Smithsonian Global Unique Identifiers (GUID) Working Group (WG) Four Year Report

GUID Working Group

August 2022

Executive Summary

The Smithsonian Global Unique Identifiers Working Group (GUID WG) was established in 2018 with the goal of bringing together staff and units working with a variety of GUIDs. The group was formed to provide education for units in the use of the most appropriate GUIDs, to promote the use of GUIDs across a variety of Smithsonian collections and services, and to ensure Smithsonian assets were discoverable in the networked environment.

Over the course of the past four years, the group has furthered these goals and created a lightweight consultative framework that relies on best practice, cross-unit communication, and established best practices for a variety of GUIDs.

Looking forward, the group will continue working and is in the process of establishing a charter for the next round of GUID work. See the Next Steps section of this report for the direction of the new charter.

Identifier Implementation at the Smithsonian: Looking Back Looking Forward

The Smithsonian Global Unique Identifiers (GUID) Working Group (WG) was formed in March 2018 with representatives from across the Institution to define how an enterprise GUID service might be implemented at the Smithsonian.

The use of unique identifiers is mandatory to expose Smithsonian information as linked data, an aspiration of the 2018-2022 Smithsonian Strategic Plan.The rationale for the formation of the GUID WG was the need for centralized service, streamlining efforts, avoiding multiple subscriptions for the same services, ensuring standardized naming, and providing guidance and best practices for when to use which type of identifier.

The work is critical in positioning Smithsonian to participate in/comply with the global Open Science Initiative. The Smithsonian is also able to explore GLAM (galleries, libraries, archives, museums) linked data initiatives, which utilize GUIDs, to integrate content/collections with external parties including Wikidata and others.

Participation has included representatives from the Biodiversity Heritage Library, Digital Asset Management System, Digital Program Office, Library and Archives System Support Branch, National Museum of American History, National Museum of Natural History, Office of the Chief Information Officer, Provost Office, Research Computing, Smithsonian American Art Museum,and Smithsonian Libraries and Archives. The original 2018 charter for this group is attached at the end of this report as Appendix III.

Identifiers Implemented

The GUID WG's initial focus was on implementing five identifiers. More detail on each is provided below. For the comprehensive list of identifiers in use at the Smithsonian see Appendix I.

- **Digital Object Identifier (DOI).** DOIs, for documents, texts, publications, datasets, and scientific instruments, managed by Smithsonian Libraries and Archives (CrossRef DOI) and Office of the Chief Information Officer (Datacite);.
- **Archival Resource Key (ARK).** ARKs, for collection items including archival, managed by SLA;
- **Open Researcher and Contributor Identifier (ORCID).** ORCIDs, for currently active researchers, managed by SLA;
- **Research Organization Registry (ROR).** RORs, for Smithsonian research facilities, managed by SLA.

Accomplishments

Some of the accomplishments of this group achieved during the initial years:

- Research and inventory of identifiers actively used across the Institution.
- Discussions with various national/international organizations, federal departments, and other GLAM organizations on registration systems and adoption of GUIDs.
- Documenting the structure, use cases, and implications of using various identifiers.
- Consensus on GUIDs and registration agencies for documents (text), datasets, scientific instruments, research facilities, collection items, and living people.
- Centralizing membership to Crossref, DataCite, ORCID, and ARK, managed by SLA.
- Building successful working relationships with representatives to the various registration agencies (Crossref, DataCite, ORCID, EZID, ROR), including membership voting and attending monthly and annual meetings.

- Implementation of ARKs for collections, archival collections, and multimedia items, in diverse systems, and in multiple units across the Institution.
- Implementation of DataCite DOIs for datasets stored on the OCIO-licensed Figshare platform.
- Implementation of DataCite DOIs for Smithsonian research instruments.
- Implementation of ROR IDs for Smithsonian research facilities.
- Successful implementation of interaction of active researchers in the SI Profiles system with the ORCID registration.
- Establishing and maintaining a public facing website of SI Global Unique Identifiers (GUIDs) https://library.si.edu/research/guids-help-make-your-data-findable, providing information and guidance on GUIDs at SI.

Collection Objects and the use of ARKs

The Smithsonian Open Access (OA) Initiative launch deadline of 2020 was used as the goal post for bringing on the SI wide use of ARK identifiers for collection systems. Initially used by NMNH, the group discussed, reviewed, and determined that the ARK system would be the most beneficial for all units to use within their collection information systems (CIS). The Smithsonian worked with EZID to implement the concept of “shoulders”, assigned to each CIS, which are associated with the SI ARK (or NAAN). Up to the SI implementation, this configuration was rarely used for ARKs. Logistics of shifting ARK management with the EZID registration agency from NMNH to a central process managed by SLA was accomplished. All units participating in the OA launch made adjustments to their CIS to generate and store ARKs for their metadata and images. Over the next year more units’ CIS’s were configured to generate and store ARKs. ARKS are now implemented in the SI Digital Asset Management System (DAMs), The Museum System (TMS), EMu, MimsyXG, and ArchivesSpace (SOVA). ARKs are now ingested into the Smithsonian’s Enterprise Digital Access Network (EDAN) central metadata storage and delivery system. The consistent use of ARKs for collection metadata and images has made the discovery and navigation of SI open access collections a streamlined and integrated success. See Appendix II for the inventory of current units and the CISs that have been modified to automatically assign and store ARKs.

Documents and Datasets

The Smithsonian registers Digital Object Identifiers (DOIs) with two registry agencies: Crossref (for documents, texts, publications, and fund IDs), and DataCite (for datasets and scientific instruments). Each registration agency was chosen because their schema allows for the metadata needed to describe the data type (i.e. text or dataset) requesting a DOI.

The SI DataCite account includes the Figshare repository, and The Chandra Data Archive. SLA has granted management of the DataCite Figshare repository to OCIO. SI FigShare registers DOIs for SI research datasets and scientific instruments (The Smithsonian Institution research repository - Browse (figshare.com)). SLA has granted management of the DataCite Chandra

Data Archive repository to The Smithsonian Astrophysical Observatory (SAO). As of August 1st, 2022, SAO has registered 26,899 DOIs for datasets which are stored on the Chandra Data Archive (CDA - Home (harvard.edu).

SLA manages and registers DOIs with Crossref for SI documents, texts, publications, and fund IDs. Most texts requesting an SI DOI are stored in the SLA managed Smithsonian Research Online DSpace Repository (DSpace Home (si.edu). Some SI DOIs are stored on other SI managed systems, such as the Bulletin of Global Volcanism (Global Volcanism Program | Bulletin of the Global Volcanism Network (si.edu)). The Smithsonian Press (SISP), with the support of SLA, uses the SI FigShare system as their CIS (The Smithsonian Institution research repository - Browse (figshare.com). The SISP FigShare installation has been configured to register DOIs with Crossref.

Administration

This Working Group evaluates standards and application feasibility of identifiers. The group can provide best practices and suggestions on use and implementation. The details of managing the registration of identifiers falls on SI units. Smithsonian Libraries and Archives have been able to support the Institution and work with other units to provide the required services needed to date.

The SLA manages and pays (with support from other SI units) for Institutional membership to CrossRef (as well as an Institutional membership for the Biodiversity Heritage Library). SLA staff are voting representatives for SI at the annual meetings, and attend monthly and annual meetings.

The SLA manages an institutional membership to EZID, and manages payment for the membership using funds transferred from participating SI units.

The SLA manages an institutional membership to ORCID.

The SLA administers the Institution's membership to DataCite. SLA manages payment for the membership using funds from SAO and OCIO, who participate in a cost-share agreement. SLA manages ownership of the SI repositories in DataCite, and SLA staff attend monthly and annual meetings and participate in yearly membership voting.

SLA and OCIO collect and publish metrics on SI GUIDs.

- For Crossref DOIs: DOI Resolutions Dashboard (sharepoint.com)
- CNRI Resolutions for SLA (Updated): Crossref Summary - Tableau Server (si.edu)
- For DataCite DOIs registered through Figshare: The Smithsonian Institution research repository - Stats (figshare.com)

Researchers (Living People)

Researchers at SI have used a variety of systems for self-registration. Discipline specific fields have favorite systems and services for sharing information within a field of study. Most of the systems reviewed by the GUID WG were not considered capable of meeting the minimum guidelines of a persistent identifier for the breadth of research conducted at SI. The SI GUID WG decided on ORCIDs as the best for creating a sustainable and interoperable research identifier for our current researchers. This conforms to the OSTP Subcommittee on Open Science standards of persistent person identifiers (see Information from the White House, below). SLA manages the SI ORCID membership, and is registered as a trusted institution. This allows for the SI researchers with ORCID IDs to connect to the SLA managed Scholarly Profile system (which runs on VIVO). ORCID IDs allow for connections between the Scholarly Profile system, Scholarly Research Information Management system (Smithsonian Research Online) and ORCID by pulling and pushing information between the different systems.

SLA collects metrics on SI ORCIDs. As of April, 2022, there are 2,119 SI ORCIDs, and 247 of these have authenticated their ORCID via the SI Scholarly Profile system. There are 4,212 publications SI has added to ORCID via authenticated SI ORCIDs.

Organizations, Facilities, and Instruments

The SI research community requested GUIDs for organizations, facilities, and instruments. After investigating, the Working Group decided on the Research Organization Registry (https://ror.org/01pp8nd67), and SI now has many of its units assigned RORs. SLA recently received RORs for a group of eight STRI facilities, including Barro Colorado Island (https://ror.org/02f4ya153), and Bocas del Toro Research Station (https://ror.org/02ftc8763).

For research instruments, the GUID WG chose to use DataCite DOIs. Both the OCIO High Performance Computing Cluster (HPCC, https://doi.org/doi.org/10.25572/SIHPC) and the STRI canopy cranes (https://doi.org/10.25572/SIHPC, https://doi.org/10.25572/W3BX-M202, https://doi.org/10.25572/269B-SD50) have been assigned DOIs. To date, the HPCC has been cited by this identifier multiple times in the published literature. With the use of this unique identifier, a list of publications can be generated and alerts can be created. Please see Research Outputs from the SI High Performance Computing Cluster | Research Computing to view the publications. With the consistent use of SI's identifier, a Google Scholar alert can be set up on the DOI to track its use in citations.

Additional GUIDs

There is a growing need for a GUID to identify the Smithsonian's historic, or legacy, authors. The GUID WG has been investigating different identifiers to fill this need, including the Virtual International Authority File (VIAF), and Wikidata (a project from Wikimedia). VIAF combines

international library authority files but is not a system to create new identities. The Smithsonian is currently engaged in several data sharing projects with Wikidata, and this may be a logical progression. The GUID WG is investigating both of these.

Information from The White House

Members of the SI GUID WG have participated in subcommittees of The White House Office of Science and Technology Policy (OSTP); specifically, the Subcommittee on Open Science (SOS) Working Group on Persistent Identifiers (PIDs). The PID Working Group developed specific guidance on the use of Persistent Identifiers. With the involvement of SI GUID members, the best practices were shared with fellow federal agencies and ensured that SI was adopting the recommended best practices.

Next Steps

Next steps for the GUID WG include

- Review and update of the GUID Working Group Charter (2022-2026)
- Continue to advocate for best practices and needs of GUIDs at the Institution
- Review of membership
- Continue to onboard units with ARK registration
- Draft best practices and methods for workflow in assigning, transferring, or tombstoning identifiers used
- Explore the use of Wikidata identifiers already established and possible integration
- Explore the use of inflections with ARKs for specific results in resolving to collection item representation
- Continue to monitor the needs for types and formats of identifiers for the Institution
- Collaborate with other units on the needs for standard GUIDs
- Work with researchers and the public on proper citation of SI, and SI outputs, with appropriate GUIDs
- Champion the use of GUIDs in the systems and services that the Institution provides internally and externally
- Plan for future successful continuation of GUID Working Group with a sustainability plan

The GUID Working Group participants have included:

- Alberto Accomazzi (SAO)
- Jon Blundell (OCIO-DPO)
- Alicia Cutler (NMAH)
- Raffaele D'Abrusco (SAO)
- Torsten Dikow (NMNH)

- Riccardo Ferrante (SLA)
- Carolyn Stern Grant (SAO)
- Andrew Gunther (OCIO)
- Alvin Hutchinson (SLA)
- Martin Kalfatovic (SLA) - Co-Chair
- Nancy Kennedy (OCIO-LASSB)
- Amy Marino (USSR)
- Isabel Meyer (OCIO-DAMS)
- Ginger Minkiewicz (SISP)
- Bess Missell (SLA-SIL)
- Edward L Monk (OCIO)
- DucPhong (Ducky) Nguyen (OCIO/NMNH)
- Thomas Orrell (NMNH) - Co-Chair
- Suzanne Pilsk (SLA)
- Vince Rossi (OCIO-DPO)
- Arnold Rots (SAO)
- Jackie Shieh (SLA)
- Sara Snyder (SAAM)
- Adam Soroka (OCIO-RC)
- Beth Stern (OCIO-RC)
- Rebecca Snyder (NMNH)
- Keri Thompson (OCIO-RC)

Identified Stakeholders (ongoing list)

- Deron Burba (OCIO)
- Monique Chilsom (USEducation)
- Ron Cortez (USAdmin)
- Tamar Evengelista-Doughtery (SLA)
- Kevin Gover (USMC)
- Becky Kobberod (DSCOO-HDT)
- Scott Miller (USSR)
- Ginger Minkiewicz (SISP)
- Rebecca Snyder (NMNH)
- Beth Stern (OCIO-RC)
- Ellen Stofan (USSR)
- Bill Tompkins (NCP)
- Diane Zorich (OCIO-DPO)

Appendix I: Identifiers in use across the Institution

Name	Preferred Registration Agency	Types of Use	Units Using	Contact
DOI	CrossRef	Text, SI Fund Codes	SISP SLA	SLA
DOI	DataCite	Datasets Astronomical Datasets Instruments Multi-instrument services	SISP OCIO-RC SAO	OCIO
ORCID	ORCID	Current SI Researchers	SLA	SLA
VNums	SI Global Volcanism Program	Volcano numbers and primary names. Secondary numbers for sub-features	SI Global Volcanism Program (NMNH)	NMNH-Global Volcanism Program
STNs	Smithsonian Trinomial System	Archeological digs site numbers in conjunction with state level archeologists	NMNH (Anthropology) NMAI	
ROR	ROR: Research Organization Registry	Research Facilities Units/Affiliations	SLA, SISP	SLA
ARKS	EZID	Collections objects (metadata and media)	See SI_ARK_DatasetIDs.xlsx	SLA

Wikidata Q	Wikidata	Testing of organizations affiliated with SI	SLA, NMNH	SLA
AphiaID		World Register of Marine Species	NMNH	
VIAF	OCLC.org	Authority lists traditionally from international libraries	SLA	SLA
IGSN		Physical samples	NMNH	NMNH
ISSN/ISBN	ISSN/Bowker	International Standard Numbers for print and electronic text	SLA	
LSID/ZooBank LSID		Life Science Identifier	NMNH	NMNH

Appendix II: ARK Inventory of current units and the CIS

Unit	CIS Software	Unit	CIS Software	Unit	CIS Software	Unit	CIS Software
AAA	ArchivesSpace	HMSG	TMS	NMAH Archives Center	ArchivesSpace	NPM	TMS
ACM Archives	ArchivesSpace	HSFA	ArchivesSpace	NMAH-AF	ArchivesSpace	NZP	DAMS
ACM	TMS	NAA	ArchivesSpace	NMAH	MimsyXG	SAAM Archives	ArchivesSpace
CFCH Archives	ArchivesSpace	NASM Archives	ArchivesSpace	NMAI Archives	ArchivesSpace	SAAM	TMS
CFCH	TMS	NASM	TMS	NMAI	EMu	SG (AAG)	ArchivesSpace
CHNDM	TMS	NMAAHC Archives	ArchivesSpace	NMNH	EMu	SG	TMS
EEPA	ArchivesSpace	NMAAHC	TMS	NPG Archives	ArchivesSpace	SGPE	IrisBG
F|S	TMS	NMAFA	TMS	NPG	TMS	SLA	ArchivesSpace

Appendix III: Original Charter

Original Charter

May 2018

Background

Globally Unique Identifiers (GUIDs) are used by multiple units within the Institution to identify publications, datasets and items in the collections. The three key characteristics of a GUID are: 1) that it is an alphanumeric string that uniquely identifies an object/publication, 2) that it must be resolvable, so that once clicked, the GUID is a link to the identified object online, and 3) that the link is persistent, meaning it will resolve even if the host website changes. There are several types of GUIDs. Three currently used at SI are DOIs, ARKs and ORCIDs. Smithsonian Libraries issues DOIs for publications authored by Smithsonian staff or published by the SI Scholarly Press and SAO uses DOIs for astronomical datasets produced by the Chandra X-ray Observatory. The ADS at SAO issues CrossRef DOIs through SIL. NMNH uses ARKs for items in their collections, and soon for their collections images and multimedia. Smithsonian Libraries has joined ORCID to assist researchers and others at the Institution to establish a personal identifier. SAO is preparing to use ORCIDs to unambiguously identify authors and investigators.

The SI GUID working group has formed with representatives from around the Institution to define how an enterprise GUID service might be implemented at SI. This centralized service will streamline efforts, ensure standardized naming, and provide guidance and best practices for when to use which type of ID. Additionally, the use of unique identifiers is mandatory to expose Smithsonian information as linked data, an aspiration of the 2018-2022 SI Strategic Plan. Many information organizations have also begun using linked data to integrate their content/collections with external parties including the British Museum, Wikipedia, SAO, and others.

Purpose

The SI GUID working group is a self-appointed group working to understand the steps necessary to implement pan-Institutional GUID services at SI. As part of this process, the working group has identified the existing services in use, identified potential use cases for GUIDs at the Institution, and consulted with the California Digital Library, the Department of Energy (OSTI), and other research organizations to understand the services they provide. The objective of the working group is to produce a best practices

recommendation and identify a governance structure to manage the roles and responsibilities for implementing and maintaining the enterprise GUID services.

Subcommittee Activities

1. The working group is conducting the following activities:
    a. Identify GUIDs in use at the Institution
    b. CrossRef DOIs for publications administered by Smithsonian Libraries
    c. CrossRef DOIs for publications administered by ADS at SAO
    d. DataCite DOIs for datasets administered by Smithsonian Libraries
    e. DataCite DOIs for SAO datasets administered by SAO
    f. ARKs for collection items administered by NMNH
    g. ORCID for collecting personal names
2. Develop SI GUID governance structure
    a. Identify pan-Institutional members
    b. Identify group Chair (ideally should be a user/consumer of GUIDs)
    c. Identify responsibilities of members
3. Develop Guidelines for use
    a. Document best practices for when to use one type of GUID vs another
    b. Guidelines for naming authorities
4. Develop GUID Policy
    a. Develop policy for when a GUID is needed
    b. Develop cost sharing policy
5. Develop GUID Service
    a. Service design
    b. Service technical architecture
    c. Integration with existing Smithsonian systems
    d. Market appropriately for integration and education to SI staff

Membership

Members actively participate as equal and contributing technical and/or business resources. Members attend all of the meetings prepared, having performed all background reading and offer constructive comments and suggestions. Interested Parties have the

right to attend working group meetings and may be requested to provide information that will assist in requirements definition or implementation, but are not required or expected to do a large amount of technical research or attend all meetings. Interested parties will not formulate the final recommendation.

The following staff are Members:

- Alberto Accomazzi (SAO)
- Jon Blundell (OCIO-DPO)
- Alicia Cutler (NMAH)
- Raffaele D'Abrusco (SAO)
- Riccardo Ferrante (SIA)
- Andrew Gunther (OCIO)
- Alvin Hutchinson (SIL)
- Martin Kalfatovic (BHL)
- Isabel Meyer (OCIO-DAMS)
- Bess Missell (SIL)
- Edward L. Monk (OCIO)
- Thomas Orrell (NMNH)
- Suzanne Pilsk (SIL)
- Vince Rossi (OCIO-DPO)
- Arnold Rots (SAO)
- Jeffrey Smith (FSG)
- Rebecca Snyder (NMNH)
- Sara Snyder (SAAM)
- Adam Soroka (OCIO-RC)
- Beth Stern (OCIO-RC)

The following Interested Parties are informed of progress.

- Deron Burba (OCIO)
- Al Horvath (OUSFA)
- Scott Miller (DUSCIS)
- Ginger Minkiewicz (SISP)
- Tammy Peters (SIA)

- Bill Tompkins (NCP)
- Diane Zorich (OCIO-DPO)

Deliverables

- The following are suggested deliverables for the SI GUID working group:
- Governance structure
- Guidance documentation for GUID type selection
- Policy or guidelines on acceptable materials/objects for which GUIDs should be created
- Clear communication to stakeholders and interested parties at the Institution for adoption
- Implementation Plan
- List of SI Systems that support use of GUIDs and recommendations for systems that do not

Timeline

The SI GUID working group will start in Q3 of FY 2018. The working group should plan on completing all of its work and submitting an Assessment and Recommendation along with all associated documentation within a calendar year.

Resources

- Starr, Joan and Chodacki, John. “Identifiers and Citations: Frequently Asked Questions” (PowerPoint).
- California Digital Library. http://www.cdlib.org/services/infrastructure/docs/Identifiers_Citation.pptx