Introduction

This document is a summary of information collected during a requirements gathering meeting with the Senior Solutions Architect at the Smithsonian Institution office of the CIO held on 4/10/2018. The institutional representative provided information on the digital repository systems currently in use, strengths and limitations of the current systems, and desired digital repository features. [Descriptions of representative collections from individual SI units are pending.]

Summary of Key Findings

Topics covered in the interview included the factors of SI's institutional context that have a bearing on digital repository use across the Institution, the current systems and their limitations, and desired features in a new system. Summary information on each of those is presented below.

Organizational Context

Several characteristics of SI are relevant to its digital repository use and needs. First, as a public institution, the Smithsonian has a legal responsibility (as well as a strong institutional commitment) to preserve and provide access to the national collections. This means that SI cannot rely on outside organizations to manage their digital collections, but must apply policy and management of digital collections locally. This is difficult to achieve, however, due to constraints imposed by another main characteristic: With its 19 museums (referred to as "units") which themselves contain libraries, galleries, and research centers, SI is large, highly diverse, and "radically decentralized." Its institutions present a wide variety of collections and collection creation workflows.

A successful digital collections management policy and digital repository system(s) should be adaptable to a wide variety of use cases – able to handle large and complex objects as well as large collections of small objects – and make drawing connections between collections possible.

The Current Systems

The various digital repository systems currently in use across the Smithsonian Institution, including DSpace, SIdora (built on Fedora 3), and the Digital Assets Management System (DAMS), a commercial enterprise system, are all failing to accommodate SI's digital collections storage and management needs in several ways.

Although some systems (like the DAMS system) have not reached their maximum storage capacity, these systems cannot adequately manage complex resources like 3D models, which are both large and require specialized systems and workflows to adequately store and manage them. The SI collections present two different kinds of scale challenges: collections contain both individual objects that are very large (in the scale of terabytes and petabytes) and collections of small objects that in aggregation form very large collections. A successful digital repository system would have to handle both these challenges, and manage millions of objects and up to 10 petabytes of data at a time. Because of the large size of some assets, the ability to compute on objects inside the repository is a highly desired feature in a new system.

Aside from scale issues, a major limitation is the difficulty of interoperability among the current repositories. The current systems are silos and do not adequately support collaborative efforts between SI's various sub-organizations, such as virtual exhibits that draw from the collections of various units to present thematically related but organizationally distributed artefacts. One reason this is difficult is that the digital repositories do not share any metadata. Although SI is not interested in managing all resources in one system, it would value a degree of compatibility between different systems that would make working across digital repositories easier. [Using linked data]

Because of the size and decentralized nature of SI's units, it is not possible to present a limited number of workflows or collection structures that are representative of the average collection throughout the whole institution. More research is needed on individual units to compile accurate representations of individual use cases.

Desired Features

As mentioned above, the ability to compute on objects inside the repository is a high priority in a potential digital repository system. Other valued features are easy scalability and flexibility of metadata requirements. There is a strong direction towards making use of linked data, both from central units like the office of the CIO and from individual institutions within SI. A new system should enable connection between collections across SI without imposing a uniformity that would undermine the decentralized way SI's units operate.

More detailed information on the Smithsonian's digital repository use and needs is presented in the table below.

Variable	Response Summary
CURRENT SYSTEM	
Current digital repository software	SIdora – Built on Fedora 3 and Islandora DSpace – In use at SI libraries DAMS – Digital Assets Management System (commercial product)
New software under consideration	Undecided.

Collection creation workflows	Very diverse – workflows are specific to the domain and people who are executing them, and the nature of digital objects managed varies widely across different SI units. Some uniformity in digitized collections introduced by a central Digital Programs Office - Digital Assets Management Systems (DAMS) increasingly in use. Custom workflows sometimes necessary due to the absence of an existing solution (as in 3D imaging).
Limitations of the current system – what to change	The current systems are silos, and make it difficult to group thematically interconnected but organizationally distributed artefacts and collections across SI.
Failing point of current storage strategies	SI collections present scale challenges both in terms of: 1) *Mass* – "in aggregation, how much stuff do you have?" and 2) *Weight* – large size of individual objects (in the scale of TBs). Most limitations are social/organizational, not technical. Current storage strategies are already failing – projects are halted because of limits in storage capacity. Enterprise systems like DAMS still have some space, but are not an adequate storage and management solution for all digital collections because they cannot accommodate customized workflows
Workarounds in case of requirements exceeding system capabilities	- Providing disk space outside the institutional repository - Offsite storage with external partners in collaborative projects - Gathering support through grants
Linked data use	SI units do use linked data, and plan to use it more heavily in the future (strong push for this from the CIO's office). CIO's office aims to provide central support to linked data projects SI units undertake independently. Observation: although there is awareness of the potential of linked data projects, they often do not go farther than producing "a pile of triples." No interesting applications are developed due to lack of incentives.

Desired digital repository functions	- Ability to perform analysis functions inside the repository - Easy scaling - Flexibility of metadata (to enable precise and rich description of a wide variety of resources and metadata use cases (i.e. both access and preservation))
Most important features of the current system	Most important features vary depending on the type of stakeholder and the system in question. Stakeholders include: - Researchers (might prioritize easy input/output and analysis capabilities) - Museum curators (might prioritize description) - Central unit staff (might prioritize durability) These sometimes-competing priorities must be reconciled in a system that best serves the needs of all stakeholders.
REPRESENTATIVE COLLECTIONS	
Structure of the current repository	
Size of current digital collections	Tens of millions of resources; 6 – 10 petabytes of capacity
Size and structure of example collections	
Most common file formats	
Types of Metadata	
Rate of collection growth	
Areas of growth	
PERFORMANCE METRICS FOR THE NEW SYSTEM	

Input/Output	Ingest/export requirements depend on whether it is possible to do compute inside the repository. Eliminating the necessity to download and upload large files during analysis activities would be a highly valued feature due to the PB scale the largest objects, which makes moving them around difficult.
ASPIRATIONS AROUND THE REPOSITORY	
Aspirations for future technical improvements	An ideal system would enable: Scalability Flexibility of metadata Computing on assets inside the repository
Technical barriers	[Not asked]