wells et al. 2014) [U34], an NSF-funded gazetteer linking information on more than one million documented archaeological and historical sites.

Together, this team has devised an innovative solution to a challenge that no individual could solve on their own: making material sample data FAIR by providing robust, scalable identifier services for a data

type that is highly heterogeneous and dispersed and has a long history of existing in silos. The team includes implementors with connections to diverse scientific communities (e.g., Open Context, Smithsonian Institution, geoscientists, biological field stations) that can drive early adoption and promote the utility of the iSamples architecture.

3.3 Social Integration: A Prospective Multi-Disciplinary User Community

iSamples will bring together existing user communities in multiple domains who are currently served by the organizations/projects that participate in iSamples as the initial test case implementations to provide improved solutions to the sample PID needs of their users.

- **Archaeologists and zooarchaeologists** currently use Open Context to publish their data, and Open Context has been looking to mint IGSNs for material samples. Open Context already contains 1.6 million items, of which ~400,000 are organic specimens (e.g. bones or bone fragments from zooarchaeologists). Multiple users have committed to be early adopters of the Open Context implementation of iSamples, including a museum collection at the Hearst Museum (LOC Porter), a lab studying ancient and modern plants (LOC Hastorf), a field excavation site at Poggio Civitate in Italy working with both new and legacy data, as well as physical archived objects in the field lab and the local museum (LOC Tuck), and archaeological faunal collections from several field sites in the Middle East with specimens analyzed in a lab at UNLV (LOC Atici).
- **Geoscientists** have been using the IGSN for nearly 15 years and adoption continues to grow in the US and world-wide. Nearly seven million samples have been registered with IGSN so far. Geoscience publishers recommend the use of the IGSN to reference samples in publications so that data, samples, and publications can be properly linked [U37]. SESAR is currently the only Allocating Agent of the IGSN in the US and has a user base of ~ 800 accounts held by individual researchers, large-scale science programs (e.g., Continental Scientific Drilling [U38], Critical Zone Observatories [U39], NEON [U40]); analytical laboratories; and sample repositories and museums (e.g., US Polar Rock Repository [U41], Scripps Institution of Oceanography [U42], Smithsonian Institution). SESAR urgently needs to re-engineer its architecture to scale to the increasing volume and diversity of samples that users want to register. Other organizations in the US, including federal agencies and national labs (e.g., USGS, DOE), want to set up their own IGSN Allocating Agent, and urgently need easy to deploy software for registering samples with the central IGSN registration authority.
- **Biologists** have a long history of assigning museum accession identifiers to samples. More recently, NCBI's BioSample IDs [U8] can be assigned to specimens for which sequence data are published. Unfortunately, neither of these systems has all the characteristics of a good PID (permanent, globally unique, resolvable, etc.), and biologists have called for better solutions (Biodiversity Collections Network 2019, Guralnick et al. 2015, Güntsch et al. 2017). In the absence of a standardized identifier solution, projects like GEOME (Deck et al. 2017) are minting ARKs for specimens at the time of collection, which are then transmitted to BioSamples. Other organizations, such as NEON and the Australian allocating agents CSIRO [U43] and ARDC [U44] are already using IGSN for biological specimens and archaeological artefacts.

We propose initial test implementations of iSamples by Open Context, SESAR, GEOME/CyVerse, and the Smithsonian Institution. Several additional organizations have already expressed their keen interest to adopt iSamples components including NEON in collaboration with the collections management software Symbiota [U45] (LOCs McKay, Franz), the USGS (LOC Powers), the National Microbiome Data Collaborative, which is an effort of the DOE (LOC Wood-Charlson), DOE's Earth System Science DeepDIVE (LOC Agarwal), and PaleoCORE, which integrates data from paleo-anthropology (LOC Reed).

A major driver for all of these organizations to adopt iSamples is that material samples often cross disciplinary bounds and discipline-specific solutions are not acceptable for them. Microbiome samples are collected by biologists and geochemists and contain both biological and mineral material. Archaeological, anthropological, and zooarchaeological sites generate specimens that can be biological, fossil, man made, rock, or soil. Rock, sediment, and soil cores collected by geoscientists often contain fossils or biological

material. Researchers and the organizations that support them are asking for identifier services that work across a range of material sample types, are scalable, and are supported by community metadata standards. iSamples services can be transformational to scientific research, because they will make it possible to realize the benefits of an Extended Specimen Network that not only includes biodiversity specimens (as originally envisioned for ESNs), but integrates any material sample, thus providing an integrated network of samples and sample based data.

5. PROPOSED WORK

Our proposed work includes three main objectives: 1) Design and develop iSamples infrastructure 2) build four initial implementations of iSamples for adoption and use case testing, 3) conduct outreach and community engagement to developers, individual researchers, and international organizations concerned with material samples. The broad architecture and implementation plans described below have emerged from knowledge of preceding systems and extensive discussion between project participants and more broadly. However, experience dictates that actual implementations evolve from initial plans due to many factors influencing the development and operation of production software services. We will therefore follow an agile development process that includes community engagement as an important element of scoping software requirements and implementation timeline. Key milestones will be identified early in the project and will be evaluated and adjusted as necessary to meet the needs of the material samples community. For more details of our development timeline, see the supplemental Management and Coordination Plan.

5.1 iSamples System Design and Development (Objective 1)

iSamples infrastructure will support creation of identifiers for material samples, ensure that identifiers and associated metadata are reliably collated by identifier authorities, facilitate tracking of provenance for material samples, enable discovery of samples, link samples to derived data and other digital content, and report usage of material sample records. There are two key components of the iSamples system (Figure 1):

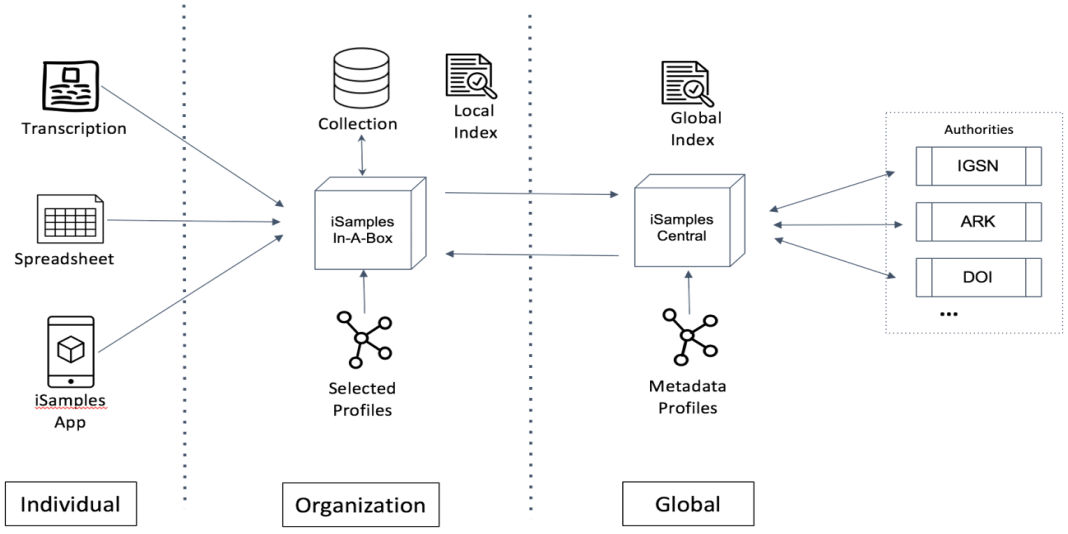

Figure 1 Design of the iSamples system: iSamples-in-a-box provides local services for identifier allocation and metadata collation and iSamples Central for global discovery, resolution and PID coordination.

- **iSamples in a Box (iSB)** is a standalone system that enables creation of identifiers and associated metadata, retrieval of the sample information, updates to the sample metadata (e.g. augmenting or correcting metadata or appending provenance statements), sample identifier resolution, and discovery of samples. iSB will support different scenarios – from installations like SESAR, which has a commitment

to provide a widely used, reliable community service, to installations with a subset of capabilities ('sub-box') delivered to a smaller community of researchers, for example at a field station, to create identifiers and collect metadata that are later synchronized with the organization's iSB.

- **iSamples Central (iSC)** is designed as a permanent Internet service that preserves and indexes sample metadata to ensure reliable discovery and retrieval, provides a gateway between iSB instances and identifier authorities to ensure that remote iSB content is fully synchronized with the relevant authorities (e.g., IGSNs generated on iSB are synchronized with iSC and the IGSN central authority). By providing these services that augment existing identifier authority capabilities, iSC enables support of identifier types such as ARKs or DOIs that are not traditionally associated with material samples, but are already in use by some organizations.

Integration with collection management systems is enabled through the provision of **iSB adapters.** These adapters ensure that the content collated on the iSB (e.g., identifier creation and metadata collection) is seamlessly available to existing collection management systems as part of the sample ingest and curation process. Metadata collection at the iSB can be configured to comply with community defined metadata standards and is coordinated through iSC where the respective standards are promoted and supported for indexing and sample record validation.

5.1.1 iSamples Central (iSC) *(Task 1.1)*

The iSamples Central (iSC) service is a critical component of the infrastructure. Therefore its development will be emphasized during early stages of the project. This component has key roles of providing identifier agnostic resolution services, enabling sample discovery, and supporting allocating agents. It will be implemented as a hosted web service that can be accessed programmatically through RESTful APIs.

Internally, we anticipate that iSC will build upon the identifier resolution capabilities of the N2T service [U14], refining and augmenting those capabilities as needed for the material samples community. Metadata indexing will leverage and adapt the multi-standard, extensible indexing functionality implemented by DataONE [U46] to populate the iSC search index. Identifiers generated by allocating agents will be accumulated using the web-publishing pattern successfully adopted by large content indexers such as Google and Microsoft to retrieve schema.org (Guha et al. 2015) described samples from allocating agents (such as iSB and other compatible services). The iSC service will also expose all content through the same schema.org approach, hence making the content openly available to all interested parties.

The iSC service shall provide a web interface that facilitates resolution, discovery, and display of sample records. The discovery interface will also emphasize relationships between samples, and so facilitate ease of discovery of the complete lineage of a material sample, from its original collection, through subsamples, analyses, and subsequent publication.

Multiple instances of iSC may exist that would mirror each other to help ensure reliable access to the services. These iSC instances would typically be fully synchronized with each other and therefore redundant. Each iSC has a complete record of metadata, which is indexed and made available through a search API and user interface. Additionally, each metadata record can be accessed via its landing page that containings the same schema.org markup describing the resource retrieved from the iSB instances.

5.1.1.1 iSC Architecture Design: Design of the iSC service will be an early emphasis of the project, and will consider not just the programming interfaces, but also the supporting infrastructure and instrumentation necessary to monitor resource use. Deliverables include functional use cases to be supported at different phases of the implementation, high level API specifications supporting the use cases, initial estimates of hardware requirements, expected third party dependencies, and other considerations such as implementation language, revision control and integration testing workflow, and documentation framework. The functional details of the Minimum Viable Product (MVP) will be determined early in the project.

5.1.1.2 iSC API Design & Development: Detailed APIs for iSC will be defined using the OpenAPI specification, which will be integrated into the development workflow such that API changes can be easily incorporated into server and client capabilities. Implementation of iSC will follow best practices and include automated unit and integration testing coupled with the revision control workflow to help ensure a high

quality product. An iterative development process coupled with automation provides rapid feedback to developers and other stakeholders. Deliverables include detailed API documentation, test and eventually production implementation instance of iSC API, and outlines for operational procedures.

5.1.1.3 iSC System Deployment & Administration: Deployment of iSC will occur in at least two environments to isolate development and testing from production systems. The deployments will utilize virtualized servers and be designed to be readily deployable and easy to update. Hardware and other resource utilization will be monitored to provide metrics of actual use that can assist with scaling and cost considerations. Deliverables include documentation of hosting requirements and agreements, administrative and operational procedures, and production deployment of iSC hosted by CyVerse.

5.1.2 iSamples-in-a-Box (iSB) *(Task 1.2)*

A key component of the iSamples infrastructure is "iSamples-in-a-Box" (iSB), which is an offline-first application that supports the creation of sample identifiers and collects metadata associated with the identifier. Interaction with iSB will be through a REST API, and will initially (for the Minimum Viable Product, MVP) include a command line client. An important consideration for iSB is ease of deployment in a variety of scenarios with varying levels of expertise. As such, we anticipate more than one packaging solution will be needed, including at least source and containerized (e.g., Docker [U47]) deployment options.

Technically, the iSB is composed of a web-server service, an instance of CouchDB [U48] or similar data store supporting offline-first functionality with synchronization, a full text search index (e.g., Apache Solr, [U49]), and a browser-based client. When connected to the Internet, an iSB instance can communicate with iSamples Central to provide updates via mechanisms including the schema.org publishing pattern and retrieve additional information to be locally cached. For example, iSB is anticipated to support identifier resolution both directly (identifier is available in the local store) and indirectly through redirection to another resolution service. When disconnected from the Internet, indirect resolution is not possible. Hence, when planning for a field session, relevant identifiers and associated metadata may be cached locally by the iSB to be available for specifying inter-sample relationships while collecting and editing metadata in the field. The set of relevant identifiers is identified by searches against the iSC catalog.

5.1.2.1 iSB Architecture Design: The iSB component is an identifier allocating agent intended to be deployed at many locations, where it can serve in the role of identifier creation and metadata sharing. It interacts with iSC, and so the development of the two will be coupled, though with more emphasis on serving the needs of specific communities. Deployments of iSB are expected to be capable of being integrated with existing collection management systems and so must be flexible and easily deployable. Design of iSB will entail collation of functional use cases, requirements from the intended audiences and consideration of the iSC implementation and respective milestones such as the MVP. Deliverables include documented use cases and requirements, system design documentation, an implementation schedules and guidelines for packaging, installation and configuration.

5.1.2.2 iSB API Design & Development: It will be necessary for iSB to interact with iSC to retrieve configuration and operating parameters, and provide collected sample metadata for collation by iSC. Like iSC, iSB will expose a RESTful API developed using the OpenAPI specifications. We anticipate that iSB will be implemented with the core as a web-server application with command line and web-browser interfaces. This approach enables both desktop installation for individuals or small groups and institutional or broader level installation to serve entire communities, and has proven successful for a number of diverse applications (e.g., the pgAdmin 4 tool for Postgres database administration [U50]). Deliverables include API documentation, test cases, and interaction with iSamples Central, initially fulfilling the MVP requirements and subsequently evolving with additional features and capabilities.

5.1.2.3 iSB User Interfaces (Command Line, basic Graphical UI): User interaction with iSB for the MVP milestone is anticipated to be principally through command-line and library level actions that fulfill the necessary functions. An exemplar web-based user interface will be provided as a later milestone to provide a basis for institutional or local customization to suit the requirements of specific groups. Developers of any extensions or customizations will be strongly encouraged to share their work through open source

contributions, and so continue to build a knowledgeable and engaged technical community to help sustain the product. Deliverables include a command-line client with documentation and an exemplar GUI.

5.1.2.4 iSB Plug-ins for Supported Identifiers (IGSN, ARK): The iSamples infrastructure will principally support IGSN identifiers though it is designed to be identifier agnostic since different identifier schemes may be more applicable for different parts of the research workflow. Identifier support will be abstracted to use a plugin model. Initial implementation will focus on full support of IGSN identifiers and shall provide the ability to register for creation of IGSNs and subsequent creation and management of those identifiers and associated metadata. Later plugin development will support other identifiers (e.g., ARKs) to meet community demand. Deliverables include full support for IGSNs and later development of plugins for additional identifiers including ARKs.

5.1.2.5 iSB Plug-ins for Collection Management Systems: Identifier management must be tightly coupled with sample collection management systems, of which there is great diversity of implementation. Interaction between iSB and a collection management system will be implemented using a plugin system that provides specific adaptations on the core iSB APIs. Initial plugin implementations will support partner collection management systems (i.e. EMu at the Smithsonian Institution, see 5.2.3) while also considering the broader community requirements. Deliverables include detailed specifications for the repository plugin API and plugins for the partner collection management systems.

5.1.2.6 iSB System Deployment & Administration: Operation of iSB must be straightforward and require minimal resources. Good design, development, and implementation practices along with comprehensive documentation can help reduce software maintenance costs. Operators of iSB will be required to register with iSC to enable identifier creation, and the provided contact information will be used to notify operators of any issues or updates that need to be addressed. Operators and developers will be encouraged to participate in open communication mechanisms including issue tracking via GitHub and chat environments such as Slack. Deliverables include administrative and operations documentation for iSB, hardware and associated resource recommendations, and a publicly accessible and up-to-date release schedule.

5.1.2.7 Field (off-line) version of iSB: Some field sample situations may require use of an iSB instance without Internet connectivity (e.g., a remote field site with an intensive sampling activity). For these cases, iSB will be accessible by multiple clients through the hosted web UI (i.e. by connecting with the web server hosted by the iSB instance on the LAN). Roving researchers may continue to collect samples and assign identifiers regardless of connectivity status with the principal iSB instance. Content will be fully synchronized when the roving iSB UI connects with the principal iSB. A roving iSB UI has limited functionality, though it is able to create identifiers and collect metadata input and store that information for later synchronization with the primary iSB. In situations where a roving iSB is not practical (e.g., harsh physical conditions precluding practical operation of a hand held device), basic time and location information may be collected using a GPS enabled digital camera and capturing an image of the sample. The image and associated EXIF metadata can be later imported by iSB. Where samples are collected in the field with handwritten metadata, that information can be transcribed to a spreadsheet and later uploaded to the primary iSB. In this way, the iSB provides a graceful degradation of capabilities while providing the core functionality necessary to assign identifiers to material samples. After returning from field sampling, the iSB instance becomes connected with the Internet. At this time, the iSB connects with the primary iSB and synchronizes its new records and any updates. Deliverables include administrative and operations documentation including examples for common field collection workflows and recommendations for hardware and associated resources to help streamline field data collection.

5.1.3 Connecting iSamples to Community Standards *(Task 1.3)*

The metadata collected by an iSB instance is expected to vary by domain of practice, and each instance of iSB may support multiple metadata schemes used for different types of samples. Metadata schemes shall be determined by community consensus through adoption or modification of existing standards (see section 5.3.3). Despite the considerable diversity of metadata standards, it is anticipated that a core set of elements shall be identified for material samples, and that set shall contain at least the four dimensions describing a

point in space-time (i.e. a 4-D VOXEL) and the identifier (Fig. 2). This core set of properties enables later joining of samples by location and/or time and will enhance discoverability across disciplines. Various efforts are already underway to establish a core metadata profile for the IGSN through organizations such as ESIP [U51] and RDA [U52, U21]. Deliverables include a publicly available mapping of core metadata elements from different communities including at least MIxS (Yilmaz et al. 2011), Darwin Core (Wiezcorek et al. 2011), IGSN metadata, DataCite [U18], and Open Context (Kansa & Kansa 2010, Kansa 2010).

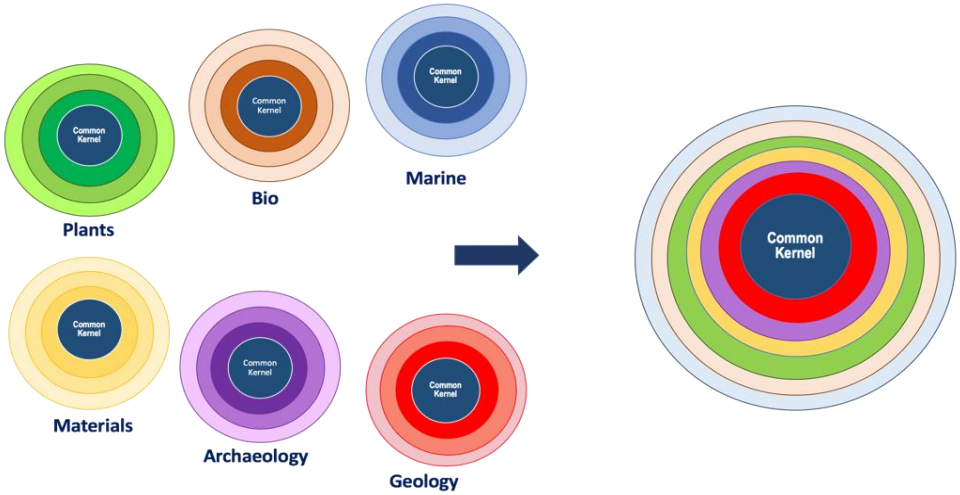

Figure 2 "Bulls-eye" concept of sample metadata: Various domain-specific metadata and ontologies can be wrapped around a common metadata kernel for samples (Credit: Lesley Wyborn)

iSamples is aims to provide a shared core set of metadata fields for material samples across disciplines, however, as a development project, we are somewhat limited in the extent to which we can drive the creation of such a standard. Therefore, some of the PIs of this proposal are also proposing a Research Coordination Network (RCN) that will deliver an interdisciplinary community-driven metadata standard for material samples, with recommendations for its short- and long-term governance. Even though the efforts are independent, we believe that both the social developments of the RCN and the technical developments of this proposal are needed to make material sample data FAIR.

5.2 Implementation and Use Case Testing (Objective 2)

We will create four initial instances of iSB. These implementations will support existing user communities who are waiting for improved access to PID services and allow us to refine iSB based on users' needs. These implementations represent much of the diversity of material samples generated for research, and each case includes the need to identify material samples of different types (organic, mineral, man-made).

5.2.1 Open Context *(Task 2.1)*

Open Context reviews, edits, annotates, publishes and archives research data and digital documentation, with a specialization in the field of archaeology (Kansa & Kansa 2010, Kansa 2010) [U6]. Open Context richly annotates and integrates analyses, maps and media, linking data to the wider world and broadening the impact of the ideas of the data creators. For Open Context to be integrated into iSamples, it would become both an Allocating Agent for IGSN for the field of archaeology and a retrofitter of existing Open-Context records into records of material samples. As an Allocating Agent for IGSN, Open Context will harmonize multiple metadata sources: the lab-specific incoming metadata, the community standards in archaeology, and the standard metadata used by IGSN. Additional developments will include:

Open Context currently has relevant persistent identifier and metadata annotation tools that will be augmented to serve IGSN workflows. Open Context mints ARKs, and will be able to leverage this existing workflow for generating ARK records that conform to the iSamples Archaeology metadata profile and

registering those records to iSamples Central. Open Context will make use of the ARK iSamples plugin. Additional developments will include:

- Open Context will support the IGSN metadata kernel, while adding relevant archaeological metadata. Open Context will work with the community toward archaeological metadata profiles that can be promoted as community best practices.
- This support of IGSN compliant metadata will require crosswalking existing Open Context metadata or producing new metadata.

Open Context currently has relevant persistent identifier and metadata annotation tools that will be augmented to serve IGSN workflows. IGSN will be promoted as the preferred sample identifier to users of Open Context, but options for registration of ARK or other identifiers will remain available for users who need to comply with institutional rules or practices. We will install an Open Context instance of iSB in the same Google Cloud infrastructure we use to host opencontext.org. Opencontext.org (most of which is written in Python) will interact with iSB through a Python wrapper around the iSB REST API.

5.2.2 SESAR *(Task 2.2)*

SESAR [U4] provides services, tools, and APIs for sample metadata cataloguing; metadata management and preservation; IGSN registration; and discovery and access of sample metadata, i.e. for making their samples FAIR. Users can obtain or register user-generated IGSN by filling out metadata forms online (Individual Sample Registration), uploading spreadsheets with sample metadata (Batch Registration), or using web services to submit sample metadata. Users are provided an authenticated workspace that features a dashboard with links to the user's registered samples and tools that support the management of the metadata and samples. SESAR has a well-established workflow for quality control of submitted metadata with scripts and an administrative user interface for data curators, also used to maintain controlled vocabularies. SESAR currently catalogs 4.355 million samples including rocks, sediments, minerals, soils, biological samples, water samples, volcanic gas samples, experiments (synthetic materials), and more.

Participation in iSamples will allow SESAR to provide users with the ability to register diverse sample types with metadata profiles that comply with community standards, e.g., fossil samples that need to follow the Darwin Core metadata standards. It will allow SESAR to integrate its metadata catalog with a much broader sample metadata at the iSC (see section 5.1.3). Using the iSB codebase, which we envision to be maintained by an open source developer community, will make SESAR infrastructure more sustainable and resilient. SESAR tasks will comprise mapping ISGN metadata to iSamples metadata kernel and deployment of an iSB instance for development and testing for IGSN registration, replacing the current direct communication with the IGSN central registry. Expanded services will include the ability to offer users a choice of discipline-specific or sample-type specific metadata profiles. We expect that these developments will attract an even wider array of new users, especially those who so far found it difficult to incorporate SESAR into their existing workflows. New users are likely to include new museums and repositories, who - as we learned by operating SESAR - may prefer to use their own collection management systems to catalog and create virtual representations of their samples.

5.2.3 GEOME *(Task 2.3)*

The Genomic Observatories Meta-Database (GEOME) is a web-based database that captures the who, what, where, and when of biological samples and associated genetic sequences, and represents all sample metadata as material samples (Deck et al. 2017) [U28]. GEOME helps users ensure that biological samples are FAIR, improves the quality of user data and compliance with global standards, and helps integrate with downstream systems including NCBI's sequence read archive and LIMS systems.

GEOME's participation in the iSamples network will enable cross-disciplinary utilization of GEOME metadata as it extends to other collaborators and domains participating in the iSamples project. GEOME has aligned with the Biological Collections Ontology in adopting the definition of "material sample" (Walls et al. 2014), which is able to extend to different domains (e.g., earth science, archaeology). Therefore,

integration with iSamples is a natural fit and extension of current capabilities. Proposed activities for GEOME entail:

- Harmonize GEOME metadata, currently based on Darwin Core and MIxS, with IGSN metadata.
- Replace the current GEOME identifier registration system that communicates directly with the EZID ARK infrastructure, a bespoke codebase, with the iSB codebase, which will be shared among a broader set of collaborators. Initially, GEOME will continue to supply ARKs for material samples, but the adoption of iSamples will allow it to extend or switch to IGSNs in the future.
- Transition identifier metadata already submitted to GEOME to adopt the most recent standards.
- Create a framework for the GEOME network to adopt and register new syntax and semantics.

We will install a GEOME iSB instance as part of the GEOME infrastructure deployed at [U24]. GEOME is written in Java and will interact with iSB through a java wrapper around the iSB REST API. GEOME is currently hosted on iDigBio infrastructure at the University of Florida, and we propose to migrate GEOME to CyVerse late in year 2 and have requested CyVerse server support for years 3 and 4 for server hosting. The migration of GEOME to CyVerse will align GEOME with the CyVerse Data Commons, which supports standardized and custom metadata use in the CyVerse Data Store and publication to the Data Commons Repository. The inability to record metadata about material samples used in CyVerse analyses has been a major limitation of the CyVerse Data Commons which this work will overcome.

5.2.4 Museum Collection: Smithsonian Institution *(Task 2.4)*

A core functionality of iSamples is to engage existing archival repositories (e.g. museums, herbaria, biorepositories) that are inheriting third-party identifiers, minting their own set of identifiers, and creating new material samples (and subsequent identifiers) from existing collections. The Smithsonian Institution Museum of Natural History (NMNH) is a significant partner because it maintains the largest natural history collection in the nation and covers all three domains targeted in this proposal: biology, geosciences, and archaeology/anthropology. The NMNH will provide subject matter expertise on collections-based identifier practices and applications, and technical support for test implementation of iSamples interactivity with the Smithsonian NMNH collections database. PI-Meyer will work with NMNH's Informatics team to support a programmer in years 2-4 to develop a strategy to leverage existing protocol interface options to provide identifier services to iSamples and to support adopted relationships between both third-party and NMNH-generated identifiers. The Smithsonian will also coordinate and host a museum-centric hackathon (Year 2) to engage representatives from permanent archival institutions for an implementation strategy session.

NMNH's collections database currently uses both ARKs and IGSNs to uniquely identify samples. These identifiers can overlap with inherited identifiers from external systems such as Field Information Management Systems (FIMS). NMNH will support an interface that exposes the relationships between these identifiers to the iSamples cyberinfrastructure. Collections/sample repositories such as natural history museums are not "dead ends" to the sample chain. New samples are created from existing samples within these repositories and new identifiers are created in-house to track relationships between them. These practices and relationships will be exposed to the iSamples infrastructure for discovery and use.

5.3 Community engagement and outreach (Objective 3)

Community engagement and outreach activities are essential to drive adoption beyond the communities that will use our initial implementations described in Objective 2, and to ensure interoperability with the international research community. Information on metrics used to assess the impact of these activities is included in the supplemental Delivery Mechanism and Community Usage Metrics document.

5.3.1 Outreach to Developer Communities *(Task 3.1)*

Our goals in engaging the scientific developer community are to encourage additional implementations of iSB by organizations supporting their own users and to enhance the iSamples code base, thereby making iSamples more robust, broadly useful, and sustainable. We have planned two hackathons (year 2 at Columbia and year 3 at Berkeley) for developers, who are looking for a way to provide material sample PIDs to their communities (iSB) or need a way to search for material sample metadata (iSC) as well as a museum-

focused focused workshop at the NMNH in year 3. The following organizations have committed to participating: NEON (LOC McKay), USGS (LOC Powers), DOE's ESS-DIVE (LOC Agarwal), DOE's National Microbiome Data Collaborative (LOC Wood-Charlson), PaleoCore (LOC Reed), USDA (LOC Parr), and Symbiota - also working with NEON (LOC Franz). The expected outcomes include the creation of additional iSB instance, immediate contributions to the iSamples code including new features and hardening, and recruitment of ongoing collaborators to the iSamples code base.

5.3.2 Outreach to researchers generating material samples *(Task 3.2)*

Our initial implementations will support adoption by a large and diverse group of users across archaeology, biology, zooarchaeology, physical anthropology, geosciences, and more. However, there remains an even larger community of researchers, who are either unaware or do not realize the importance of using PIDs for material samples. We plan a series of outreach activities aimed at individual researchers with the goals of educating on the importance of PIDs and comprehensive metadata for material samples, driving adoption of our implementations as well as those built by collaborators (task 3.1), and stimulating interdisciplinary sample-based research. Activities include hosting early career workshops at conferences and user testing meetings hosted at Columbia University, peer reviewed publications, and promotion of iSamples at international meetings like Research Data Alliance (RDA), American Geophysical Union (AGU), Biodiversity Information Standards (TDWG), the Genomics Standards Consortium (GSC) and the International Council for Archaeozoology (ICAZ). At these meetings, we will reach representatives of research communities who can promote iSamples to their individual scientists. The expected outcomes are to have more samples registered with PIDs and appropriate metadata across multiple disciplines and increased usage of iSC to discover samples for reuse in research.

5.3.3 Coordinate with International Standards and Infrastructure Organizations *(Task 3.3)*

A key justification for iSamples is to promote the creation of FAIR samples via the use of PIDs and standardized machine-readable metadata. To do so, iSamples must rely on the international standards community. The iSamples team will continue their interactions with the standards organizations mentioned in task 3.2 (RDA, TDWG, GSC, ICAZ) to ensure that iSamples users have access to the most current standards for material samples. For biological samples that will be sequenced, the National Center for Biotechnology Information (NCBI) maintains the BioSamples Archive. In order to ensure the interoperability between NCBI BioSample and organizations developing platforms to capture sample metadata and that develop standards for sample metadata, BioSample interacts closely with these groups. BioSample has agreed to provide input to the iSamples project as it has in the past with other groups. (LOC Mizrachi). We will also work to align iSamples development with other international efforts such as CSIRO and ARDC in Australia (LOC Klump) and Europe via DISSCo [U53] (LOC Koureas). We are fortunate that organizations in other countries share our vision for FAIR material samples and are developing corresponding infrastructure. The outcomes of this objective are interoperability among international sample identifier and metadata efforts, and compatible use of sample metadata across organizations, countries, and disciplines.

5.4 Administration and project management (Objective 4)

Overall administration and project management for iSamples will be carried out by the lead organization, Lamont Doherty Earth Observatory at Columbia University, under the direction of Lead PI Lehnert. As a collaborative proposal, each additional funded organization (The University of Arizona - Walls, University of Kansas - Vieglais, UC Berkeley - Davies) will be responsible for completing its assigned tasks. Walls and Vieglais will have primary responsibility for core iSamples development (Objective 1). iSamples implementations (Objective 2) and outreach activities (Objective 3) will be managed by the responsible personnel listed in the supplemental Management and Coordination Plan.

6. SUSTAINABILITY PLAN

iSamples is envisioned as a critical step toward a national-level cyberinfrastructure that will support the integration of material samples into modern-day digital and data-driven research, and advance cross-

disciplinary access and study of samples and the mining of observational data resulting from these studies. Anticipated outcomes of the iSamples developments are software products (iSC, iSB) that will be deployed and operated by existing organizations, agencies, and research infrastructure providers. Major scientific programs (e.g., NEON, CZO, IODP [U52]) and national agencies in the US and abroad (e.g., USGS, DOE, British Geological Survey) are interested in collaborating with iSamples because they are eager to implement and use software tools that allow them to readily assign globally unique and persistent, resolvable identifiers to their samples and publish sample metadata in a consistent, machine-readable form across domains. This level of adoption of the iSamples cyberinfrastructure is the key to the project's sustainability.

While it will be the responsibility of iSB adopters to maintain individual deployments, ongoing development and maintenance of the software will be required. Such support will be achieved principally through community contributions, governed by a community agreed process that will likely identify a small group of experts to review and merge proposed contributions. All source contributions will be managed through the openly accessible source code repository environment (e.g., GitHub). iSC will initially be hosted within the CyVerse infrastructure, but will need to find a persistent home with sustained funding. We envision this to happen by either transferring iSC to a research institution with the necessary infrastructure and interest in sample CI; to a national agency such as the Smithsonian Institution, DOE, or USGS; or to a long-term research infrastructure. Several members of the iSamples team have experience with non-profit organizations, and we will also explore this option. Open Context has operated continuously since 2006, financed by a mixture of revenue sources including: data management service and curation charges, fee-for-service software engineering and technical consulting, grants and philanthropic donations. Open Context could incorporate iSamples as part of an expanded set of fee-for-service offerings. The IGSN is a non-profit organization and is currently developing solutions for the long-term sustainability and scalability of its technical and organizational architecture (IGSN 2040) through a grant from the Alfred P. Sloan Foundation.

Meeting outcomes, documentation, source code, project web site, and similar content will be stored in an openly accessible revision control system such as GitHub. Community contributions will be encouraged through outreach activities including but not limited to participation at various national and international meetings of relevance. Community participation and support will be further encouraged through ample documentation and open, responsive administrative workflows that accept or reject "pull requests" with appropriate feedback. The general goal is to encourage participation and attract in-kind support through intellectual and operational investment in the project and its outcomes.

7. Broader Impacts

The relevance of scientific collections for science and society has been emphasized in various studies and reports over the past decades (NRC 2002; IWGSC 2009). In 2005, the White House Office of Science and Technology Policy (OSTP) and Office of Management and Budget (OMB) called on Federal agencies to focus attention on integrated support and planning for the care and use of Federally held scientific collections with the goal of ensuring that this "vital research infrastructure is preserved and strengthened for the benefit of both our country and the global scientific research enterprise" (IWGSC 2009). iSamples delivers enhanced infrastructure for STEM research and education, decision-makers, and the general public. It provides the missing link among physical collections (e.g., Smithsonian), research stations and observatories (e.g., LTER sites, NEON, field stations, marine laboratories), cyberinfrastructure for data (e.g., DataONE, CyVerse), and publications, leveraging significant national investments in each. In the longer term, iSamples technologies can be applied beyond the initial focus on the sciences of natural history to fields such as biomedicine, agriculture, astronomy, and manufacturing. iSamples can benefit national security and resource management by offering a means to adequately track samples, easing the task of assessing their status, permits, and origin (automated provenance of physical objects auditing), including sensitive archaeological or biodiversity specimens, and specimens containing controlled substances. iSamples aims to maximize diversity and inclusiveness. Two of the four collaborating efforts are led by women, an underrepresented group in cyberinfrastructure leadership. Senior personnel at Lamont is 80% female and cul-

turally diverse. The project will train a graduate student and a post-doc in the technical aspects of infrastructure design and development, communication, and outreach, and will host workshops specifically aimed at early career scientists.

8. Results from Prior NSF Support

Lehnert: *Geoinformatics Facility Support: Integrated Data Collections for the Earth & Ocean Sciences: The Marine Geoscience Data System and the Geoinformatics for Geochemistry Program.* Award: OCE-0950477; Period: 12/1/2010-9/30/2017; Award Amount: $12,623,687; PI: Lehnert, Co-PIs: Carbotte, Ferrini, Richard, Ryan, Block). *Intellectual Merit:* The grant supported operation of the Interdisciplinary Earth Data Alliance (IEDA) facility, which maintains data systems and services for the solid earth sciences, including primary collections of global geochemical data, marine geoscience data, and for Earth Science samples (SESAR) that ensure long-term preservation and support the discovery, retrieval, and analysis of these sensor and sample-based analytical data. *Broader Impacts:* IEDA systems provide free and open access to all data holdings for the broad geoscience research and education community. Usage metrics show broad adoption and substantial impact on scientific productivity and new discoveries. Tools for data visualization and analysis (GeoMapApp, Virtual Ocean, EarthChem Portal) support multi-disciplinary research and education. SESAR and the IGSN drive a global movement toward the adoption of persistent identifiers for samples. Products: operational data repositories, products, & services; >40 abstracts at major conferences (AGU, GSA, EGU, ESIP, etc.); peer-reviewed articles (e.g. Hsu et al. 2016, McNutt et al. 2016).

Ramdeen: No prior support from NSF.

Walls: *EAGER. Collaborative Research: Evaluating Identifier Services for the Life Cycle of Biological Data.* Award: EF-1554930; Period: 2015-2017; Award Amount: $65,378. *Intellectual Merit:* This project built a proof of concept portal that solved key challenges in managing large distributed datasets while concurrently conducting research into the functionality of different identifier types and their use by biologists. *Broader Impacts:* Identified major issues in managing large datasets that are not being addressed by current data repositories. Proposed solutions from this project can improve national cyberinfrastructure for managing big data across many disciplines. One graduate student was trained as part of this project (currently working at Google). Products: Five conference abstracts, three peer-reviewed conference papers and one journal article were published (e.g., Esteva et al. 2019, Xu et al. 2016).

Vieglais: *DataONE: Data Observation Network for Earth*, #0830944, 8/1/2009-7/31/2014, $20M, and #1430508, 10/1/2014-3/30/2020, $15M. *Intellectual Merit*: DataONE [U27] developed cyberinfrastructure federating repositories globally and undertook community engagement and education to enable the long-term preservation of complex, cross-scale, multi-disciplinary, and multi-national science data. Currently, 45 data repositories share a common API, making >8000,000 data sets searchable in DataONE [U27], which are accessed hundreds of thousands of times every month [U27]. *Broader Impacts:* DataONE stimulates workforce development through training workshops, development of educational materials, and community needs assessments. Products: operational data services, dozens of software products supporting the DataONE cyberinfrastructure, and 73 peer-reviewed publications (including Michener et al. 2011, Murillo 2014, King 2007, Allard 2012, Michner 2015).

Davies: *Collaborative Research: Biological Science Collections (BiSciCol) Tracker: Towards a tagging and tracking infrastructure for biodiversity science collections* (Berkeley PI Davies, Deck, Smithsonian - Meyer) NSF DEB-0956426 10/01/10-09/30/14, $597,474. *Intellectual Merit:* The BiSciCol Project drew on experience gained through the Moorea Biocode Project (Check 2006) to enable the semantic integration of sample-based databases by employing persistent identifiers and standards-based conceptual frameworks. *Broader Impacts*: BiSciCol has helped bring together the Genomic Standards Consortium (Field et al. 2011) and Biodiversity Informatics (TDWG) [U28] standards organizations through a common Genomic Biodiversity Working Group as part of TDWG, co-chaired by Deck and Walls, with outcomes influencing community standards (Deck et al. 2013). Products: BiSciCol has produced various data products, such as the BiSciCol Triplifier application (Stucky et al. 2014) and the Biological Collections Ontology (Walls et al. 2014).

References

URLs of Resources cited in the Project Description:

[U1] https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp
[U2] http://www.igsn.org/
[U3] https://n2t.net/e/ark_ids.html
[U4] http://www.geosamples.org/
[U5] https://www.cyverse.org/
[U6] https://opencontext.org/
[U7] http://scicoll.org/research.html
[U8] https://www.ncbi.nlm.nih.gov/biosample/
[U9] http://www.earthchem.org/library
[U10] http://www.geosamples.org/
[U11] https://www.iedadata.org/
[U12] https://ezid.cdlib.org/
[U13] https://n2t.net/
[U14] https://en.wikipedia.org/wiki/LSID
[U15] http://igsn.github.io/organisation/
[U16] https://www.doi.org/
[U17] http://igsn.github.io/metadata/
[U18] https://schema.datacite.org/meta/kernel-4.0/
[U19] http://www.earthchem.org
[U20] http://www.astromat.org
[U21] https://www.rd-alliance.org/groups/physical-samples-and-collections-research-data-ecosystem-ig
[U22] https://schema.org/
[U23] http://gensc.org
[U24] https://www.tdwg.org/
[U25] http://mcr.lternet.edu/
[U26] https://geome-db.org/
[U27] https://www.dataone.org/
[U28] https://dwc.tdwg.org/
[U29] https://www.gbif.org/
[U30] http://www.fishnet2.net/aboutFishNet.html
[U31] http://www.herpnet.org/
[U32] http://www.ornisnet.org/
[U33] http://vertnet.org/
[U34] https://tools.ietf.org/id/draft-kunze-bagit-16.html
[U35] https://pythonhosted.org/Pairtree/pairtree.pairtree_store-module.html
[U36] https://alexandriaarchive.org/dinaa/
[U37] https://copdess.org/enabling-fair-data-project/commitment-statement-in-the-earth-space-and- environmental-sciences/
[U38] https://www.icdp-online.org/home/
[U39] http://criticalzone.org/national/
[U40] https://www.neonscience.org/
[U41] https://research.bpcrc.osu.edu/rr/
[U42] https://scripps.ucsd.edu/
[U43] https://www.csiro.au/
[U44] https://ardc.edu.au/
[U45] http://symbiota.org/docs/
[U46] https://indexer-documentation.readthedocs.io/en/latest/
[U47] https://www.docker.com/

[U48] https://en.wikipedia.org/wiki/Apache_CouchDB
[U49] https://lucene.apache.org/solr/
[U50] https://www.pgadmin.org/
[U51] https://www.esipfed.org/
[U52] https://www.rd-alliance.org/
[U53] https://www.dissco.eu/
[U54] https://www.iodp.org/
[U55] https://search.dataone.org/data

Allard S. (2012). "DataONE: Facilitating eScience through collaboration". Journal of eScience Librarianship, 1(1), 3. Biodiversity Collections Network (2019). "Extending U.S. Biodiversity Collections to Promote Research and Education". American Institute of Biological Sciences, Washington, DC, 8 pp. URL: https://www.aibs.org/home/assets/BCoN_March2019_FINAL.pdf

Car N. J., Golodoniuc P., Klump J. (2017). "The Challenge of Ensuring Persistency of Identifier Systems in the World of Ever-Changing Technology." Data Science Journal, 16, 13. http://doi.org/10.5334/dsj-2017-013

Check, E. (2006). "Treasure Island: Pinning down a Model Ecosystem." Nature 439 (7075): 378–79.

Constable H., Guralnick R., Wieczorek J., Spencer C., Townsend Peterson A. (2010). "The VertNet Steering Committee. VertNet: A New Model for Biodiversity Data Sharing." PLoS Biology. https://doi.org/10.1371/journal.pbio.1000309

Conze R, Lorenz H., Ulbricht D., Elger K., Gorgas T. (2017). "Utilizing the International Geo Sample Number Concept in Continental Scientific Drilling During ICDP Expedition COSC-1." Data Science Journal, 16, 2. https://doi.org/10.5334/dsj-2017-002

Deck J., Barker K., Beaman R., Buttigieg P. L., Dröge G., Guralnick R., Miller C., Tuama E., Murrell Z., Parr C., Robbins B., Schigel D., Stucky B., Walls R., Wieczorek J., Morrison N., Wooley J. (2013). "Clarifying Concepts and Terms in Biodiversity Informatics." Standards in Genomic Sciences 8 (2): 352-59.

Deck, John, Michelle R. Gaither, Rodney Ewing, Christopher E. Bird, Neil Davies, Christopher Meyer, Cynthia Riginos, Robert J. Toonen, and Eric D. Crandall. (2017). "The Genomic Observatories Metadatabase (GeOMe): A New Repository for Field and Sampling Event Metadata Associated with Genetic Samples." PLoS Biology 15 (8): e2002925.

Dutton S. P., Goldstein S. L. (2004). "Curation of Terrestrial Scientific Cores, Samples, and Collections." Workshop Report.

Esteva M.A., Walls R.L., McGill A., Xu W., Huang R., Song J., Carson J. (2019). "Identifier Services: Modeling and Implementing Distributed Data Management to Cyberinfrastructure." Data and Information Management 3:26–39. https://doi.org/10.2478/dim-2019-0002

Field D., Amaral-Zettler L., Cochrane G., Cole J. R., Dawyndt P., Garrity G. M., Gilbert J., Glöckner F. O., Hirschman L., Karsch-Mizrachi I., Klenk H.-P., Knight R., Kottmann R., Kyrpides N., Meyer F., San Gil I., Sansone S.-A., Schriml L. M., Sterk P., Tatusova T., Ussery D. W., White O., Wooley J. (2011). "The Genomic Standards Consortium." PLoS Biology 9 (6): e1001088. Golodoniuc P., Car N., Klump J. (2017). "Distributed Persistent Identifiers System Design." Data Science Journal, 16, 34. DOI: http://doi.org/10.5334/dsj-2017-034

Güntsch A., Hyam R., Hagedorn G., Chagnoux S., Röpert D., Casino A., Droege G., Glöckler F., Gödderz K., Groom Q., Hoffmann J., Holleman A., Kempa M., Koivula H., Marhold K., Nicolson N., Smith V. S., Triebel D. (2017). "Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects." Database 2017:bax003.. https://doi.org/10.1093/database/bax003

Guha R. V., Brickley D. MacBeth S. (2015). "Schema. org: Evolution of structured data on the web." Queue, 13(9), 10. 1 https://doi.org/0.1145/2857274.2857276

Guralnick R.P., Cellinese N., Deck J., Pyle R. L., Kunze J., Penev L., Walls R. L., Hagedorn G. Agosti D., Wieczorek J., Catapano T., Page R.D. (2015). "Community next steps for making globally unique identifiers work for biocollections data." Zookeys (494):133-54. https://doi.org/10.3897/zookeys.494.9352

Hasterok D., Gard M., Cox G., Hand M. (2019). "A 4 Ga record of granitic heat production: Implications for geodynamic evolution and crustal composition of the early Earth." Precambrian Research, 331. https://doi.org/10.1016/j.precamres.2019.105375

Hazen R., Downs R., Eleish A., Fox P., Gagne O., Golden J., Grew, E., Hummer D., Hystad G., Krivovichev V., Li C., Ma X., Morrison S., Pan F., Piers A., Prabhu A., Ralph J., Runyon S., Zhong H. (2019). "Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization." Engineering, 5(3), 397-405. https://doi.org/10.1016/j.eng.2019.03.006

Hobern D., Hahn A., Robertson T. (2018). "Options to Apply the IGSN Model to Biodiversity Data." Biodiversity Information Science and Standards, 2: e27087. https://doi.org/10.3897/biss.2.27087

Hsu L., Mayorga E., Horsburgh J.S., Carter M., Lehnert K., Brantley S. (2016), "Enhancing Interoperability and Capabilities of Earth Science Data using the Observations Data Model 2 (ODM2)." Data Science Journal, 16, p.4. DOI: http://doi.org/10.5334/dsj-2017-004

IWGSC: Interagency Working Group on SCientific Collections (2009). "Scientific Collections: Mission-Critical Infrastructure for Federal Science Agencies A Report of the Interagency Working Group on Scientific Collections.". https://obamawhitehouse.archives.gov/sites/default/files/sci-collections-report-2009-rev2.pdf

Kansa E. C. (2010). "Open Context in Context: Cyberinfrastructure and Distributed Approaches to Publish and Preserve Archaeological Data." SAA Archaeological Record, 10(5):12-16.

Kansa E. C., Whitcher Kansa, S. (2010): "Publishing Data in Open Context: Methods and Perspectives." Center for the Study of Architecture Newsletter. Vol. XXIII, No. 2 (September 2010). Bryn Mawr: Center for the Study of Architecture. Keller B.C., Schoene B. (2018). "Plate tectonics and continental basaltic geochemistry throughout Earth history." Earth and Planetary Science Letters, 481, 290-304. https://doi.org/10.1016/j.epsl.2017.10.031

King G. (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." Sociological Methods & Research, 36(2), 173–199. https://doi.org/10.1177/0049124107306660

Klump J., Wyborn L., Lehnert K. (2018). "IGSN - Status and Future Development." EGU General Assembly Conference Abstracts, Geophysical Research Abstracts 20, EGU2018-4076,

Lehnert K., Klump J. (2012). "The Geoscience Internet of Things." EGU General Assembly Conference Abstracts, Geophysical Research Abstracts 14, EGU2012-13370 Lehnert K. A., Klump J., Arko R. A., Bristol S., Buczkowski B., Chan C., Chan S., Conze R., Cox S. J., Habermann T., Hangsterfer A., Hsu L., Milan A., Miller S. P., Noren A. J., Richard S. M., Valentine D. W., Whitenack T., Wyborn L. A., Zaslavsky I. (2011). "IGSN e.V.: Registration and Identification Services for Physical Samples in the Digital Universe." American Geophysical Union, Fall Meeting 2011, abstract id.IN13B-1324

Liu H., Sun W-D., Deng J. (2019). "Statistical analysis on secular records of igneous geochemistry: Implication for the early Archean plate tectonics." Geological Journal. doi: 10.1002/gj.3484 McNutt M., Lehnert K. A., Hanson B., Nosek B., Ellison A. M., King J. L. (2016). "Liberating field science samples and data." Science, 351, 1024-1026.

Michener W., Vieglais D., Vision T., Kunze J., Cruse P., Janée G. (2011). "Dataone: Data observation network for earth-preserving data and enabling innovation in the biological and environmental sciences." D-Lib Magazine, 17(1/2), 3. Michener W. K. (2015). "Ecological data sharing." Ecological Informatics, 29, 33-44.

Murillo A. P. (2014). “Examining data sharing and data reuse in the DataONE environment.” Proceedings of the American Society for Information Science and Technology, 51(1), 1-5. National Research Council (U.S.) (2002). “Geoscience Data and Collections — National Resources in Peril.” Washington, D.C.: National Academies Press.

Stein B. R., Wieczorek J. R. (2004). “Mammals of the World: MaNIS as an example of data integration in a distributed network environment.” Biodiversity Informatics:1. https://doi.org/10.17161/bi.v1i0.7

Stucky B. J., Deck J., Conlin T., Ziemba L., Cellinese N., Guralnick R. (2014). “The BiSciCol Triplifier: Bringing Biodiversity Data to the Semantic Web.” BMC Bioinformatics, 15 (July): 257. Ueki K., Hino H., Kuwatani T. (2018). “Geochemical discrimination and characteristics of magmatic tectonic settings: a machine learning-based approach.” Geochemistry, Geophysics, Geosystems, 19(4):1327-1347. https://doi.org/10.1029/2017GC007401

Walls R. L., Deck J., Guralnick R., Baskauf S., Beaman R., Blum S., Bowers S., Buttigieg P. L., Davies N., Endresen D., Gandolfo M. A., Hanner R., Janning A., Krishtalka L., Matsunaga A., Midford P., Morrison N., Tuama E. O., Schildhauer M., Smith B., Stucky B. J., Thomer A., Wieczorek J., Whitacre J., Wooley J. (2014). “Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies.” PloS One 9 (3): e89606.

Walls R.L., Buttigieg P.L., Deck J., Guralnick R.P., Wieczorek J. (2018). “Integrating and Managing Biodiversity Data with the Biocollections Ontology.” in Application of Semantic Technologies in Biodiversity Science, Anne Thessen, editor. IOS Press. ISBN978-1-61499-853-2 (print) | 978-1-61499-854-9 (online).

Webster M. S. (2017). “The extended specimen: emerging frontiers in collections-based ornithological research.” CRC Press.

Wells J. J., Kansa E. C., Kansa S. W., Yerka S. J., Anderson D. G., Bissett T. G., Myers K. N., DeMuth R. C. (2014). “Web-based discovery and integration of archaeological historic properties inventory data: The Digital Index of North American Archaeology (DINAA).” Literary and Linguist Computing fqu028. https://doi.org/10.1093/llc/fqu028

Wilkinson, Mark D., Dumontier, Michel, Aalbersberg, IJsbrand Jan, Appleton, Gabrielle, Axton, Myles, Baak, Arie, Blomberg, Niklas, Boiten, Jan-Willem, da Silva Santos, Luiz Bonino, Bourne, Philip E., Bouwman, Jildau, Brookes, Anthony J., Clark, Tim, Crosas, Mercè, Dillo, Ingrid, Dumon, Olivier, Edmunds, Scott, Evelo, Chris T., Finkers, Richard, Gonzalez-Beltran, Alejandra, Gray, Alasdair J.G., Groth, Paul, Goble, Carole, Grethe, Jeffrey S., Heringa, Jaap, ’t Hoen, Peter A.C, Hooft, Rob, Kuhn, Tobias, Kok, Ruben, Kok, Joost, Lusher, Scott J., Martone, Maryann E., Mons, Albert, Packer, Abel L., Persson, Bengt, Rocca-Serra, Philippe, Roos, Marco, van Schaik, Rene, Sansone, Susanna-Assunta, Schultes, Erik, Sengstag, Thierry, Slater, Ted, Strawn, George, Swertz, Morris A., Thompson, Mark, van der Lei, Johan, van Mulligen, Erik, Velterop, Jan, Waagmeester, Andra, Wittenburg, Peter, Wolstencroft, Katherine, Zhao, Jun, Mons, Barend. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” Scientific Data 3:160018, https://doi.org/10.1038/sdata.2016.18

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. (2012) “Darwin Core: An Evolving Community-Developed Biodiversity Data Standard.” PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715

Xu, Weijia, Ruizhu Huang, Maria Esteva, Jawon Song, Ramona Walls, (2016) “Content-based Comparison for Collections Identification.” Proceedings of the 2016 IEEE International Conference on Big Data, December 5-8, Washington DC. https://doi.org/10.1109/BigData.2016.7840987

Yilmaz P., Kottmann R., Field D., Knight R., Cole J. R, Amaral-Zettler L., Gilbert J. A, Karsch-Mizrachi I., Johnston A., Cochrane G., Vaughan R., Hunter C., Park J., Morrison N., Rocca-Serra P., Sterk P., Arumugam M., Bailey M., Baumgartner L., Birren B. W., Blaser M. J., Bonazzi V., Booth T., Bork P., Bushman F. D, Buttigieg P. L., Chain P. S. G. Charlson E., Costello E. K., Huot-Creasy H., Dawyndt P., DeSantis T., Fierer N., Fuhrman J. A., Gallery R. E, Gevers D., Gibbs R. A., Gil I. S., Gonzalez A., Gordon J. I.,

Guralnick R., Hankeln W., Highlander S., Hugenholtz P., Jansson J., Kau A. L, Kelley S. T., Kennedy J., Knights D., Koren O., Kuczynski J., Kyrpides N., Larsen R., Lauber C. L., Legg T., Ley R. E., Lozupone C. A., Ludwig W., Lyons D., Maguire E., Methé B. A., Meyer F., Muegge B., Nakielny S., Nelson K. E., Nemergut Di., Neufeld J. D., Newbold L. K., Oliver A. E., Pace N. R., Palanisamy G., Peplies J., Petrosino J., Proctor L., Pruesse E., Quast C., Raes J., Ratnasingham S., Ravel J., Relman D. A., Assunta-Sansone S., Schloss P. D., Schriml L., Sinha R., Smith M. I. Sodergren E., Spor A., Stombaugh J., Tiedje J. M, Ward D. V., Weinstock G. M., Wendel D., White O., Whiteley A., Wilke A., Wortman J. R., Yatsunenko T., Glöckner F. O. (2011). "Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications." Nature Biotechnology 29(5): 415-420. https://doi.org/10.1038/nbt.1823