Introduction

This document is a summary of information collected during a requirements gathering meeting with the Georgetown University Library's Head of Digital Services, the library's Senior Programmer Analyst, and the Senior Systems Administrator held on 3/9/2018. The institutional representatives provided information on the digital repository systems currently in use, strengths and limitations of the current systems, desired digital repository features, and descriptions of representative library collections.

Summary of key findings

DigitalGeorgetown, the University's digital repository, currently uses DSpace 5.8 to manage the digital collections of six university communities. DSpace successfully meets the library's needs and it is expected to be a sustainable digital repository solution for Georgetown. Some of the main success factors cited for the software are its high reliability, the customization and automation capabilities made possible by its open source nature, and the availability of in house developers and systems administrators to tailor the repository software's performance to the Library's needs. The repository's growth has so far been predictable and it has been possible to expand the repository's storage capacity as needed, made easier by the fact that streaming media is stored in a server outside DSpace and that DSpace is not used as a preservation repository – preservation is done separately in partnership with AP Trust.

Some general requirements for a digital repository were identified as open source software (enabling customization, automation, and control over system maintenance), easy migration to new versions, and easy scaling. Other desirable features include full text searching with highlighting of word matches and side-by-side transcription.

Overall, DigitalGeorgetown houses around 590,000 items belonging to six separate communities:

- The Bioethics Research Library of the Kennedy Institute of Bioethics

○ Includes the largest collection of ~278,000 items, comprised mostly of bibliographic citations

- The University Library's Digital and Special Collections
- The Law Library
- The University's Institutional Repository
- The University Publications
- The collections of the Initiative for Technology Enhanced Learning (ITEL)

10 collections were added to the repository last year, and a constant or faster rate of growth is anticipated. The most common file types used are PDFs and JPEGs. The LIT team is currently introducing IIIF compliant processes and systems and is interested in developments in linked data.

More detailed information on the library's current systems, digital repository needs, and representative collections is presented in the table below.

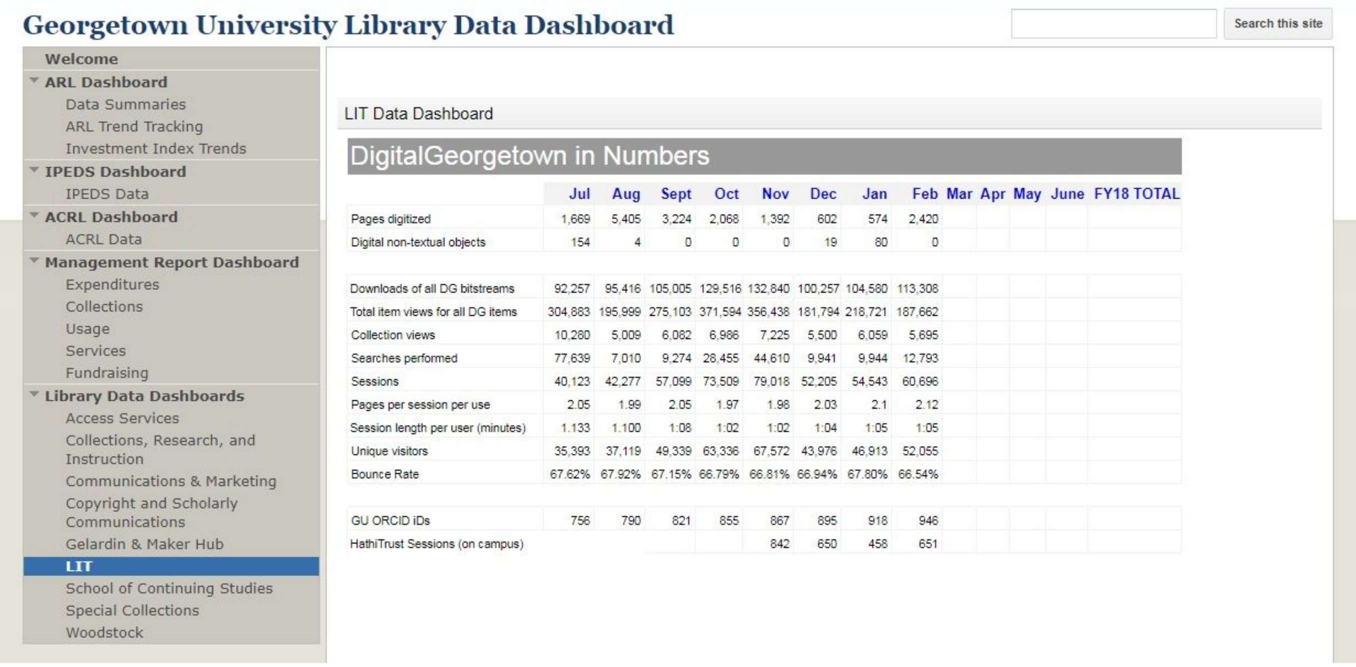
Variable	Response Summary
CURRENT SYSTEM	
Current digital repository software	DSpace 5.8 Tools integrated with DSpace: - FlexPaper document viewer - File Analyzer tool built by the library's application developer for creating inventory and metadata management
New software under consideration	- Introducing IIIF compliant processes and platforms (currently in progress) - Migration to DSpace 6.x in the next 12 months and eventually to DSpace 7
Collection creation workflows	**<u>Digitized materials</u>** - Digitized in house 1. Create preservation files for ingest into a separate preservation repository 2. Create derivatives (JPEGs and PDFs) for access copies to store in DigitalGeorgetown, the DSpace repository 3. Reformat files using tools like Photoshop, Bridge, and Media Encoder. - Digitized by vendors **<u>Born digital content</u>** - File collection 1. Received from donors through email or Box (example donors: institutes on campus) 2. Some donors submit directly to the repository (for content like faculty papers) 3. For Proquest: submitted through an SFTP dropbox - File staging 1. Content submitted directly to the repository goes through vetting, approval, and description processes

	2. File Analyzer tool used for a variety of processes, including creating inventories, generating metadata, and authenticating metadata **Note:** Automation was introduced whenever possible, and users of DSpace given access to the system through a web interface to eliminate the need for using command line Preservation system (with AP Trust) 1. Preservation copies of files uploaded into AP Trust 2. An e-tag is created for the file and is automatically written back to the metadata for the item in DSpace, thus identifying items that have been uploaded to the preservation repository
Limitations of the current system – what to change	No major limitations – DSpace satisfies the library's digital repository needs with the help of automation and customization. Success factors: - Availability of systems administrators and developers in house to introduce tools and processes tailored to the library's needs - Hosting audio and video content in a streaming media server outside DSpace Some minor limitations: - DSpace might not be able to accommodate dramatic growth in collections or users, but this scale of growth is unlikely - Similarly, DSpace might not be able to support a higher amount of processing on all content if that becomes necessary - A policy change requiring release of high-resolution files (e.g. tiffs instead of jpegs) would be hard to accommodate in DSpace, but this scenario is unlikely. - Implementing IFFF compliance is leading to complicated workarounds, e.g. storing objects outside DSpace - The FlexPaper PDF viewer is not reliable in displaying irregular PDFs. - Future challenge: migrating to DSpace 7 will be a labor intensive process expected to take 4-6 months. - DSpace does not handle compound objects well o Example: hierarchical archival collections with a series-box-folder structure. DSpace is designed to have item level metadata, which is not always necessary or feasible for all collections

	- DSpace cannot support fixity checks for large collections (example: a collection with 500,000 images) but performing fixity checks in DSpace is not currently a priority since the library uses a separate preservation repository
Workarounds in case of requirements exceeding system capabilities	The library team would work further on customizing DSpace so that it works for them. In the past they have been able to add storage and memory to keep up with growth in collections and users.
Desired digital repository functions	- Open source o Customizable o Allows for automation o Allows for control over system maintenance - Easy scaling - Full-text searching and indexing, highlighting word matches within documents - Side-by-side transcription of documents - Seamless automated ingest of faculty publications
Most important features of the current system	- Stable and reliable – functions as expected, never crashes. This enables the library to meet the expectations of patrons and donors. - Open source o Easily customizable Enables use of a web interface so that team members can use DSpace tools Enables creating custom themes and facets for collections o Ability to introduce automation as needed o Control over system maintenance - Ability to interact with DSpace tools using a web interface, enabling team members to use tools without relying on a systems administrator - DSpace has scaled well using added storage and memory. This is due to the constant and steady rate of growth of the repository that could be anticipated and planned for.
REPRESENTATIVE COLLECTIONS	

Structure of the current repository	Hierarchical structure: 6 communities sub-communities (archives & special collections, rare books, manuscripts, art, and university archives) Collections Items
Size of current digital collections	~590,000 items in total
Size and structure of example collections	Largest collections: - 278,000 item bioethics bibliographic collection - Art history collection with 150,000 items (digitized slides and metadata) Remaining content: ~120,000 items - The average large collection contains ~2,400 items Collections digitized in house range from 50-200 items.
Most common file formats	PDF and JPEG files
Types of Metadata	10 – 15 Dublin Core elements used consistently for all objects; further metadata varies widely depending on the type of object/collection; different standards because partner organizations administer their own collections.
Rate of collection growth	10 collections created in 2017 – more rapid (roughly linear) growth projected in coming years.
Areas of growth	- In progress: a born-digital policy for electronic records with the university archives and the manuscripts division - Accepting born digital content (e.g. video) from the university's communications department - Growth in large file formats like video
ASPIRATIONS AROUND THE REPOSITORY	
Aspirations for future technical improvements	- Working with IIIF (pilot project released March 5th) - Innovations around linked data

	- Open access buttons (using DSpace more as an institutional repository for faculty publications – not currently the case)
Technical barriers that make it hard to work with repository content	Outdated streaming media system stored on a server separate for DSpace does not run well, is "clunky" and "hard to use"



Georgetown University Library Data Dashboard

LIT Data Dashboard

DigitalGeorgetown in Numbers

	Jul	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	FY18 TOTAL
Pages digitized	1,669	5,405	3,224	2,068	1,392	602	574	2,420					
Digital non-textual objects	154	4	0	0	0	19	80	0					
Downloads of all DG bitstreams	92,257	95,416	105,005	129,516	132,840	100,257	104,580	113,308					
Total item views for all DG items	304,883	195,999	275,103	371,594	356,438	181,794	218,721	187,662					
Collection views	10,280	5,009	6,082	6,986	7,225	5,500	6,059	5,695					
Searches performed	77,639	7,010	9,274	28,455	44,610	9,941	9,944	12,793					
Sessions	40,123	42,277	57,099	73,509	79,018	52,205	54,543	60,696					
Pages per session per use	2.05	1.99	2.05	1.97	1.98	2.03	2.1	2.12					
Session length per user (minutes)	1.133	1.100	1:08	1:02	1:02	1:04	1:05	1:05					
Unique visitors	35,393	37,119	49,339	63,336	67,572	43,976	46,913	52,055					
Bounce Rate	67.62%	67.92%	67.15%	66.79%	66.81%	66.94%	67.80%	66.54%					
GU ORCID iDs	756	790	821	855	867	895	918	946					
HathiTrust Sessions (on campus)					842	650	458	651					

Monthly statistics for DigitalGeorgetown, July 2017 - February 2018.