Introduction

This document is a summary of information collected during a requirements gathering meeting with the Digital Programs & Initiatives Manager and the Software Systems Development and Research Manager at the University of Maryland Libraries held on 2/27/2018. 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 two large scale library collections.

Summary of key findings

The library is faced with managing increasingly large digital collections both for research data and UMD scholarly production (currently housed in DSpace 5) and unique digital collection materials originating from the library (managed using Fedora 2 and 4). The current total size of digital collections is estimated to be 100 TB, with a separate collection not included in that total that is itself estimated to be in the 100s of TBs (the Gordon W. Prange Collection). The estimated rate of growth of digital collections is 30 TB per year.

The new digital repository software should have the capacity to comfortably manage petabytes of data as well as the flexibility to be applied to a wide range of collection types – from research datasets to humanities projects containing a wide variety of file formats. The most common file format in the library's collections is image (.tiff) files, but they are increasingly also working with audiovisual materials like .wav files.

One important goal for the library's digital repository systems is to consolidate the management of all digital assets under one centralized repository rather than the current system using three separate repository softwares. A desired feature is the ability to develop specific applications that can be used with a single underlying digital repository system to support the diverse needs of collections managers and systems administrators as well as the collections' end users. This is becoming an increasingly urgent necessity as the growth in volumes of collections continues to outpace the capacity of the current systems in place and as large and complex collections from all three repository systems present similar challenges that could be better met with a single robust but flexible repository software.

Other desired features are the presence of a strong community and collaborative development around the software. The ability to develop applications that meet diverse user needs and share the applications for reuse would be valuable as a tool to adapt the capabilities of the centralized repository software for a range of institution-wide use cases.

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	- **Fedora 2** for digital collections; simultaneously introducing **Fedora 4** (one collection currently in Fedora 4); - **ShareStream** (loosely coupled with Fedora) for streaming media. Plans to pilotusing Avalon. - **DSpace 5** for UMD scholarly production/research data, upgrading to DSpace 6 later this year
New software under consideration	- Samvera as a replacement for DSpace – not ready to transition to it because of issues in the Samvera community (lack of central governance, system requires frequent updates, uncertain commitment to Fedora as backed repository). - Avalon for streaming media (for better integration with Fedora)
Collection creation workflows	Described as “nascent.” **Born digital content:** packaged using BagIt and stored on a restricted NFS mount, with backups.. Electronic records archivist working on establishing a more formal process (but none in place currently [verify]). **Research Data:** Data Services Librarian ensures datasets are managed according to standards of individual funding agencies; not much processing applied to data deposits in DRUM. **Physical materials:** digitized in house or by vendors. Materials digitized by vendors (where quality assurance has been applied to completed metadata records) are loaded into Fedora 4. Materials digitized in-house are staged by the Digital Conversion and Media Reformatting (DCMR) department. Systems librarian applies archival processing to the materials and queues them up for ingest into the appropriate repository. Materials digitized in-house are currently loaded into Fedora 2 because it has an administrative interface where objects can be created, files loaded, and metadata added, which is not available in Fedora 4.
Limitations of the current system – what to change	- The current system is too fragmented – 3 separate repositories are used for the library’s two use cases – scholarly production (DSpace) and unique materials originating from the library (Fedora 2 & 4). Move towards one centralized repository for all digital assets.

	- A higher capacity for managing large research data collections (possibly in the petabytes)
	- Fedora 2 has rigid/complex metadata requirements. A system with minimal requirements (just a title and a file) where further metadata can be added after ingesting materials would be preferred.
	- Currently one developer who knows/supports the repository
	- Metadata currently time consuming to update.
	- It is currently difficult to manage AV materials. Challenges include:
	  o Size of assets
	  o Must play through contents in order to accurately describe them as they often come with minimal description
	  o The extent (play time) of physical media is hard to assess visually

Desired digital repository functions	(Emphasis on the need for scale and performance that enable the system to comfortably handle petabytes of data at a time)
	- General priorities: scale and performance (can handle petabytes of data)
	- There is a community around the software – the repository system comes with a community-supported set of tools; collaborative development
	- Support for AV materials
	- Support for disaster recovery and preservation options
	- Open standards
	- Support for linked data
	- Web Access Control
	- Integration with IIIF
	- Interoperability with other systems
	- The system enables democratized access, e.g. unmediated API level access for researchers
	- The system enables increased access/preservation of materials through intellectual property rights management capabilities; the system enables implementation of formally established permissions mechanisms
	- Metadata:
	  o Flexible/minimal requirements for compulsory metadata
	  o Easy to add mass updates without introducing mass errors

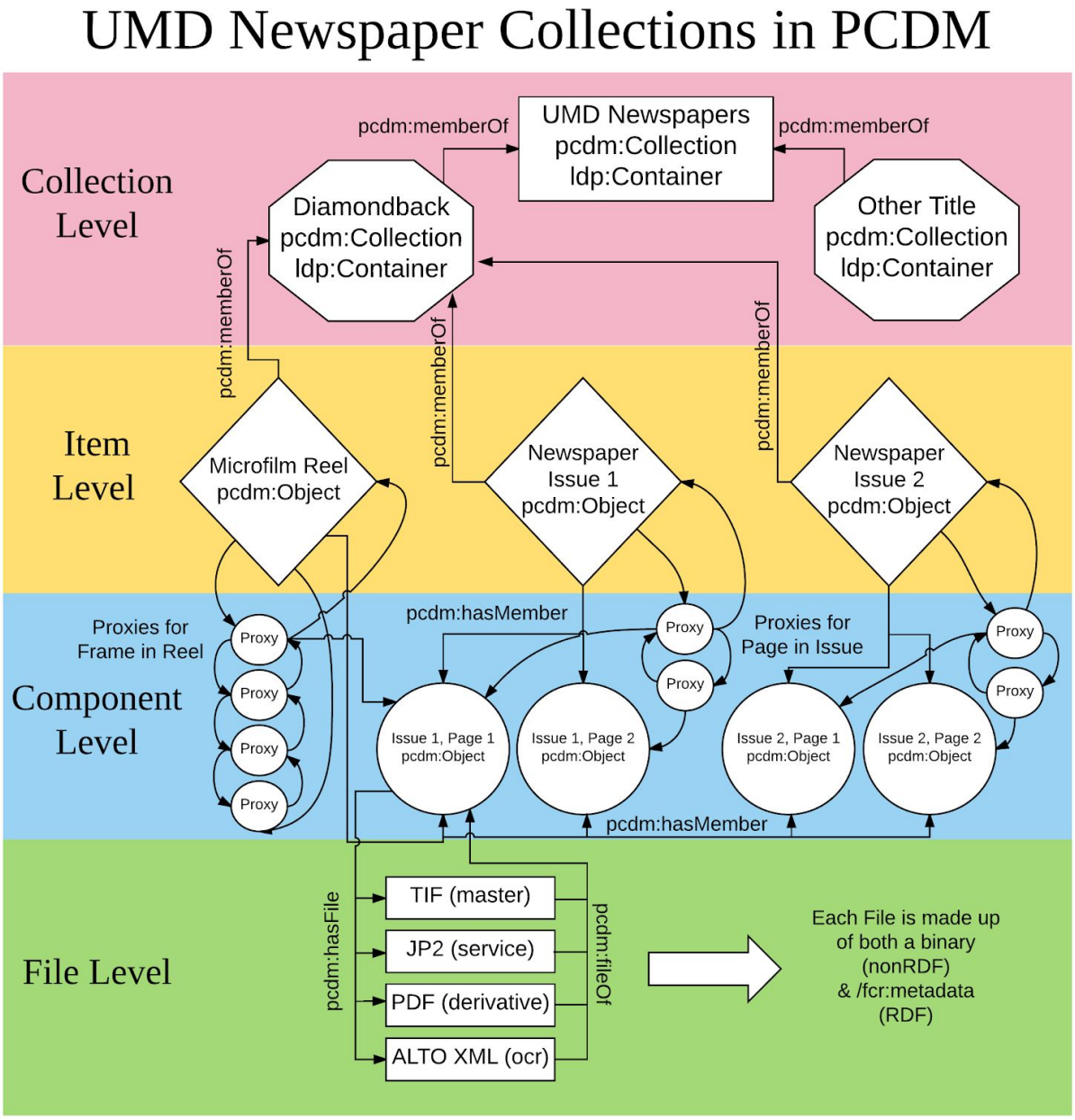
Most important features of the current system	Pros of DSpace: ○ Stable ○ Easy to update ○ Easy to maintain Pro of Fedora 2: ○ Administrative interface (not available in library's Fedora 4 system) ○ Web AC model for access control o Flexibility of the content model o Native support for Linked Data Platform and RDF
REPRESENTATIVE COLLECTIONS	
Size of current digital collections	100 TB total
Size and structure of example collections	1) Student newspaper collection in Fedora 4 ○ 2.5 million resource nodes ○ 700 issues ○ 77,000 pages ○ 324,000 binaries (2.6 TB) ○ 1.6 million annotations 2) Gordon W. Prange Collection ○ Size: hundreds of TBs (pending more specific estimate) ○ 40 – 80 TB digitized already
Most common file formats	Text/converted print materials (.tiff files) and increasingly AV formats (.wav)
Types of Metadata	Mostly descriptive and administrative (not a lot of preservation metadata)
Rate of collection growth	30 TB a year (expected to be the consistent rate for coming years) DRUM (UMD scholarly production): 7-8 IR collections per year Slow collections growth currently due to ongoing transition from Fedora 2 to Fedora 4.
ASPIRATIONS AROUND THE REPOSITORY	

Aspirations for future technical improvements	- Use of cloud computing (Division of IT currently has an AWS account but it is unclear if it would be feasible to use it for the library's digital collections) - The system enables increased access/preservation of materials through intellectual property rights management capabilities; the system enables implementation of formally established permissions mechanisms - Non-synchronous, non-HTTP ingests as collections to ingest get larger – possibly a peer-to-peer solution for submitting files.

Example Collection 1: The UMD Student Newspaper Collection

The library maintains a database with all the issues of the UMD student newspapers from 1910 to the present. This collection is housed in Fedora 4 (See the table above for some of the collection characteristics).

Simplified model diagram



Source: The project's page on the UMD Libraries Digital Systems and Stewardship Knowledge Base Website. Link:

https://confluence.umd.edu/pages/viewpage.action?title=Newspapers+in+PCDM+diagram&spaceKey=LIB.

Example Collection 2: The Gordon W. Prange Collection

As described on the collection's website (https://www.lib.umd.edu/prange): "the most comprehensive archive in the world of Japanese print publications issued during the early years of the Occupation of Japan, 1945-1949." The estimated total size of the digitized materials (the full collection will be digitized) to be in the 100s of TBs. 40 – 80 TB have been digitized already.

The collection's holdings include:

- 71,000 book and pamphlet titles
- (including 1,000 book and pamphlet titles with censorship action taken and 1,500 galley proof fragments)
- 140 ephemera items
- 13,800 magazines titles
- 18,000 newspaper titles
- 10,000 news agency photos
- 640 maps
- 90 posters

The collection is managed using separate workflows from the library's other collections, which is why it is not included in the total estimated size of the library's digital collections. A more detailed description of the TB size and structure of the collection should be available in the coming weeks.

Possible follow up questions/areas:

- Gather more information on collection creation workflows (file deposit, staging, and ingest) for physical materials, born digital materials, and research data. Represent workflows for each type of material diagrammatically.

- Obtain a more specific description of the size and structure of the Prange Collection (given as an example of one of the largest collections)

- Democratized access is described as a priority. Unmediated API access for researchers is an example capability given. What other capabilities of a digital repository can support democratized access?

- What system features are needed to support digital humanities projects?

- What system features are needed to support intellectual property rights and access management?

- Get a clearer idea of the library's current metadata use.