Permanent Objects, Disposable Systems

Stephen Abrams, Patricia Cruse, John Kunze

The California Digital Library (CDL) preservation program is re-envisioning its curation infrastructure as a set of loosely-coupled, distributed micro-services. There are many monolithic systems that support a range of preservation activities but also require the user and the hosting institution to buy-in to a particular system culture. The result is an institution that becomes, say, a DSpace, Fedora, or LOCKSS "shop", with a specific worldview and set of object flows and structures that will eventually need to be abandoned when it comes time to transition to the next system. Experience shows that these transitions are unavoidable, despite claims that once an object is “in” the system, it will be safe forever.

In view of this it is safer and more cost-effective to acknowledge from the outset the inevitable transient nature of systems and to plan on managing, rather than resisting change. The disruption caused by change can be mitigated by basing curation services on simple universal structures and protocols (e.g., filesystems, HTTP) and micro-services that operate on them. We promote a “mix and match” approach in which appropriate content- and context-specific curation workflows can be nimbly constructed by combining necessary functions drawn from a granular set of independent micro-services. Micro-services, whether deployed in isolation or in combination, are especially suited to exploitation upstream towards content creators who normally don’t want to think about preservation, especially if it’s costly; compared to buying into an entire curation culture, it is easy to adopt a small, inexpensive tool that requires very little commitment.

We see digital curation as an ongoing process of enrichment at all stages in the lifecycle of a digital object. Because the early developmental stages are so critical to an object's health and longevity, it is desirable to push curation "best practices" as far upstream towards the object creators as possible. If preservation is considered only when objects are close to retirement, it is often too late to correct the structural and semantic deficiencies that can impair object usability. The later the intervention, the more expensive the correction process, and it is always difficult to fund interventions for "has been" objects.

In contrast, early stage curation challenges traditional practices. Traditionally, preservation actions are often based on end-stage processing, where objects are deposited "as is" and kept out of harm's way by limiting access (i.e., dark archives). While some systems are designed to be dark or “dim”, with limited access and little regard for versioning or object enrichment, enrichment and access are now seen as necessary curation actions, that is, interventions for the sake of preservation. In particular, the darkness of an entire collection can change in the blink of an eye, for example, as the result of a court ruling or access rights purchase; turning the lights on for a collection should be as simple as throwing a switch, and not require transferring the collection from a "preservation repository" to an "access repository". Effective curation services must be flexible and easily configurable in order to respond appropriately to the wide diversity of content and content uses.

To be most effective, not only should curation practices be pushed upstream but also they should be pushed out to many different contexts. The micro-services approach promotes the idea that curation is an outcome, not a place. Curation actions should be applied to content where it most usefully exists for the convenience of its creators or users. For example, high value digital assets in access repositories, or even scholars’ desktops, would certainly benefit from such things as

persistent identification or regular audits to discover and repair bit-level damage, functions usually available only in the context of a "preservation system" but now easily applied to content where it most usefully resides without requiring transfer to a central location.

The design the CDL curation infrastructure is predicated on the following assumptions:

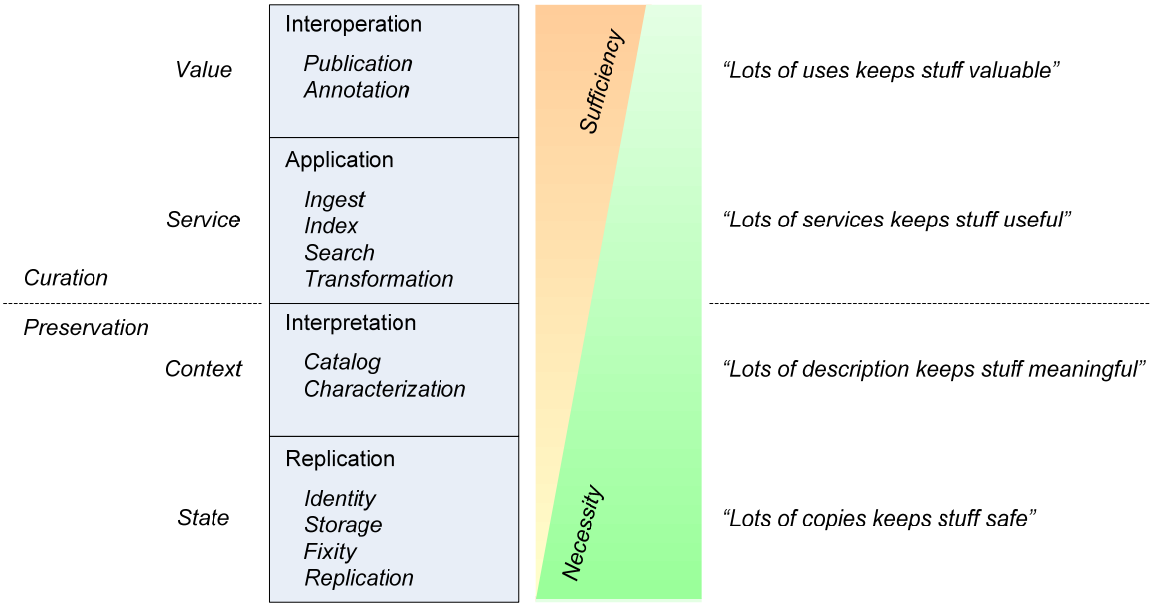
- An accelerating diversity of content types and behavioral expectations regarding the use of that content
- A growing demand for curation of content originating in non-library-centric contexts
- The inevitability of disruptive technological change

In response to the first assumption, the curation infrastructure is designed to be inherently flexible and responsive to changing conditions and user needs. As a consequence of the second assumption the functions provided by the infrastructure are deployable in contexts beyond traditional workflows. In particular, curation services are applicable throughout the full digital asset lifecycle and make minimal demands on sophisticated computing environments or skills. In view of the third assumption the infrastructure is designed to evolve gracefully over time without disruption of service function.

The range of curation function provide by the infrastructure is suggested by the following imperatives reflective of community best practices:

- Lots of copies keeps stuff safe
- Lots of description keeps stuff meaningful
- Lots of services keeps stuff useful
- Lots of uses keeps stuff valuable

As a consequence, the CDL infrastructure is organized into four conceptual levels, providing assurance of digital content state through distributed replication, content context through interpretive description, service through comprehensive application, and value through free interoperation. Replication ensures the safety and accessibility of curated content and the availability of services built around that content; abundant description allows content to be exposed to user communities in meaningful contexts; user-facing services facilitate the widespread integration of curated assets into the discourse of the University; and the multiplier effect of the creative use and re-use of curated content enriches that discourse.



The individual micro-services include:

- *Identity*. To identify uniquely digital objects of curation interest.
- *Storage*. To manage the files that express the abstract information content of digital objects.
- *Fixity*. To verify the bit-level integrity of files managed by the Storage service and, when necessary, repair any such damage that is uncovered.
- *Replication*. To provide a globally-fault tolerant storage environment.
- *Catalog*. To associate various types of syntactic, semantic, and pragmatic descriptive information with digital objects and their files.
- *Characterization*. To automatically derive pertinent descriptive information about objects for management by the Catalog service.
- *Ingest*. To add new digital content into the curation environment for active management.
- *Index*. To build searchable indexes of object descriptive information and content in support of the Search service.
- *Search*. To support content request and delivery via index-based search and browse of managed content and description.
- *Transformation*. To derive new forms of object content from existing forms, as requirements change.
- *Publication*. To notify users that digital content is being managed and is available for use.
- *Annotation*. To enrich managed object through user-supplied description and comment.

The envisioned micro-services are only partially built out. Our presentation will report on working real-world examples, including applications from the California Digital Library and HathiTrust. We will also describe work in progress, with reference to publically available specifications and software.