UC Curation Center
CDL
California Digital Library

UC Curation Center / California Digital Library

UC3 Curation Foundations

Rev. 1.0 – 2010-03-25

1 Introduction

Information technology and resources have become integral and indispensable to the pedagogic mission of the University of California. Members of the UC community routinely produce and utilize a wide variety of digital assets in the course of teaching, learning, and research. These assets represent the intellectual capital of the University; they have inherent enduring value and need to be managed carefully to ensure that they will remain availed for use by future scholars. Within the UC system the UC Curation Center (UC3), one of five programmatic areas of the California Digital Library (CDL), has a broad mandate to ensure the long-term usability of the University's digital assets. UC3 was previously known within the CDL as the Digital Preservation Program. However, "preservation" is a highly elastic term applicable to a continuum of intentions, activities, and outcomes, each with its attendant level of effort and efficacy.[1] To better position itself to meet the obligations of this complex task, UC3 is engaged in a process of reinvention involving significant transformations of its outlook, effort, and infrastructure (see Table 1).

Mission	Preservation	⇒	Curation
Approach	Project	⇒	Programmatic
Emphasis	Systems	⇒	Services
Priority	Repository	⇒	Content

Table 1 – UC3 reinvention

UC3 now defines its mission in broader terms of digital *curation*, rather than preservation. Digital curation is the set of policies and practices aimed at maintaining and adding to the value of trusted digital content over time.[2] Given its academic setting, UC3 curation activities are facilitating the alignment of the scholarly and information lifecycles (see Figure 1). Curation better expresses the need for the coordinated activities of *preservation* of, and *access* to, managed assets.[3] While preservation and access were previously considered disparate functions, they are now properly seen as complementary: preservation aimed at providing access to managed content *over time*, while access depends upon preservation at a *point in time*. Curation also better connotes the ongoing *enrichment* of managed content; that is, the curation intention is not only to maintain content as originally acquired, but also to add value to it. All of these activities are relevant throughout the full digital lifecycle, ideally starting before an

[1] See, for example, Association for Library Collections and Technical Services, *Definitions of Digital Preservation*, June 24, 2007 <http://www.ala.org/ala/alcts/newslinks/digipres/index.cfm>; and Brian Lavoie and Lorcan Dempsey, "Thirteen ways of looking at … digital preservation," *D-Lib Magazine* 10.7/8 (July/August 2004) <http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>.

[2] Daisy Abbot, *What is Digital Curation?* April 3, 2008 <http://www.dcc.ac.uk/resource/briefing-papers/what-is-digital-curation/>.

[3] Chris Rushbridge, "'Digital Preservation' term considered harmful?" Digital Curation Blog, July 29, 2008 <http://digitalcuration.blogspot.com/2008/07/digital-preservation-term-considered.html>

asset is first created and lasting until its ultimate disposition.[4] Consequently, UC3 has shifted to an open-ended *programmatic*, rather than a time-bounded *project*-oriented approach. This new approach is accompanied by a shift in emphasis from *systems* to *services*. Technical systems are inherently ephemeral, their useful lifespan being constantly encroached upon by disruptive technological change. Rather than pursuing the somewhat illusory goal of long-lived systems, curation goals are better served by concentrating on persistent services that can evolve and be easily reimplemented as necessary while continuing to provide necessary function.[5] This change in emphasis is best exemplified by a concomitant deprecation of the centrality of the curation repository as *place*.[6] Rather than relying on a conceptually monolithic system as a locus, curation outcomes should be the product of a set of small, self-contained, loosely-coupled, and distributed services capable of operating on content *in situ* without a necessary precondition of being transferred to a central point for processing.

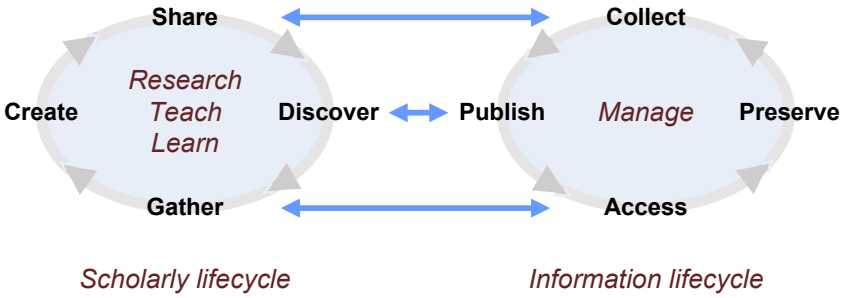

Figure 1 – Alignment of scholarly and information lifecycles

In trying to devise appropriate and attainable goals for UC3 in an environment of ever increasing demands for services on an expanding variety and volume of content, it is necessary to achieve a careful balance between desirable function and available resources. To ensure a sound theoretical basis for decision making it is useful to start from first principles and consider the application to the digital realm of historically-established principles of library and archival science.

2 Foundational principles

> The foundation of empire is art & science.
>
> – William Blake

Library and archival science have a deep, rich history of theoretical analysis that still has currency when applied to the digital realm. Ranganathan's "laws" of library science are particularly pertinent in this

[4] Sarah Higgins, "The DCC curation lifecycle model," *International Journal of Digital Curation* 1.3 (2008): 134-140 <http://www.ijdc.net/ijdc/article/view/69/69>.

[5] Peter J. Denning, Chris Gunderson, and Rick Hayes-Roth, "Evolutionary system development," *Communications of the ACM* 51.17 (December 2008): 29-31.

[6] Owen Stephens, "Thinking we need to stop thinking of the 'repository' as a 'place'," Twitter, July 3, 2008 <http://twitter.com/ostephens?page=7>; and Stephen Abrams, Patricia Cruse, and John Kunze, "Preservation is not a place," *International Journal of Digital Curation* 4:1 (2009): 8-21<http://www.ijdc.net/index.php/ijdc/article/viewFile/98/73>.

regard: [7]

- The first three laws ("*Books are for use*"; "*Every reader his book*"; "*Every book its reader*") concern use. By analogy we initially assert that digital assets are curated in order to be used. More specifically, these assets must be *discoverable* and *utile*; that is, they can be found by users and their encapsulated information content can be meaningfully exposed to those users.
- The fourth law ("*Save the time of the user*") is fundamentally concerned with service. All digital assets require technological mediation in order to be useful. So by analogy we assert that user services built around curated digital assets must be *available*, *responsive*, and *comprehensive*, that is, the services can be used at the time and place of user choosing and they conform to user performance and functional expectations.
- The fifth law ("*The library is a growing organism*") is fundamentally concerned with change. Due to the nature of mediation, digital assets are inherently fragile with respect to technological change. So again by analogy we assert that digital curation activities must *evolve* over time in a *sustainable* manner in order to continue to mitigate against risks and threats as they arise. This service obligation is facilitated by reliable and predictable administrative, financial, and professional support.

The concept of archival diplomatics stresses the importance of *provenance*, the understanding of an asset's source and relationship of the carrier to the information content it encapsulates.[8] One of the distinguishing characteristics of digital content over analog forms is its ease of undetectable mutability. So by a final analogy we can extend our initial assertion to say that curated assets must not only be accessible and usable, but also *authentic*; that is, that they are what they purport to be.

In summary, the primary curation imperative of UC3 can be expressed as *providing highly available, responsive, comprehensive, and sustainable services for access to, and use and enhancement of, authentic digital assets over time*. It is important to note that these goals merely restate a set of stewardship responsibilities that scientific and cultural memory institutions have always carried out as part of their mission: maintaining and adding value to authentic digital assets for use now and in the future.[9]

3 Curation objects

> I've information vegetable, animal, and mineral.
>
> – W. S. Gilbert, *The Pirates of Penzance*

The primary unit of curation management is the *digital object*, an encapsulation in digital form of an

[7] S. R. Ranganathan, *The Five Laws of Library Science* (Madras, 1931).

[8] Seamus, Ross, "Digital preservation, archival science, and methodological foundations for digital libraries," *ECDL 2007: The 11th European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, September 16-21, 2007.

[9] Digital Curation Centre, *About the DCC: What is Digital Curation?* April 26, 2007 <http://www.dcc.ac.uk/about/what/>.

abstract intellectual or aesthetic *work*.[10] The fundamental components of a digital object with regard to curation considerations are (see Figure 2):

- *Content*. The information content inherent to the underlying work. Within an object the abstract semantic *meaning* associated with this content is represented in tangible syntactic *form* and is exposed through pragmatic *behavior*. Content components may exist within a network of typed relationships to other components, such as *derivative-of* or *color-profile-for*.
- *Description*. Extrinsic information *about* the object and its content in terms of metadata defining their significant syntactic, semantic, and pragmatic characteristics, and the *change* of those characteristics over time.

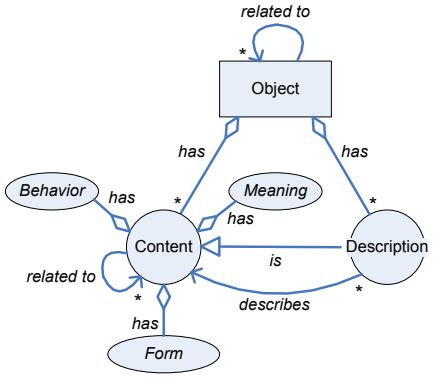

Figure 2 – Digital object composition

Objects themselves may exist within a network of typed relationships to other objects, such as *new-edition-of*.

This modeling scheme is intended to be both generic and capable of modeling complex expressions of digital content. Note that the object decomposition model (*content*, *description*) is a reasonable analogue to the OAIS (ISO 14721) information package model (see Table 2).[11] Content description fulfills the OAIS obligation that managed content be *independently understandable* by a designated community, that is, without the intervention of domain specialists.

	OAIS
Object	Information object
Content	Data object
Description	Representation information

Table 2 – Comparison with OAIS information packages

The tripartite nature of object content (*meaning*, *form*, *behavior*) follows from a semiotic conceptualization of preservation activity. Semiotics explains the transference of meaning across time

[10] IFLA, *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications – New Series Vol. 19 (Munich: K. G. Saur, 1998) <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.

[11] ISO 14721:2003, *Space data and information transfer systems – Open archival information system – Reference model*.

and space in terms of *signs*, meaning-bearing symbols. As formulated by Morris, syntax is the relationship *between* signs; semantics is the relationship between signs and the things that they *signify*; pragmatics is the relationship between signs and the *agents* who interpret them.[12] Note that this is reasonable analogue to the National Archives of Australia (NAA) performance model (see Table 3): [13]

	Morris	***NAA***
Meaning	Semantics	Performance
Form	Syntax	Source
Behavior	Pragmatics	Process

Table 3 – Comparison with NAA performance model

The OAIS concept of representation information is defined explicitly in terms of structural – that is, syntactic – and semantic information. The OAIS model does makes provision for "other" representation information as well, but the importance of managing behavioral description is not emphasized. This appears to be a significant omission since without appropriate behaviors to convert digital content into human-sensible form that content is inaccessible. It is therefore preferable to raise the importance of behavior explicitly to the level of syntax and semantics.

Two further aspects of semiotic theory bear on curation thinking.

- First, under Peirce's theory of semiosis the meaning of a sign (or in curation terms, a digital object) understood by its interpreter (or a user or designated community) is dependent upon that interpreter's semiotic ground (or knowledge base).[14] Since that ground is unique to time, place, and individual, semiosis (or preservation) is at best an approximate mechanism. In other words, preservation success should not be evaluated on an all-or-nothing basis. Rather, a continuum of outcomes is to be expected. Under certain circumstances some loss may be inevitable; in other cases loss up to a prescribed level may be anticipated by design.
- Second, the value attributed to a sign/object is ultimately determined by its interpreter/user. Value is therefore not fixed, but can be *added* to that already carried by a curated object by uses and re-uses unintended by its original creator.

It is important to note that while many curation activities involve the manipulation of tangible digital artifacts (i.e. files), the true focus goal of these activities is the underlying information meaning of which those artifacts are merely the carrier. In other words, bits are the means, content is the ends.[15]

[12] Charles W. Morris, *Foundations of the Theory of Signs* (Chicago: University of Chicago Press, 1938).

[13] Helen Heslop, Simon Davis, and Andrew Wilson, *An Approach to the Preservation of Digital Records*, National Archives of Australia Green Paper, Canberra, December 2002 <http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf>.

[14] Eugene Rochberg-Halton and Kevin McMurtrey, "The foundations of modern semiotic: Charles Peirce and Charles Morris," *American Journal of Semiotics* 2.1-2 (1983): 129-56

[15] See, for example, James Currall, "The fetish of the digital," *JISC-PoWR blog*, January 7, 2009 <http://jiscpowr.jiscinvolve.org/2009/01/07/the-fetish-of-the-digital/>.

4 Curation function

I dwell in possibility.

– Emily Dickinson

The value of a digital object – whether cultural, scientific, or economic – is predicated by its intended and actual use. However, in the context of curation over any significant time period the users and uses of a digital asset cannot be definitively known *a priori*. UC3 will therefore accept custodial responsibility for digital objects from UC-affiliated agents *regardless* of object provenance, structure, format, or characterization by accompanying metadata. However, the level of assurance of ongoing usability applicable to a given object is subject to limitations imposed by these formal factors (or their absence), the general state of curation understanding, and other CDL priorities. Preservation assurance – focused on *maintaining* the value of managed digital content – coalesces into two fundamental service levels: [16]

- When requested, a bit-faithful *copy* of the digital object originally submitted for management can be returned.
- When requested, a potentially different *version* of an original object, *representing the same underlying content* but supporting appropriate behavior in a contemporaneous technological context can be returned.

Curation value can be *added* to managed content by enabling its creative use and reuse in whole, in part, or in aggregation with other content. These uses are facilitated by:

- Persistent citation and actionable reference.
- Discovery of both content and contextual description.
- Annotation for enriched description.

It is important to emphasize that digital curation is not solely a technical process.[17] While it is certainly necessary to have a robust, secure, and sustainable technical environment in which to manage digital objects, their active curation is also dependent upon significant human competencies, analysis, and decision making, both on the part of CDL curation managers and UC collection managers and curators.

[16] Tony Hendley, *Comparison of Methods and Costs of Digital Preservation*, British Library Research and Innovation Report 106, 1998 <http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html#_Toc422714267>.

[17] See, for example, Cornell University Library, *Digital Preservation Management: Implementing Short-term Strategies for Long-Term Problems*, 2007 <http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html>.

5 Curation risk

Risk comes from not knowing what you're doing.

– Warren Buffett

Trust, but verify.

– Ronald Reagan

While incipient technical obsolescence constitutes an obvious risk to usability, there is a wide variety of other factors that should be the focus of a comprehensive curation effort, including: [18]

- Natural disaster (e.g. fire, flood, earthquake)
- Facilities infrastructure failure (e.g. power, cooling, network connectivity)
- Storage failure (e.g. media degradation, drive failure)
- Server hardware/software failure
- Application software failure
- External dependencies (e.g. PKI failure)
- Format obsolescence
- Legal encumbrance (e.g. take down request or other enforceable impediment to access)
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of institutional commitment (e.g. shifting priorities)
- Loss of financial stability (e.g. reliable and predictable funding)
- Changes in user expectation and requirements

The effective mitigation of these multifarious risks necessitates an ongoing programmatic approach to curation, meaning one that is both proactive as well as reactive, encompasses both human and technological components, is capable of addressing concerns arising throughout the digital content lifecycle, carries with it a strong institutional commitment backed by sufficient resources necessary to meet its ongoing service obligations, and maintains a close, ongoing engagement with the stakeholder communities that are service consumers.

[18] McHugh, Andrew, Raivo Ruusalepp, Seamus Ross, and Hans Hoffman, *Digital Repository Audit Method Based on Risk Assessment*, Version 1.0 (draft), February 28, 2007, Digital Curation Centre/Digital Preservation Europe <http://www.repositoryaudit.eu>; and Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito, "Requirements for digital preservation systems: A bottom-up approach," *D-Lib Magazine* 11.11 (November 2005) <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>.

6 Curation strategies

> However beautiful the strategy, you should occasionally look at the results.
>
> – Winston Churchill

Prudent programmatic curation incorporates a deliberate top-down planning approach in which technical decision making is deferred until curation intentions and activities are clearly understood and unambiguously defined. There intentions are best articulated first in terms of desirable object- and service-centric *values*; curation activities can then be properly articulated in terms of *strategies* designed to foster those values.

From an object-centric perspective, the essential curation values and strategies are (see Table 4):

Value	***Justification***	***Strategy***
Identity	To distinguish an object from all others	Unambiguous persistent naming, actionable resolution
Viability	To recover an object from its medium	Redundancy, heterogeneity, media refresh
Fixity	To ensure that an object is unchanged from its accepted state	Redundancy, error-correcting codes, message digests, periodic audit
Authenticity	To ensure that an object is what it purports to be	Provenance, cryptographically-secure signatures
Ontology	To understand the significant nature of the object	Syntactic, semantic, and pragmatic characterization
Visibility	To enable users to find objects of interest	Public discovery systems and registries, exposure for web harvesting
Utility	To expose the underlying information content of an object	Behavior-rich delivery
Portability	To facilitate content sharing and succession planning	Self-contained, self-documenting objects, packaging standards
Appraisement	To understand the consequences of the passage of time	Analysis and assessment
Timeliness	To know when a preservation value is threatened	Technology watch, stakeholder engagement

Table 4 – Object-centric preservation values and strategies

From a service-centric perspective, the essential curation values and strategies are (see table 5):

Value	*Justification*	*Strategy*
Availability	To provide access at the time of a patron's choosing	Redundancy, automated failover
Responsivity	To provide appropriate throughput in servicing requests	Redundancy, load balancing
Security	To enforce appropriate use of systems and content	Cryptographically-secure identity and role management, access control mechanisms
Interoperability	To facilitate creative reuse	Standard interfaces
Extensibility	To enable graceful evolution	Granularity, orthogonality, virtualization
Trustworthiness	To promote users' sense of predictability and reliability	Transparency, audit, certification
Sustainability	To ensure ongoing access and use	Commodity components, institutional commitment, financial cost-recovery, professional development

Table 5 – Service–centric preservation values and strategies

Operationally, these curation strategies (or groups of related strategies) are formulated first as abstract *services*, which are then in turn implemented through concrete human *activities* and automated *systems*.

7 Curation services

> все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему [Happy families are all alike; every unhappy family is unhappy in its own way].
>
> – Leo Tolstoy, *Anna Karenina*

UC3 services have applicability over the full range of the digital content lifecycle. The primary focus of these services is curation risk assessment and mitigation.[19] While some subset of necessary curation services can be performed with respect to generic curation risk, many will be highly dependent on the particular characteristics and curatorial intentions associated with specific content. Therefore these analytical and consultative services are primarily provided through human effort, albeit often with significant technological augmentation.

These services include:

- *Creation / acquisition*. Best practice recommendations for the creation of curation-amenable objects, including:
    - Selection of formats best suited for representing content.

[19] See, for example, Cornell University Library, *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library*, May 2008 <http://ecommons.library.cornell.edu/handle/1813/10903>; and Harvard University Library, *Digital Repository Service (DRS) Policy Guide*, April 26, 2007 <http://hul.harvard.edu/ois/systems/drs/policyGuide>.

  - Development of appropriate technical specifications and workflows for object creation.
  - Selection of "best edition" from multiple versions of object content for use in creating derivatives.

- *Appraisal / selection*. Consideration of factors leading towards a decision to curate content, including:
  - Assessment of intellectual, aesthetic, economic, and artifactual value.
  - Rarity (or ubiquity), ease (or difficulty) of re-acquisition, and degree of alternative access.
- *Preservation planning*. Core activities focused on ensuring ongoing usability of managed digital objects.[20]
  - Understanding the significant properties, and the formal characteristics – such as format, structural relationships, and behavior – that expose those properties, of managed objects.[21]
  - Surveying the technological environment that mediates the use of these objects.
  - Consideration of changing behavioral expectations for that use.
  - Determination of the prospective resilience of managed objects based on their formal characteristics evaluated in light of environmental monitoring and user expectation.
  - Development of action plans, with associated trigger events, to ameliorate identified preservation risks.
- *Preservation intervention*. The complex of activities involved in executing action plans following trigger events, including quality assurance testing subsequent to the execution of the plans to ensure the efficacy of the intervention and the invariance of the resulting object's significant properties.
- *Service brokerage*. Selection of appropriate curation service providers and mediation of service-level agreement negotiation, including offerings from UC3, the CDL, the UC system, and external commercial, academic, and non-profit agencies.

8 Curation infrastructure

The locus for automated curation services is conventionally defined as a *repository*, most often

[20] See, for example, Stephen Strodl, Christoph Becker, Robert Neumayer, and Andreas Rauber, "How to choose a digital preservation strategy: Evaluating a preservation planning procedure.," *7th ACM/IEEE-CS Joint Conference on Digital Libraries*, Vancouver. June 17-23, 2007.

[21] Margaret Hedstrom and Christopher A. Lee, "Significant properties of digital objects: definitions, applications, implications," *Proceedings of the DLM-Forum 2002: @ccess and Preservation of Electronic Information: Best Practices and Solutions*, Barcelona, May 6-8, 2002, pp. 218-227 <http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf>; and Gareth Knight, *Framework for the Definition of Significant Properties*, version 1, May 2, 2008 <http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf>.

considered in terms of the OAIS reference model (see Figure 3). Historically the repository has been conceived of as a unitary system. (In part this is due to viewing OAIS as an *architecture*, not a *model*.) However, if the assumption of monolithic atomicity is adhered to slavishly the result can be large, cumbersome systems that are expensive to deploy and difficult to support. Current research in software engineering quality metrics indicates that the correctness of a technical system component is, in general, inversely proportional to its complexity.[22] This suggests that unitary repository function should be devolved into a multiplicity of granular, loosely-coupled, and distributed services.[23]

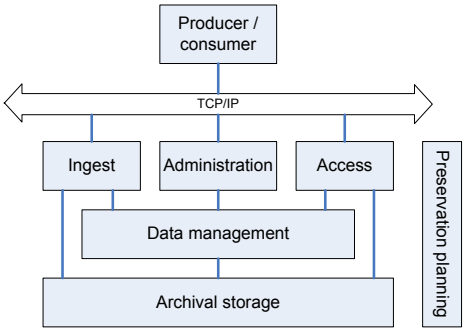

Figure 3 – OAIS repository model

Architecturally, these services should expose function through well-defined abstract interfaces that constitute the public service "contract" (see Figure 4). External agents can issue requests conforming to service interfaces without any knowledge of underlying service implementations. Service requests are made, and responses received, through concrete language bindings (for procedural interaction) or protocol bindings (for network interaction) that implement the interfaces.

[22] See, for example, Subhas C. Misra and Virenda C. Bhavsar, "Relationships between selected software measures and latent bug-density: Guidelines for improving quality," *Computational Science and Its Applications – ICCSA 2003*, Montreal, May 18-21, 2003, in V. Kumar, M. L. Gavrilova, C. J. K. Tan, and P. L'Ecuyer, eds., *Lecture Notes in Computer Science* 2667 92003): 724-32 <http://www.springerlink.com/content/4rcnfqqtn3fvetrh/fulltext.pdf>.

[23] See, for example, Sue Factor, *What makes a design "Googley"?*, The Official Google blog, posted April 23, 2008, <http://googleblog.blogspot.com/2008/04/what-makes-design-googley.html>; James Hamilton, "On Designing and Deploying Internet-Scale Services," *Proceedings of the 21st Large Installation System Administration Conference (LISA '07)*, Dallas, November 11-16, 2007, pp. 231-42 <http://www.usenix.org/event/lisa07/tech/full_papers/ hamilton/hamilton_html/index.html>; and Philipp Liegl, "The strategic impact of service oriented architectures," *Proceedings of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS '07)*, Tucson, March 26-29, 2007.

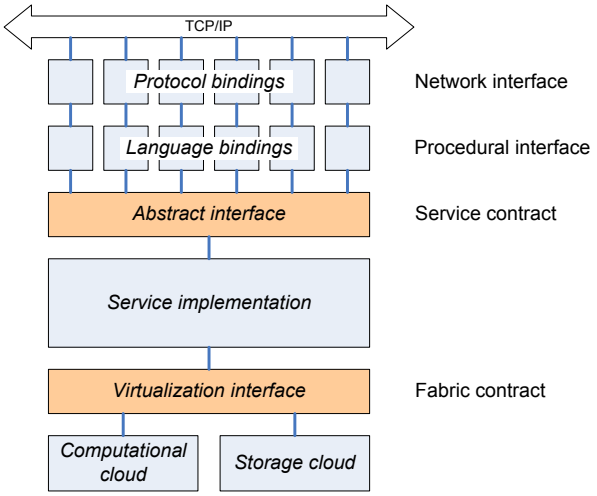

Figure 4 – Curation service stack

This decentralized approach follows from conceptualizing curation from a content-centric, rather than systems-centric perspective.[24] In other words, preservation is not a *thing* into which content is placed for safe-keeping, but rather, it is a *process* in which content evolves both proactively and reactively over time through the application of strategy-embodying services operating on the content as necessary to promote the essential curation values (see Figure 5).

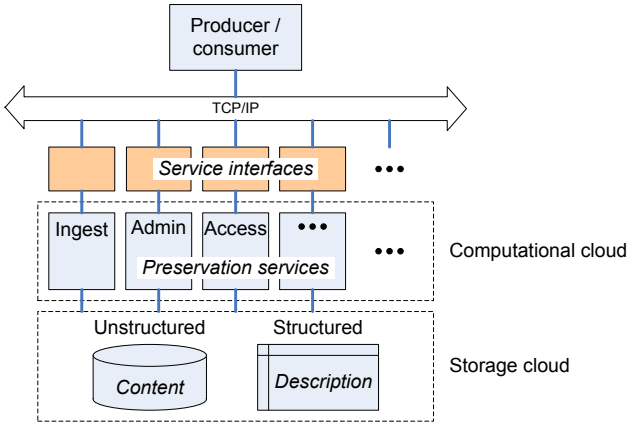

Figure 5 – Content-centric services architecture

Automated curation services are instantiated as self-contained units deployed in the context of an infrastructural fabric of ubiquitous connectivity and virtualized cloud computation and storage. In order to provide desired availability and responsivity, the computational cloud should support redundant web dispatching to service implementations that are easily and flexibly deployable in arbitrary number and location for automated load-balancing and failover (see Figure 6).

[24] H. M. Gladney, *Durable Digital Objects Rather Than Digital Preservation and Professional Implications*, preprint, May 15, 2008 <http://eprints.erpanet.org/149/01/Durable.pdf>.

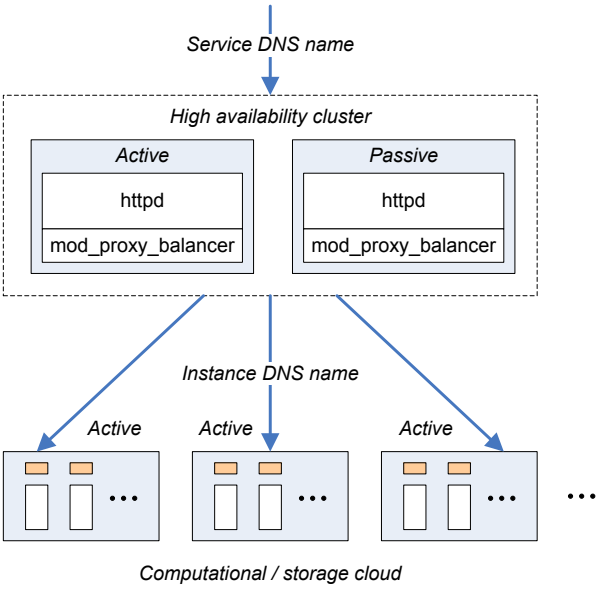

Figure 5 – High availability architecture

These infrastructural components should be designed and implemented with care to avoid unnecessary dependencies to underlying platform.

8.1 Information packaging and management

The internal Archival Information Package (AIP) for a digital object in this environment is an aggregate entity composed of one or more content files holding the formatted bit streams that collectively represent intrinsic object meaning and one of more formatted metadata files holding all extrinsic information – intellectual, administrative, technical, structural/relational, behavioral, provenancial, etc. – about that content (see Figure 6).

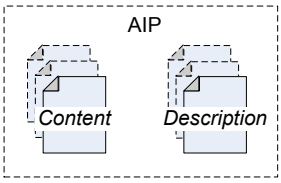

Figure 6 – Archival Information Package

From the perspective of curation services these files are managed as opaque containers in a file-level, unstructured storage abstraction; conventionally, a file system, ideally with a large uniform namespace.[25]

For operational convenience, some subset of object metadata may also be managed in a structured (or

[25] Linden, Jim, Sean Martin, Richard Masters, and Roderic Parker, *The Large-Scale Archival Storage of Digital Objects*, Digital Preservation Coalition Technology Watch Series Report 04-03, February 2005 <http://www.dpconline.org/docs/dpctw04-03.pdf>.

fielded) storage abstraction; conventionally, a relational or XML/RDF database. Regardless of this duplicative management, the metadata of record for an object is that found in the metadata files associated with that object. This is an important consideration supporting a business continuity requirement that all automated administrative and access services and systems built on top of managed content can be fully re-instantiated solely from the expression of data on the file system.

Accordingly, although object files are syntactically, semantically, and pragmatically opaque to the storage abstraction, it is important that object *coherence* – that is, the full set of files that compose an aggregate object – can be maintained by, and re-instantiated from, information managed solely by the storage abstraction.

Unstructured storage can be provided at various service levels defined by three independent facets:

Performance	:	Slow	⇔	Fast
Persistence	:	Transient	⇔	Permanent
Resilience	:	Unreliable	⇔	Dependable

Table 6 – Storage facets

In practice, not all possible positions within the three-dimensional functional space need to be provisioned. However, certain common service-level combinations are necessary to fulfill foreseeable preservation needs:

Staging	:	(Fast, Transient, Unreliable)
Managed	:	(Slow, Persistent, Quasi-dependable)
Dark archival	:	(Slow, Persistent, Dependable)
Bright archival	:	(Fast, Persistent, Dependable)

Table 7 – Storage service levels

In general there are two strategies for providing overall storage resilience:

- The use of enterprise, rather than commodity hardware, often incorporating local redundancy at the rack or component level.
- Data replication; in other words, global redundancy.

Recent research suggests that global replication using commodity components is the preferred solution to obtain the highest level of assurance at the lowest cost.[26] Note that redundancy is most effective when it

[26] Bernd Panzer-Steindel, *Data Integrity*, Draft 1.3, April 8, 2007 <http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797>; Baker, Mary, Mehul Shah, David S. Rosenthal, Mema Roussopoulos, Petros Maniatis, T. J. Guili, and Prashanth Bungale, "A fresh look at the reliability of long-term digital storage," *EuroSys '06*, Leaven, April 18-21, 2006, pp. 221-34; and Rosenthal, David S. H., "Bit preservation: A solved problem?," *Proceedings of the Fifth International Conference on Preservation of Digital Objects*, British Library, London, September 29-30, 2008, pp. 274-80.

is most uncorrelated.[27]

8.2 Curation micro-services

The individual automated curation services (known as "micro-services" in view of their fine granularity) execute in the context of a computational cloud and operate on objects managed in the storage cloud.[28] The scoping of these micro-services is based on the following REST-like principles (see Table 8): [29]

- *Granularity* and *orthogonality* of function.
- *Layering*; that is, the Unix/Linux pipe-like composition of existing atomistic function to provide complex behavior.[30]

Since each of the services is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance;[31] equally as important, since the level of investment in, and concomitantly, commitment to, any given service is small, they are more easily replaced when they have outlived their usefulness. Although the scope of any given service is narrowly focused, complex curation function can nevertheless *emerge* from the strategic combination of individual, atomistic services.[32]

Metaphors	***Preferences***	***Practices***
• Pipeline • Lego™ bricks	• Small and simple over large and complex • Minimally sufficient over feature-laden	• Define, decompose, recurse • Approach sufficiency through a series of incrementally necessary steps
Principles		
• Granularity • Orthogonality • Parsimony • Evolution	• Configurable over the prescribed • The proven over the merely novel • Outcomes over means	• Top-down design, bottom-up implementation • Code to interfaces

Table 8 – Micro-services

[27] See, for example, Eduardo Pinheiro, Wolf-Dietrich Weber, and Luiz André Barroso, "Failure trends in a large disk drive population," *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, San Jose, February 14-16, 2007; and Bianca Schroeder and Garth A. Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?" *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, San Jose, February 14-16, 2007.

[28] Moore, Reagan, "Towards a theory of digital preservation," *International Journal of Digital Curation* 3.1 (2008): 63-75 <http://www.ijdc.net/ijdc/article/view/63/82>.

[29] Roy Fielding and Richard Taylor, "Principled design of the modern web architecture," *ACM Transactions on Internet Technology* 2:2 (May 2002): 115-150 <doi:10.1145/514183.514185>

[30] Lucent Technologies, *The Creation of the UNIX Operating System: Connecting Streams Like a Garden House*, 2002 <http://www.bell-labs.com/history/unix/streams.html>.

[31] Peter J. Denning, Chris Gunderson, and Rich Hayes-Roth, "Evolutionary system development," *Communications of the ACM* 51:17 (December 2008): 29-31.

[32] David A. Fisher, *An Emergent Perspective on Interoperation in Systems of Systems*, CMU/SEI-2006-TR-003, ESC-TR-2006-03, March 2006 <http://www.sei.cmu.edu/pub/documents/06.reports/pdf/06tr003.pdf>.

The UC3 micro-services are known by the collective name "Merritt".[33] The Merritt services include:

- *Identity* service. The Identity service provides a means to uniquely identify various entities of curation interest.
- *Storage* service. The Storage service provides a means to manage digital object files.
- *Fixity* service. The Fixity service provides a means to verify the bit-level integrity of files managed by the Storage service.
- *Replication* service. The Replication service provides a means to provide a globally-fault tolerant storage environment.
- *Inventory* service. The Inventory service provides a means to associate various types of syntactic, semantic, and pragmatic descriptive information with digital objects.
- *Characterization* service. The Characterization service provides one possible source for descriptive information managed by the Catalog service.
- *Ingest* service. The Ingest service provides a means to add new digital content into the curation environment for active management by UC3.
- *Index* service. The Index service builds searchable indexes of object descriptive information and content.
- *Search* service. The Search service supports content request and delivery via index-based search and browse of managed content and description.
- *Transformation* service. The Transformation service provides a means to transcode digital object representations (that is, sets of files) from existing forms to newly required forms.
- *Publication* service. The Publication service provides a means to notify user communities that digital content is being managed and is available for use.
- *Annotation* service. The Annotation service provides a means for user-driven enrichment of managed object description.

It is expected that additional, added-value services will continue to be added to Merritt repertoire over time. The application of these services extends throughout the full digital curation life cycle (see Figure 7). The design of the services conforms to the high-level UC3 mission goals:

- Providing safety through redundancy.
- Maintaining meaning through description.
- Facilitating utility through service.
- Adding value through use.

and reflects the following high-level design goals:

- Open source deliverables.

[33] Koniklijke Bibliotheek, Deutsche Nationalbibliothek, and Niedersächsische- und Universitätsbibliothek Göttingen, *DPFuse: Report on Digital Preservation Functionalities & Services*, Version 4.0, November 10, 2008.

- Granularity and orthogonality (both *across* and *within* components).
- Persistent interfaces, evolving implementations.
- Complexity by composition, not addition.
- Flexible configuration, but meaningful default behavior ("Principle of least surprise").
- Strategic re-use ("Principle of least effort").

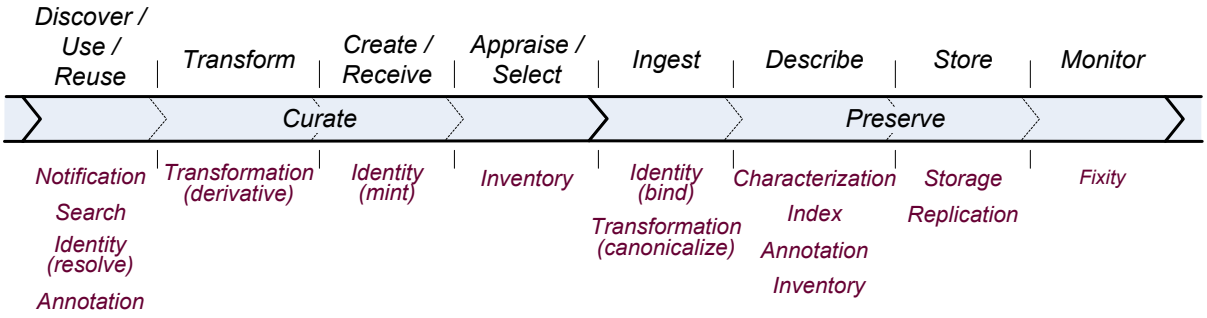

Figure 7 – Micro-service applicability throughout the curation lifecycle
[Adapted from Higgins]

Interaction with micro-services can occur in several modalities:

- *Procedural APIs*, with various language bindings.
- *Command line APIs* supported in various operating system command shells.
- *Web APIs*, composed of:
    - *Thin client GUIs*, supported in various browser platforms.
    - *RESTful service interfaces*.

The preferred language bindings for procedural APIs are Java and Perl. The preferred Java platform for RESTful APIs is Jersey (the reference implementation of JSR 311, *JAX-RS – Java API for RESTful web services*) running in a servlet container.

8.2.1 Identity service

An identifier is a persistent *association* between a unique character string and typed referents.[34] In the curation context there are three important identifier referent types:

- Information used to retrieve a representation of a digital object.
- Information descriptive of that object.
- Information descriptive of the identifier, that is, the string-referent association.

The first referent type will generally be to an actionable identifier, such as ARK, DOI, Handle, PURL, URI, URL, URN, etc. The other referents will generally be string-typed metadata elements.

The Identity service supports three fundamental operations:

- *Minting*. The process of generating a new identifier string according to the syntactic rules of an

[34] See, for example, California Digital Library, *NOID(1): Batch Identifier Infrastructure*, CDL 0.424, April 19, 2006 < http://www.cdlib.org/inside/diglib/ark/noid.pdf>.

identifier namespace.

- *Binding*. The association of an identifier string with a referent, which is itself a typed name/value pair.
- *Resolution*. The process of retrieving a requested identifier referent for a given identifier string.

The presumptive service identifier is the ARK namespace; the presumptive service implementation is the *noid* minting, binding, and resolution system.

8.2.2 Storage service

The Storage service provides unstructured, opaque storage to manage the file components of objects. Although the individual managed files are opaque to the service, object coherence – the fact that a set of files are all associated with a given object – will be maintained by the service.

The presumptive service implementation is a POSIX file system conforming to the CAN, Pairtree, D-flat, ReDD, and CLOP structuring conventions.[35] CAN (Content Access Node) defines a flexible file system-based object store. Pairtree uses a bigram decomposition of object identifiers to define a file system hierarchy with reasonable subdirectory fan-out of breadth and depth. D-flat provides a means to clearly indicate object data, metadata, versioning, and file-level parity for bit integrity error detection and correction. ReDD defines an efficient scheme for file-level compression of object versions. CLOP defines a class-based system for managing object properties.

8.2.3 Fixity service

The Fixity service verifies the bit-level integrity of files. The service should support a variety of commonly used digest types, including Adler-32, CRC-32, MD5, SHA-1, SHA-256, SHA-384, SHA-512, etc. The service should be capable of being configured to perform fixity verification on an arbitrary schedule.

8.2.4 Replication service

The Replication service creates and verifies the synchronization of replicas of digital objects managed in diverse Storage service instantiations.

8.2.5 Catalog service

The Catalog service maintains syntactic, semantic, and pragmatic information descriptive of digital

[35] IEEE 103.1-2008, *Information Technology – Portable Operating System Interface (POSIX)*; and J. Kunze, M. Haye, E. Hetzner, M. Reyes, and C. Snavely, *Pairtrees for Object Storage (V0.1)*, Internet draft, November 25, 2008 <http://www.ietf.org/internet-drafts/draft-kunze-pairtree-01.txt>; and Stephen Abrams, John Kunze, and David Loy, *D-flat: A Simple File System Convention for Object Storage*, Rev. 0.3, January 6, 2009.

objects and their files. (Note that the term "descriptive" is used here in its general sense, that is, information *about* something, and is not intended to specify strictly intellectual as opposed to technical, structural, administrative, etc. information.)

8.2.6 Characterization service

Characterization is information about a digital object that describes its character or significant nature.[36] The fundamental formal characterization property is *format*, a class of digital objects whose members share a common set of syntactic and semantic rules governing the encoding of abstract information content into tangible digital form. The process of creating characterization information has five aspects:[37]

- *Identification*. The determination of the purported format of an object on the basis of suggestive extrinsic hints and intrinsic internal and external signatures.
- *Validation*. The determination of the conformance of an object to the normative requirements of its format's specification.
- *Feature extraction*. The reporting of intrinsic properties of a digital object that can be used as a *surrogate* for the object itself for purposes of curation analysis and decision making.
- *Assessment*. The determination of the acceptability of a digital object for a specific purpose on the basis of local policy rules.

The Characterization service can be deployed in the following contexts:

- Client-side Submission Information Package (SIP) packaging
- Server-side ingest processing
- Post-transformation quality assurance testing
- Dissemination Information Package (DIP) packaging

The presumptive service implementation will use the JHOVE characterization framework (and JHOVE2 when it becomes available).[38]

8.2.7 Ingest service

The Ingest service accepts Submission Information Packages (SIPs) and converts them into AIPs. This process may involve the use of the Transformation service to transcode objects into normative forms defined by internal standards, the Characterization service to produce relevant descriptive information for

[36] Adrian Brown, "Developing practical approaches to active preservation," *International Journal of Digital Curation* 2.1 (2007): 3-11 <http://www.ijdc.net/ijdc/article/view/37/42>.

[37] Stephen Abrams, Sheila Morrisey, and Tom Cramer, "'What? So what?': The next-generation JHOVE2 architecture for format-aware characterization," *International Journal of Digital Curation* 4:3 (2009): 123-136 <http://www.ijdc.net/index.php/ijdc/article/viewFile/139/174>.

[38] Harvard University Library, *JHOVE – Harvard/JSTOR Object Validation Environment*, February 25, 2009 <http://hul.harvard.edu/jhove/>; and California Digital Library, *JHOVE2 Home: The Next-Generation Architecture for Format-Aware Characterization*, <http://confluence.ucop.edu/display/JHOVE2Info/Home>.

management in the Catalog service, and the Storage service to manage object files.

8.2.8 Index service

The Index service builds searchable indices of object descriptive information and content. The presumptive service implementation is Lucene.

8.2.9 Search service

The Search service supports content request and delivery via index-based search and browse of managed content and description. The delivery process may involve the use of the Transformation service to produce representations in the form preferred by the requesting user agent.

8.2.10 Transformation service

The Transformation service transcodes digital object representations from existing forms to newly required forms. The service can be deployed in the following contexts:

- Ingest canonicalization
- Preservation localization, desiccation, and migration
- Access derivation
- Export canonicalization

8.2.11 Publication service

The Publication service provides a means to notify user communities that digital content is being managed and is available for use. In furtherance of the goal implied by the JISC Common Repository Interfaces Group (CRIG) slogan – "The coolest thing to do with your data will be thought of by someone else" – the service will make use of various channels for publicizing available content managed by UC3:[39]

- Registration of managed digital objects in appropriate public discovery services and registries.[40]
- Exposing managed digital objects for harvesting by search engines.[41]

---

[39] JISC Common Repository Interfaces Group, *CRIG – DigiRepWiki*, October 7, 2008 <http://www.ukoln.ac.uk/repositories/digirep/index/CRIG>.

[40] See, for example, Adrian Burton and Chris Blackall, "The key role of registries and registry standards in the transition to a federated network of repositories," *3rd International Conference on Open Repositories*, Southampton, April 1,-4, 2008 <http://pubs.or08.ecs.soton.ac.uk/29/1/submission_87.pdf>; and John Mark Ockerbloom, "Promoting discovery and use of repository content: An architectural perspective," *Partnerships in Innovations (2008) II: From Vision to reality and Beyond*, College Park, October 7-8, 2008 <http://works.bepress.com/cgi/viewcontent.cgi?article=1006&context=john_mark_ockerbloom>.

[41] Roberta Fox, Michael Vandermillen, and Spencer McEwen, "Deep web content and internet discovery: Exposing Harvard University Library's digital resources to search engines," *DLF Fall Forum 2008*, Providence, November 10-12, 2008 <http://www.diglib.org/forums/fall2008/presentations/Fox.pdf>.

- Exposing managed digital objects for harvesting by OAI clients.[42]
- Syndicated notification of newly available objects.[43]

8.2.12 Annotation service

The Annotation service provides a means for user-driven enrichment of managed object description. The service will make use of various mechanisms for enabling user-driven annotation, enrichment, and aggregation:

- Social tagging.[44]
- OAI-ORE-based aggregation.[45]

[42] Herbert Van de Sompel, Michael L. Nelson, Carl Lagoze, and Simeon Warner, "Resource harvesting within the OAI-PMH framework," *D-Lib Magazine* 10.12 (December 2004) <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>.

[43] Annika Hinze, Andrea Schweer, and Geoge Buchanan, "An integrated alerting service for open digital libraries: Design and implementation," *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODCASE*, Agia Napa, Cyprus, October 31-November 4, 2005, in Robert Meersman *et al*. (eds.), *Lecture Notes in Computer Science* 3760 (Berlin: Springer, 2005): 484-501 <http://www.springerlink.com/content/48cpwtk2hbvt4vdn/fulltext.pdf>.

[44] Umer Faroog, Yang Song, John M. Carroll, and C. Lee Giles, "Social bookmarking for scholarly digital libraries," *IEEE Internet Computing* 11.6 (November 2007): 29-35 <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04376225>; and Rich Gazan, "Social annotations in digital library collections," *D-Lib Magazine* 14.11/12 (November/December 2008) <http://www.dlib.org/dlib/november08/gazan/11gazan.html>.

[45] Carl Lzgoze, Herbert Van de Sompel, Michael L. Nelson, Simeon Warner, Robert Sanderson, and Peter Johnson, *Object Re-Use & Exchange: A Resource-Centric Approach*, April 14, 2008, arXiv:0804.2273v1[cs.DL] <http://arxiv.org/ftp/arxiv/papers/0804/0804.2273.pdf>.

9 Conclusions

> *Entia non sunt multiplicanda praeter necessitatem* [Entities must not be multiplied beyond necessity].
>
> – William of Occam

> The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.
>
> – Albert Einstein

> *Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte* [I made this very long, because I did not have the leisure to make it shorter].
>
> – Blaise Pascal, *Lettres provinciales*

> How difficult it is to be simple.
>
> – Vincent van Gogh

In order to maintain its continuing relevance and viability within the University system, the UC Curation Center is pursuing a process of self-reflection and evaluation resulting in a series of significant transformations to its fundamental outlook, effort, and infrastructure. UC3 is engaged broadly in the programmatic curation of digital assets central to the University's pedagogic mission. Curation goals are best served by concentrating on long-lived content – sustained by a constantly evolving repertoire of nimble, commodified services – rather than somewhat illusory long-lived systems.[46] The cumulative effect of these transformations will better enable UC3 to remain responsive to the ever expanding needs of its University stakeholders, and to collaborate more effectively with the broader curation community.

Throughout this process the work of UC3 will proceed along four parallel strands of effort applicable across the full digital asset lifecycle (see Table 10):

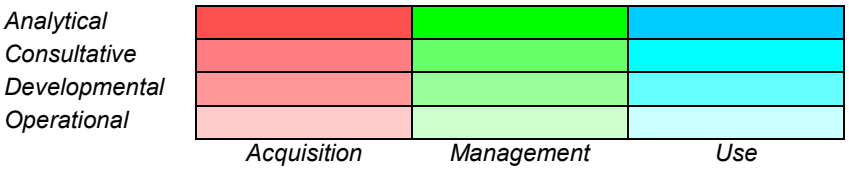

Table 10 – Programatic activity

- *Analytical*. Determining what activities UC3 should undertake.
- *Consultative*. Applying analysis to the problems of stakeholder communities and suggesting best practices for the safe, secure, and enriching management of their digital assets.
- *Developmental*. Embedding human analysis and decision making into efficient and effective

[46] Andreas Aschenbrenner, Tobias Blanke, Favid Flanders, Mark Hedges, and Ben O'Steen, "The future of repositories? Patterns for (cross-)repository architectures," *D-Lib Magazine* 14.11/12 (November/December 2008) <http://www.dlib.org/dlib/november08/aschenbrenner/11aschenbrenner.html>.

automated systems.

- *Operational*. Deploying these services and systems for use by stakeholder communities.

Taking the well known epigrams of William of Occam and Einstein to heart, the conceptual framework developed this paper started by considering the question, How simple can a curation environment be and still be effective? Simplicity was the goal of the deliberative multi-stage design proces described above (*value → strategy → service → system*), in which technical decisions are deferred until curation intentions and outcomes are clearly understood and well documented. The resulting 12 Merritt micro-services build from a base of minimal necessity and extend towards increasing sufficiency (see Figure 8). Taken together, they provide the baseline function needed to meet UC3 obligations for effective, efficient, and sustainable curation services.

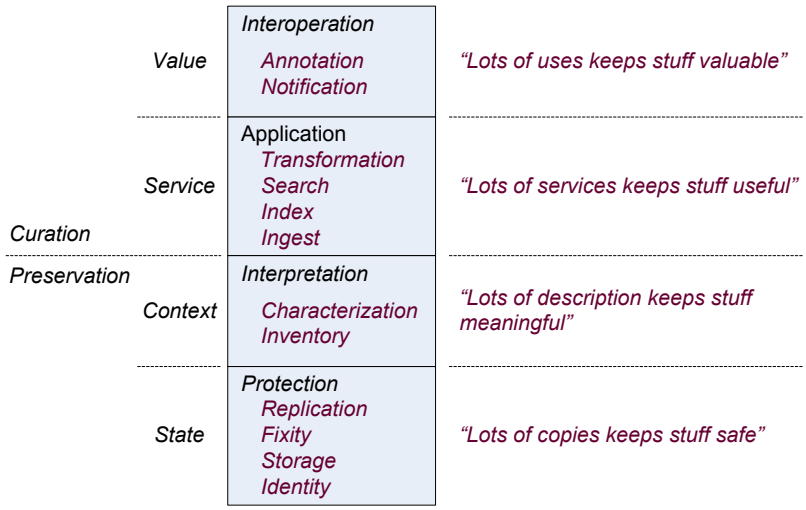

Figure 8 – Merritt micro-services

The services can be aggregated into four service levels – *protection*, *interpretation*, *application*, and *interoperation*, respectively providing maintenance of content *state* and *context*, and provision of content *service* and *value* – which suggests a pithy summarization of UC3 intentions in terms of four rather simple aphorisms:

- Lots of copies keeps stuff safe.[47]
- Lots of description keeps stuff meaningful.
- Lots of services keeps stuff useful.[48]
- Lots of uses keeps stuff valuable.[49]

[47] See, for example, Vicky Reich and David S. H. Rosenthal, "LOCKSS: A Permanent Web Publishing and Access System", *D-Lib Magazine* 7.6 (June 2001) <http://www.dlib.org/dlib/june01/reich/06reich.html>.

[48] See, for example, Gregory Crane, "On scholarly access to a million books," *Using Digital Collecitons: Open Content Alliance Annual Meeting*, San Francisco, October 27-28, 2008.

[49] See, for example, Erv Blyth and Vinod Chachra, "The value proposition in institutional repositories," *EDUCAUSE Review* 40.5 (September/October 2005): 76-77 <http://connect.educause.edu/Library/EDUCAUSE+Review/TheValuePropositioninInst/40584>; and Herbert Van de Sompel, Xiaoming Lui, Carl Lagoze, Sandy

Rendundancy ("lots of copies") is the key principle for ensuring the safety and accessibility of curated assets and the availability of services built around those assets; abundant characterization ("lots of description") is the key principle for exposing content to user communities in useful contexts; responsiveness to the needs of users ("lots of services") is the key principle for ensuring the widespread integration of curated assets into the research, teaching, and learning activities of the University; and the multiplier effect of the creative use and re-use of curated assets ("lots of uses") is the key principle for enriching that discourse.

Work on the micro-services will progress in a series of incremental development waves, with siginficant milestones at the completion of the second, fourth, and sixth waves (see Table 10).

First wave	*Second wave* ✓	*Third wave*	*Fourth wave* ✓	*Fifth wave*	*Sixth wave* ✓
Identity	Inventory	Index	Search	Notification	Annotation
Storage	Ingest	Fixity	Replication	Characterization	Transformation
Object and collection modeling			Authentication and authorization		
Policy and business model development					

Table 10 – Micro-service development plan

NOTE While this table presents a *logival* view of high-level priorities and dependencies, it should not be interpreted literally as a *schedule*. The 12 micro-services vary widely in scope and complexity and this variance will be reflected in the time necessary for individual service implementation. Additionally, parallelism and pipelining beyond what is represented above may be possible subject to the overall schedule and availability of appropriate resources.

It is important to note that what has been described above is only a programmatic and architectural *approach* to the curation problem, not a technical specification for its solution. Apropos of Pascal, a truly simple solution will only result through the diligent application of significant time and effort.

Payette, Simeon Warner, and Jeroen Bekaert, "An interoperable fabric for scholarly value chains," *D-Lib Magazine* 12.10 (October 2006) <doi:10.1045/october2006-vandesompel>.

References

Abbot, Daisy, *What is Digital Curation?* April 3, 2008 <http://www.dcc.ac.uk/resource/briefing-papers/what-is-digital-curation/>

Abrams, Stephen, Patricia Cruse, and John Kunze, "Preservation is not a place," *International Journal of Digital Curation* 4:1 (2009): 8-21 <http://www.ijdc.net/index.php/ijdc/article/viewFile/98/73>.

Abrams, Stephen, Sheila Morrissey, and Tom Cramer, "'What? So what?': The next-generation JHOVE2 architecture for format-aware characterization," *International Journal of Digital Curation* 4:3 (2009): 123-136 <http://www.ijdc.net/index.php/ijdc/article/viewFile/139/174>.

Abrans, Stephen, John Kunze, and David Loy, *D-flat: A Simple File System Convention for Object Storage*, Rev. 0.3, January 6, 2009

ALCTS, *Definitions of Digital Preservation*, June 24, 2007 <http://www.ala.org/ala/alcts/newslinks/digipres/index.cfm>.

Aschenbrenner, Andreas, Tobias Blanke, Favid Flanders, Mark Hedges, and Ben O'Steen, "The future of repositories? Patterns for (cross-)repository architectures," *D-Lib Magazine* 14.11/12 (November/December 2008) <http://www.dlib.org/dlib/november08/aschenbrenner/11aschenbrenner.html>.

Baker, Mary, Mehul Shah, David S. Rosenthal, Mema Roussopoulos, Petros Maniatis, T. J. Guili, and Prashanth Bungale, "A fresh look at the reliability of long-term digital storage," *EuroSys '06*, Leaven, April 18-21, 2006, pp. 221-34.

Blyth, Erv, and Vinod Chachra, "The value proposition in institutional repositories," *EDUCAUSE Review* 40.5 (September/October 2005): 76-77 <http://connect.educause.edu/Library/EDUCAUSE+Review/TheValuePropositioninInst/40584>.

Brown, Adrian, "Developing practical approaches to active preservation," *International Journal of Digital Curation* 2.1 (2007): 3-11 <http://www.ijdc.net/ijdc/article/view/37/42>.

Burton, Adrian, and Chris Blackall, "The key role of registries and registry standards in the transition to a federated network of repositories," *3rd International Conference on Open Repositories*, Southampton, April 1,-4, 2008 <http://pubs.or08.ecs.soton.ac.uk/29/1/submission_87.pdf>.

California Digital Library, *JHOVE2 Home: The Next-Generation Architecture for Format-Aware Characterization*, <http://confluence.ucop.edu/display/JHOVE2Info/Home>.

California Digital Library, *Merritt: An Emergent Micro-services Approach to Digital Curation Infrastructure*.

California Digital Library, *NOID(1): Batch Identifier Infrastructure*, CDL 0.424, April 19, 2006 < http://www.cdlib.org/inside/diglib/ark/noid.pdf>.

Christensen, Erik, Francisco Curbera, Greg Meredith, and Sanjiva Weerawarana, *Web Services Description Language (WSDL) 1,1*, W3C Note, March 15, 2001 <http://www.w3.org/TR/wsdl>.

Cornell University Library, *Digital Preservation Management: Implementing Short-term Strategies for Long-Term Problems*, 2007 <http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html

Cornell University Library, *Digital Research Data Curation: Overview of Issues, Current Activities, and Opportunities for the Cornell University Library*, May 2008 <http://ecommons.library.cornell.edu/handle/1813/10903>.

Currall, James, "The fetish of the digital," *JISC-PoWR blog*, January 7, 2009

<http://jiscpowr.jiscinvolve.org/2009/01/07/the-fetish-of-the-digital/>.

Dale, Robin, and Bruce Ambacher, eds. *Trustworthy Repositories: Audit & Certification Criteria and Checklist*, Version 1.0, February 2007 <http://www.crl.edu/PDF/trac.pdf>.

Denning, Peter J., Chris Gunderson, and Rick Hayes-Roth, "Evolutionary system development," *Communications of the ACM* 51.17 (December 2008): 29-31

Digital Curation Centre, *About the DCC: What is Digital Curation?* April 26, 2007 <http://www.dcc.ac.uk/about/what/>.

Factor, Sue, *What makes a design "Googley"?*, *The Official Google blog*, posted April 23, 2008, <http://googleblog.blogspot.com/2008/04/what-makes-design-googley.html>.

Faroog, Umer, Yang Song, John M. Carroll, and C. Lee Giles, "Social bookmarking for scholarly digital libraries," *IEEE Internet Computing* 11.6 (November 2007): 29-35 <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=04376225>.

Farquhar, Adam, and Helen Hockx-Yu, "Planets: Integrated services for digital preservation," *International Journal of Digital Curation* 2:2 (November 2007) <http://www.ijdc.net/ijdc/article/view/46/59>.

Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, *Hypertext Transfer Protocol – HTTP/1.1*, RFC 2616, June 1999 <www.ietf.org/rfc/rfc2616.txt>.

Fielding, Roy, and Richard Taylor, "Principled design of the modern web architecture," *ACM Transactions on Internet Technology* 2:2 (May 2002): 115-150 <doi:10.1145/514183.514185>.

Fisher, David A., *An Emergent Perspective on Interoperation in Systems of Systems*, CMU/SEI-2006-TR-003, ESC-TR-2006-03, March 2006 <http://www.sei.cmu.edu/pub/documents/06.reports/pdf/06tr003.pdf>.

Fox, Roberta, Michael Vandermillen, and Spencer McEwen, "Deep web content and internet discovery: Exposing Harvard University Library's digital resources to search engines," *DLF Fall Forum 2008*, Providence, November 10-12, 2008 <http://www.diglib.org/forums/fall2008/presentations/Fox.pdf>.

Gazan, Rich, "Social annotations in digital library collections," *D-Lib Magazine* 14.11/12 (November/December 2008) <http://www.dlib.org/dlib/november08/gazan/11gazan.html>.

Gladney, H. M. *Durable Digital Objects Rather Than Digital Preservation and Professional Implications*, preprint, May `5, 2008 <http://eprints.erpanet.org/149/01/Durable.pdf>.

Hamilton, James, "On Designing and Deploying Internet-Scale Services," *Proceedings of the 21st Large Installation System Administration Conference (LISA '07)*, Dallas, November 11-16, 2007, pp. 231-42 <http://www.usenix.org/event/lisa07/tech/full_papers/hamilton/hamilton_html/index.html>.

Harvard University Library, *Digital Repository Service (DRS) Policy Guide*, April 26, 2007 <http://hul.harvard.edu/ois/systems/drs/policyGuide>.

Harvard University Library, *JHOVE – Harvard/JSTOR Object Validation Environment*, February 25, 2009 <http://hul.harvard.edu/jhove>.

Hedstrom, Margaret, and Christopher A. Lee, "Significant properties of digital objects: definitions, applications, implications," *Proceedings of the DLM-Forum 2002: @ccess and Preservation of Electronic Information: Best Practices and Solutions*, Barcelona, May 6-8, 2002, pp. 218-227 <http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf

Hendley, Tony, *Comparison of Methods and Costs of Digital Preservation*, British Library Research and Innovation Report 106, 1998 <http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html#_Toc422714267

Heslop, Helen, Simon Davis, and Andrew Wilson, *An Approach to the Preservation of Digital Records*, National Archives of Australia Green Paper, Canberra, December 2002 <http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf>.

Higgins, Sarah ,"The DCC curation lifecycle model," *International Journal of Digital Curation* 1.3 (2008): 134-140 <http://www.ijdc.net/ijdc/article/view/69/69>.

Hinze, Annika, Andrea Schweer, and Geoge Buchanan, "An integrated alerting service for open digital libraries: Design and implementation," *On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODCASE*, Agia Napa, Cyprus, October 31-November 4, 2005, in Robert Meersman *et al*. (eds.), *Lecture Notes in Computer Science* 3760 (Berlin: Springer, 2005): 484-501 <http://www.springerlink.com/content/48cpwtk2hbvt4vdn/fulltext.pdf>.

IEEE 103.1-2008, *Information Technology – Portable Operating System Interface (POSIX)*.

IFLA, *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications – New Series Vol. 19 (Munich: K. G. Saur, 1998) <http://www.ifla.org/VII/s13/frbr/frbr.pdf>.

ISO 14721:2003, *Space data and information transfer systems – Open archival information system – Reference model*.

ISO/PDTR 15801, *Document management – Information stored electronically – Recommendations for trustworthiness and reliability*, July 3, 2008.

Knight, Gareth, *Framework for the Definition of Significant Properties*, version 1, May 2, 2008 <http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf>.

Koniklijke Bibliotheek, Deutsche Nationalbibliothek, and Niedersächsische- und Universitätsbibliothek Göttingen, *DPFuse: Report on Digital Preservation Functionalities & Services*, Version 4.0, November 10, 2008.

Kunze, J., M. Haye, E. Hetzner, M. Reyes, and C. Snavely, *Pairtrees for Object Storage (V0.1)*, Internet draft, November 25, 2008 <http://www.ietf.org/internet-drafts/draft-kunze-pairtree-01.txt>.

Lagoze, Carl, Herbert Van de Sompel, Michael L. Nelson, Simeon Warner, Robert Sanderson, and Peter Johnson, *Object Re-Use & Exchange: A Resource-Centric Approach*, April 14, 2008, arXiv:0804.2273v1[cs.DL] <http://arxiv.org/ftp/arxiv/papers/0804/0804.2273.pdf>.

Lavoie, Brian F. "The Fifth Blackbird: Some Thoughts on Economically Sustainable Digital Preservation," *D-Lib Magazine* 14:3/4 (March/April 2008) <http://www.dlib.org/dlib/march08/lavoie/03lavoie.html>.

Lavoie, Brian, and Lorcan Dempsey, "Thirteen ways of looking at … digital preservation," *D-Lib Magazine* 10.7/8 (July/August 2004) <http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>.

Liegl, Philipp, "The strategic impact of service oriented architectures," *Proceedings of the 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS '07)*, Tucson, March 26-29, 2007

Linden, Jim, Sean Martin, Richard Masters, and Roderic Parker, *The Large-Scale Archival Storage of Digital Objects*, Digital Preservation Coalition Technology Watch Series Report 04-03, February 2005 <http://www.dpconline.org/docs/dpctw04-03.pdf>.

Lucent Technologies, *The Creation of the UNIX Operating System: Conecting Streams Like a Garden House*, 2002 <http://www.bell-labs.com/history/unix/streams.html

McHugh, Andrew, Raivo Ruusalepp, Seamus Ross, and Hans Hoffman, *Digital Repository Audit Method Based on Risk Assessment*, Version 1.0 (draft), February 28, 2007, Digital Curation Centre/Digital Preservation Europe <http://www.repositoryaudit.eu/>.

Misra, Subhas C., and Virenda C. Bhavsar, "Relationships between selected software measures and latent bug-density: Guidelines for improving quality," *Computational Science and Its Applications – ICCSA 2003*, Montreal, May 18-21, 2003, in V. Kumar, M. L. Gavrilova, C. J. K. Tan, and P. L'Ecuyer, eds., *Lecture Notes in Computer Science* 2667 92003): 724-32 <http://www.springerlink.com/content/4rcnfqqtn3fvetrh/fulltext.pdf>

Moore, Reagan, "Towards a theory of digital preservation," *International Journal of Digital Curation* 3.1 (2008): 63-75 <http://www.ijdc.net/ijdc/article/view/63/82>.

Morris, Charles W., *Foundations of the Theory of Signs* (Chicago: University of Chicago Press, 1938).

Ockerbloom, John Mark, "Promoting discovery and use of repository content: An architectural perspective," *Partnerships in Innovations (2008) II: From Vision to reality and Beyond*, College Park, October 7-8, 2008 <http://works.bepress.com/cgi/viewcontent.cgi?article=1006&context=john_mark_ockerbloom>.

Panzer-Steindel, Bernd, *Data Integrity*, Draft 1.3, April 8, 2007 <http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797>.

Pinheiro, Eduardo, Wolf-Dietrich Weber, and Luiz André Barroso, "Failure trends in a large disk drive population," *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, San Jose, February 14-16, 2007.

Ranganathan, S. R., *The Five Laws of Library Science* (Madras, 1931).

Reich, Vicky, and David S. H. Rosenthal, "LOCKSS: A Permanent Web Publishing and Access System", *D-Lib Magazine* 7.6 (June 2001) <http://www.dlib.org/dlib/june01/reich/06reich.html>.

Rochberg-Halton, Eugen, and Kevin McMurtrey, "The foundations of modern semiotic: Charles Peirce and Charles Morris," *American Journal of Semiotics* 2.1-2 (1983): 129-56.

Rosenthal, David S. H., "Bit preservation: A solved problem?," *Proceedings of the Fifth International Conference on Preservation of Digital* Objects, British Library, London, September 29-30, 2008, pp. 274-80.

Rosenthal, David S. H., Thomas Robertson, Tom Lipkis, Vicky Reich, and Seth Morabito, "Requirements for Digital Preservation Systems: A Bottom-up Approach," *D-Lib Magazine* 11.11 (November 2005) <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>.

Ross, Seamus, "Digital preservation, archival science, and methodological foundations for digital libraries," *ECDL 2007: The 11th European Conference on Research and Advanced Technology for Digital Libraries*, Budapest, September 16-21, 2007.

Chris Rushbridge, "'Digital Preservation' term considered harmful?" Digital Curation Blog, July 29, 2008 <http://digitalcuration.blogspot.com/2008/07/digital-preservation-term-considered.html>.

Schroeder, Bianca, and Garth A. Gibson, "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?" *Proceedings of the 5th USENIX Conference on File and Storage Technologies*, San Jose, February 14-16, 2007.

Stephens, Owen, "Thinking we need to stop thinking of the 'repository' as a 'place'," Twitter, July 3, 2008 <http://twitter.com/ostephens?page=7>.

Strodl, Stephen, Christoph Becker, Robert Neumayer, and Andreas Rauber, "How to choose a digital preservation strategy: Evaluating a preservation planning procedure," *7th ACM/IEEE-CS Joint Conference on Digital Libraries*, Vancouver. June 17-23, 2007.

Van de Sompel, Herbert, Xiaoming Lui, Carl Lagoze, Sandy Payette, Simeon Warner, and Jeroen Bekaert, "An interoperable fabric for scholarly value chains," *D-Lib Magazine* 12.10 (October 2006)

<doi:10.1045/october2006-vandesompel>.

Van de Sompel, Herbert, Michael L. Nelson, Carl Lagoze, and Simeon Warner, "Resource harvesting within the OAI-PMH framework," *D-Lib Magazine* 10.12 (December 2004) <http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html>.