Towards a Design Space for a Commons Provenance System

DRAFT 2015-01-21

Tessa Askamp, Paul Keller, Mike Linksvayer, Catharina Maracke, Maarten Zeinstra

Published by Project Octopus <http://project-octopus.org>

Please send feedback to hello@project-octopus.org

In this paper we examine the demand for and attempts to provide provenance systems for the digital commons and the design space for a new provenance system. We find that direct attempts have a poor track record, particularly those conceptualized primarily as copyright registries. But much useful provenance information concerning the digital commons is provided as a side effect of hosting, developing, and curating works, particularly within mass collaboration communities. We also note that there is much provenance and use information concerning works in the digital commons on the web that is not aggregated as such anywhere. We conclude that a web observatory is a high potential design frame for a new provenance system powered by mass collaboration with growing assistance from machine analysis and bots. Such an observatory could provide the reliable provenance information, access to heretofore lacking metrics and discovery mechanisms for the digital commons, and provide a sound basis for adding features such as content-derived lookup, whitelists, and registration.

Table of Contents

Introduction

This paper attempts to identify a design space for a provenance system yet to be developed by the authors under the working title Project Octopus.

Provenance systems (often under the name “digital copyright registries”) have been put forth as a solution to several challenges faced by the digital commons. These challenges include identifying works, credit, license, or copyright status of works, and uses and remixes of works — all with a level of reliability that makes it practical for others to make, host, and promote further uses of the digital commons, including in long-lived, institutional and commercial contexts where reliable provenance is required. The paper analyzes existing and abandoned efforts to create a provenance system for the digital commons: both where these efforts have been explicitly aimed at creating a provenance system or where the provenance system is a by-product of another service.

We acknowledge extensive work on provenance for free and open source software[1] and open data,[2] both of which are vital components of the digital commons. Further we do not endorse a conception of “culture”, “data”, and “software” as non-overlapping magisteria. However we are limiting our focus in this paper to “culture”; in part this reflects the extent to which existing work has treated these domains separately, and in part a practical matter of limiting our scope. This leaves an opportunity for future distillation of provenance practices and lessons from free and open source software and open data that ought be applied to cultural provenance.

While this paper is scoped to focus on provenance for the cultural digital commons, some of the analysis may apply to the design of provenance systems which are focused on proprietary works, especially where those provenance systems are built for the web. One dimension of cross-cutting analysis involves commons methodologies (e.g., open data, open source, open governance) *for provenance systems*, in contrast with the focus of this paper: provenance methodologies *for digital commons*. We hope to address these other dimensions in separate, more legal- and policy-oriented, papers.

Definitions

- **Provenance.** Information about origins, uses, and the legal status of works and digital objects. For example creator, creation date, sources and derivations, publication dates, locations and licenses, and use metrics.
- **Provenance system.** Any system that is used to obtain provenance of works or digital objects. For example — from degree of intentionality to use as a provenance system: a copyright registry, an archive, a general web search engine.
- **Digital object.** Any digital item (files and streams). Synonymous with manifestation.
- **Work.** An intellectual creation, which may be subject to copyright or similar restrictions, or in the public domain. A work may be embodied in any number of different digital objects (see above).[3]
- **Digital Commons.** Works that are substantially free[4] of copyright, related restrictions and communities/individuals/institutions/services[5] which create and curate such works.

---

[1] For example http://www.fossology.org/ but more often as a side effect of packaging software for “distributions”, e.g., through one lens, work on reproducible builds is work on increasing the reliability of provenance information about software; see https://wiki.debian.org/ReproducibleBuilds

[2] For example http://www.gbif.org/infrastructure/summary and generally efforts to enable “data citation” such as https://www.datacite.org

[3] This is an important distinction for provenance systems. In some cases (e.g. “lookup”) the subject is a digital object, the objective is to determine what works the digital object embodies. In other cases, the work is the subject: what terms is the work subject to, who created the work, etc. Note digital objects might also be associated with agents and processes: who encoded a specific file, where is that specific file published, etc. Finer distinctions can be drawn (e.g. Functional Requirements for Bibliographic Records’ (FRBR) work/expression/manifestation/item hierarchy) but for our initial purposes digital object and work suffice. Note that “work” also is the word typically used for, and corresponds with, the coarsest granularity of subject for copyright. From the perspective of rights management, some fields (e.g. music) may require granularity between “work” and “digital object” (copies); the extent to which this perspective forms a design requirement it should be explained in Section 1.

[4] Intentionally vague. We may wish to initially focus on public domain or Open Definition compliant works, but a provenance system for the commons ought be helpful for works with less substantial freedom (e.g. noncommercial-only permissions) or even entirely different reasons for freedom (e.g. long record of web availability with no take down).

[5] “No commons without commoners.”

- **Collective Management organizations (CMOs).** Collective Management Organizations or Collecting Societies exercise copyright and related rights on behalf of the owners of rights.[6]
- **Intermediaries.** Entities that do not hold or represent their own collections but act as an conduit between other stakeholders.

Stakeholder Definitions

To assess different functionalities on their demands and usefulness, we will be using the following categories of stakeholders.[7]

- **Creators**. All makers of original works, including individuals who reuse works originally created by others. Synonymous with “Authors.”
- **Right holders.** Those other than creators who hold rights, e.g. publishers.
- **Users.** Individuals or Institutions who want to make use of works and digital objects, and would require a request for permission if content is subject to copyright.
- **GLAMs***:* Galleries, Libraries, Archives and Museums. Synonymous with “Cultural Heritage Institutions.”
- **Consumers.** End users that use works on the Internet without creating a new work.
- **Platforms.** Intermediaries that host or process content online, e.g. Flickr or YouTube.
- **Commons advocates.** Commons policy advocates and organizations.[8]

The remainder of this document reviews (1) wished-for provenance system-based solutions and workarounds in their absence, (2) specific demands from provenance systems from different groups, and the relationship of those demands to aforementioned solutions, (3) fulfillments or lack thereof by explicit provenance systems, and (4) in what way the same wishes are fulfilled by other services not primarily intended as provenance systems. The review helps to (5) determine a design space for a provenance system.

[6] Defintion derived from http://www.wipo.int/copyright/en/management/

[7] In this phase of the project we are interested in their specific wishes regarding provenance systems for the Digital Commons, not e.g., everything they might think about formalities or copyright.

[8] This includes the authors of this paper.

1. Solutions for which provenance systems have been posited to provide for the Digital Commons

This section describes functionalities that a provenance system could provide, and workarounds that people take in the absence of an effective and intentional provenance system. We document such workarounds to indicate demand; in some cases workarounds could inspire features of a new provenance system. Workarounds also clarify which problems are best addressed outside of a provenance system. We also identify negative effects that the implementation of these functionalities could have on the realization of the Digital Commons. Finally we list the requirements implied by each functionality.

1.1 Content-derived Lookup

Definition: Obtain provenance information of a digital object using only the "content" of the digital object, e.g. using a content-derived hash[9] *as lookup key.*

Description of problem this solution addresses

Digital objects are often shared outside of their original publishing context and without links to the original context, especially on social media platforms. In many cases they are stripped of embedded metadata or resized, distorted, remixed. One the most often wished-for features of a provenance system is the ability to obtain provenance information for arbitrary digital objects.

Imagine a provenance system without this solution

Most provenance systems do not support content-derived lookup. Some bits of information about a work is required to initiate a search for provenance information, such as creator or title or an authority-designated identifier (e.g. ISBN).

Partial solutions and workarounds

- Ask whoever shared the digital object of interest where it was obtained.
- Ask around in a community venue (e.g. forum, wiki) that seems to be pertinent to digital object of interest about the provenance information of the work.
- Do a web search based on whatever limited context provided with the shared digital object, or what can be inferred from the content of the digital object itself.
- In the case of still images, do a similar images search on the web in a reversed image search engine.
- Research the embedded metadata for more details about the owner.
- Ignore the lack of provenance information for digital object, use it anyway without understanding the risk or think that the risk is minimal.
- Appropriate digital objects of unknown provenance (passing them on as one's own).

Downsides of realization for Digital Commons

- Digital objects for which a lookup obtains no provenance information, or just have no indication of permission, could be by default treated as blacklisted, even if they

[9] Usually a cryptographic hash or perceptual hash, but other methods are possible, especially given access to digital object for ongoing analysis relative to other digital objects, e.g. improbable phrases or word frequencies.

actually are freely licensed or in the public domain. To mitigate this risk a lookup system must strive to include as much of the Digital Commons as possible.
- Lookup can serve as the basis for enforcement of licenses, but can potentially also be misused to build new business models and sources of income around aggressive enforcement (look up “copyright troll”).

Requirements (scale)

- The system should be large, and include digital objects of GLAMs but also of individual users of the Internet. A lookup system that primarily returns a “this digital object is not known” type of response has no or very limited value for Users. Scale does not necessarily need to be measured in absolute terms or relative to the web as a whole, it can also be determined in coverage of a specific subset (relative to all works available via a particular Platform or from a particular aggregator).
- In the long term a lookup system would need to contain a substantial proportion of the Digital Commons as a whole.

Requirements (partnerships)

- Partnerships would be helpful on at least 2 levels. The first level concerns making the system aware of a substantial proportion of the Digital Commons. This implies partnerships with
    - Public institutions (GLAMs) with a mission to share their collections
    - Platforms for user contributed works (e.g. Flickr, Wikimedia Commons)
    - Providers of tools and services for content creation (Camera manufacturers, providers of software for content creation etc)
- The second level concerns the lookup functionality itself. Ideally it needs to be integrated with as many publication services as possible. This implies partnerships with:
    - Blogging platforms (e.g. WordPress; lookup could be integrated into the application’s media manager)
    - Platforms for user contributed works or digital objects (e.g. Flickr, Wikimedia Commons)
    - Web search services
- In many cases formal partnerships are not required (e.g. information can be harvested from many websites with no formal partnership; plugins can be provided for some platforms again with no formal partnership), but on a case-by-case basis, where the cost of formal collaboration is low and additional impact high, it should be pursued.

Requirements (features)

- To lookup works from arbitrary digital objects, possibly with no embedded or contextual metadata, the provenance system must associate content-derived identifiers with works.
- It should be possible to retrieve the information returned by a lookup at a specific point in time in the past; effectively these means all such information should be publicly versioned.

1.2 Registration

Definition: Intentional provision of provenance information about a work or digital object to provenance system for the purposes of establishing a relationship between a work or digital object and the registrant.

Description of problem this solution addresses

Establishing a relationship between an agent and a work or digital object on the web predominantly happens in an ad hoc manner as a side-effect of publishing on the web: publishing a digital object and associated provenance information on a site or under an account one controls. Such mechanisms generally do not have built-in provision of the provenance information to other systems or external agents in mind and as a consequence such ad hoc registrations are potentially not as reliable and searchable as those made to a dedicated provenance system. This also means that it is not obvious to creators or other interested parties that such ad hoc registrations serve the purpose of establishing relationships between agent and work. This means that ad hoc registrations might be underused relative to a dedicated provenance system with an explicit registration feature.

Imagine a provenance system without this solution

A provenance system without a registration feature could rely entirely on provenance information harvested from other systems, e.g. the ad hoc registrations created by web publication mentioned above. A provenance system that allowed any kind of non-harvested input (e.g. corrections and rating of provenance information) would to some extent itself serve as an ad hoc registration system, as parties interested in establishing relationships to a work or digital object would use such input mechanisms. Even a harvest-only provenance system, assuming it is both widely used and understood, would effectively provide an incentive for registration to systems it harvests from, i.e. make such registration less ad hoc and implicit.

Partial solutions and workarounds

- Online platforms (YouTube, flickr, Wikimedia Commons, Europeana, institutional repositories etc.) generally contain provenance information as part of their metadata. This information is often unavailable in structured formats outside of these web pages but can be obtained by browsing or scraping or via APIs.
- Collective Rights Management organizations generally maintain their own databases with provenance information. This information is rarely disclosed (other than to other CMOs).
- There are a number of specific attempts to set up registration information for works or digital objects that constitute the Digital Commons (e.g., Registered Commons, Safe Creative, see below) but these are limited in scope.
- Collection management systems in the GLAM sector contain provenance information related to works in their collections. This provenance information is often of dubious quality and often outdated. Such provenance information increasingly is published via the web and APIs as open data.
- The United States Copyright office maintains a registration system for copyrighted works.
- In Europe the Office for the Harmonisation of the Internal Market is hosting a registration system for Orphan Works.

Downsides of realization for Digital Commons

- Registration creates an additional burden on creators wishing to contribute to the Digital Commons.
- Registration requires substantial infrastructure and organizational investments that need to be covered.
- Depending on the implementation, in general the idea of registering works or digital objects as being part of the Digital Commons is undermining the general position that copyright is the exception and the Digital Commons / the Public Domain is the rule.[10]

Requirements (partnerships)

- See the section on Lookup above: in order to not be just another empty database on the web with a schema corresponding to registration, an explicit registration system needs critical mass and thus partnerships, just as a lookup system needs critical mass in order to not always respond “unknown” to lookups. The same set of intermediaries is pertinent.

Requirements (features)

- A registration system needs to support multiple sources of registration information (both manual and derived from existing systems).
- A registration system requires a conflict detection and resolution mechanism that is based on transparent criteria.
- A registration system needs to support initial registration and updates of provenance information.
- A registration system must be run by (a) trustworthy organization(s).
- Registration procedures must be transparent.
- A registration system must support all sorts of works and digital objects (most current registration systems listed in the partial solutions section above only support certain types of works or digital objects).
- Registration systems must be interoperable with existing identifiers for works/digital objects (ISAN, ISSN, ISBN etc) and authority files (such as VIAF).

1.3 Reliability

Definition: Provenance information provided by provenance system is reliable, both perceived to be generally accurate, and with a high probability of information to prove accurate if subjected to more in-depth investigation.

Description of problem this solution addresses

There are many digital objects on the Internet that are known to be in the commons (or that are officially in the commons but this fact is unknown). For these digital objects it is important to provide accurate, trustworthy provenance information. What should a provenance system claim about digital objects that are known to be in the commons? And why would anyone believe these claims?

[10] See the Public Domain Manifesto and COMMUNIA policy recommendation #8

Imagine a provenance system without this solution

A provenance system without reliability as an explicit goal or without any mechanisms intended to achieve this goal would still have *some* (possibly negative) level of reliability, determined by the presence of mechanisms that have the unintended side effect of increasing reliability and the extent to which the system is directly used or information from the system is disseminated. A system that is not reliable is thus a target for spammers or parties interested in injecting false information.

Partial Solutions and workarounds

- Check with the person who shared the digital object of interest if the claim about the provenance information is true (which undermines the purpose of commons tools such as CC licenses[11]).
- Ask around in the community venue (e.g. forum, wiki) that seems to be pertinent to the digital object of interest about the digital object.
- Do a web search based on whatever limited context is provided with shared digital object, or what can be inferred from content of digital object itself.
- In the case of still images, do a similar images search on web.
- Trust any information that is found.
- Disregard any form of provenance. Break the law and don't comply with copyrights. Use the work without proper attribution or agreement.
- Some content aggregation platforms (notably wikimedia commons) have developed elaborate (crowd sourced) procedures for verifying and correcting provenance information.
- Some GLAMs and cultural heritage aggregation platforms have internal rights clearance procedures that are aimed at creating reliable rights/provenance information.
- There are specialized tools that are aimed at identifying the rights status of a work and aim to recreate or establish provenance information (ARROW, FORWARD, CRMS at U Michigan, outofcopyright.eu).

Downsides of realization for Digital Commons

- No system will be 100% reliable (nor comprehensive); similar to the potential downside of largely successful content-derived lookup: if information from a provenance system is taken to be only source of truth, and anything not represented in there is effectively blocked from distribution, that would harm both formal commons and informal sharing.

Requirements (features)

- Any system for storage of provenance information about works or digital objects must have mechanisms for dispute resolution (both internal and in relation to other systems for storage of provenance information).
- It must be possible to dispute the correctness of lookup information returned by the lookup service, and to make corrections.

[11] One annoyance that is frequently expressed by creators that offer works under open licenses on on platforms like flickr that that they get asked if the licensed works can be used for a specific purpose. A percentage of such requests seems to be intended to confirm the licensing terms.

- It must be possible to update provenance information for works to reflect changes such as ownership transfers and new license offers.
- All provenance information must be public, and otherwise positioned for maximum exposure to entities holding accurate information about works and digital objects.
- Reliability of provenance information for digital objects can be enhanced if robust content-derived identifiers and whole-content analysis are used.

1.4 Whitelist

Definition: Content-derived lookup returning highly reliable provenance information that makes it possible for platforms (e.g. YouTube) to check ownership claims against a whitelist of works and digital objects in the Digital Commons.

Description of problem this solution addresses

It addresses inaccurate takedown requests[12] and other false-positive filtering of content which is actually free to share.

Imagine a provenance system without this solution

No provenance system currently provides this solution directly. Without content-derived lookup, a provenance system with reliable information could still serve as a non-automated whitelist, requiring human intervention but at a lower cost than an ad hoc takedown response.

Partial Solutions and workarounds

- Some platforms are reported to use internal whitelists.

Downsides of realization for Digital Commons

- A whitelist needs to be comprehensive otherwise only a subset of the Digital Commons is protected from appropriation/unjustified take down requests. Creating a comprehensive white list for the Digital Commons is difficult.

Requirements (scale)

- A whitelist needs to be comprehensive (at least in relation to the types of digital objects that are being published by a particular platform).
- A whitelist requires a high level of reliability of the provenance information that it is in the system.

Requirements (Partnerships)

- Whitelist implementations are primarily useful for (online) content aggregation platforms. Additionally due to scale requirement, see again lookup partnership requirement.

Requirements (Features)

- Platforms need to be able to access the whitelist via an API. For some platforms it may also be necessary for the whitelist to convey information to the platform via APIs specified by the platform in question.

[12] For example the case between the Blender Foundation and Sony.

1.5 Metrics

Definition: A metrics function provides indicators of how, where and/or by whom works or digital objects that are part of the Digital Commons are being used. Such metrics could leverage unique identifiers (and possibly the lookup functionality described above).

Description of problem this solution addresses

A lack of information about use of works in Digital Commons makes it more difficult to demonstrate effectiveness of (and thus to fund) commons. It makes it more difficult to discern what is popular (see more on this below, under discovery), and loses a source of provenance information which potentially could increase both the depth and reliability of a system.

Imagine a provenance system without this solution

Most provenance systems do not have this feature. It is expected nor essential for provenance systems in general, but might be considered an opportunity especially for a provenance system intended to facilitate Digital Commons, given paucity of metrics.

Partial Solutions and workarounds

- Many platforms (YouTube, flickr, Wikimedia Commons etc) provide statistics for views and embeds.
- There are a number of web analytics (e.g. Google Analytics, Piwik, server log analysis) tools that allow Creators and Publishers to measure uses in case those uses include links back to creator/publisher website.
- Using a reverse image search functionality of web scale image search engines (TinEye and Google).[13]
- An indication of use is provided in cases where content is both hosted and reused on platforms / in software (primarily WordPress) that receives and sends trackbacks/pingbacks/webmentions.

Downsides of realization for Digital Commons

- “What gets measured gets done.” If what is measured is not actually a positive indicator for Digital Commons, more metrics could encourage funders and others who care about commons to misallocate resources.
- Some value of the commons could be destroyed by emphasizing metrics which imply rank, thus can be taken as zero or even a negative sum for most participants.

---

[13] Isabella Kirton, et Melissa Terras. « Digitization and Dissemination: A Reverse Image Lookup Study to Assess the Reuse of Images of Paintings from the National Gallery’s Website ». *Journal of Digital Humanities*, 27 janvier 2014. http://journalofdigitalhumanities.org/3-1/reverse-image-lookup-paintings-digitisation-reuse/.
———. « Where Do Images of Art Go Once They Go Online? A Reverse Image Lookup Study to Assess the Dissemination of Digitized Cultural Heritage | MW2013: Museums and the Web 2013 », 2013. http://mw2013.museumsandtheweb.com/paper/where-do-images-of-art-go-once-they-go-online-a-reverse-image-lookup-study-to-assess-the-dissemination-of-digitized-cultural-heritage/.

Requirements (scale)

- There are no hard requirements, but some metrics will benefit from scale of some sort. For example, if a provenance system includes adaptation relationships, a relatively comprehensive system will capture more of those relationships. Importantly, more scale may give metrics more credibility; lack of scale implies rankings may be idiosyncratic.

Requirements (Partnerships)

- Content hosts are in the best position to gather data, and are in possession to use data. Partnerships could be useful, even necessary, where the platform is the dominant location of use and its use data is not already readily available for automated consumption.

Requirements (Features)

- Aggregate use data for works and digital objects should come from content hosts.
- Metrics about interactions should be retained with provenance information within system.
- Metrics should be published alongside other provenance information for works and digital objects.

Requirements (Other)

- Use data is easy to misreport and misrepresent and may be often be inflated. Multiple well-documented and comparable sources would be beneficial.

1.6 Discovery

Definition: Given that information about works and digital objects in the commons is in a registry, and this information is reliable and includes use metrics, what design would have to be implement to add a functionality that uses the registry to discover quality/relevant works and digital objects?

Description of problem this solution addresses

Digital Commons often compete with proprietary systems on production, but not on distribution/marketing/promotion. The nexus of provenance and use information about works in the commons could position a provenance system to fill part of this gap, at least making popularity[14] evident.

Imagine a provenance system without this solution

Most provenance systems do not have this feature. It is not expected or essential for provenance systems in general, but might be considered an opportunity especially for a provenance system intended to facilitate Digital Commons, given paucity of marketing for same.

[14] Self-fulfilling cultural relevance, see e.g.,*David E. Giles (2005/07) Increasing Returns to Information in the U.S. Popular Music Industry.; Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts (2006/02/10) Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market.*

Partial Solutions and workarounds

- Some Digital Commons collections can be searched (but this is only a portion of discovery).
- Site/platform-specific ratings and awards, e.g. featured images and articles on Wikimedia Commons and Wikipedia.
- Media-specific aggregated popularity and recommendation, e.g. libre.fm and cchits.net
- Individual ad-hoc recommendation, e.g. reviews on blogs.
- Ask others for recommendation, e.g. in person or in relevant forum.

Downsides of realization for Digital Commons

- Could exaggerate the downsides mentioned under metrics.

Requirements (scale)

- The system need not be comprehensive; “low quality” and “unpopular” works are unnecessary, that is, the vast majority of works need not be known to the system. However a critical mass of “quality” works known to the system are necessary for a good experience, and system should be comprehensive with regard to any actually “popular” work in the Digital Commons.

Requirements (Partnerships)

- Platforms that can provide rating and popularity data for works in commons.
- Platforms that can provide distribution for discovery mechanisms of provenance system.
- Similar to partnership requirements under the content-derived lookup case, formal partnerships may be helpful, but are not strictly necessary. To extent data, it may be harvested on the public web and provenance system may be used without tight integration.

Requirements (features)

- Search weighted by use and quality information.
- Top lists or other use/quality-enhanced browsing of information about works in the commons.
- Use and quality information published/available in any context other, or provenance information that is published/available for a digital object or work in commons.

Requirements (other)

- Use and rating information must be considered in-scope.

1.7 Sustainability

Definition: Provenance system is profitable, source of funds for additional commons activities.

Description of problem this solution addresses

Historically it has been imagined that a copyright registry could be a profit center. However, the funding of an organization like Creative Commons has never demonstrated the feasibility

of this approach. A provenance system though may provide value directly and indirectly to other parties contributing to the commons in some fashion.

Imagine a provenance system without this solution

This solution assumes a provenance system is a) a profit center and b) controlled by an organization dedicated to fostering the Digital Commons. These may be desirable attributes, but most provenance systems have neither. The solution is included as it has been a primary wish for past commons-oriented “registry” investigations.

Partial Solutions and workarounds:

- Funding achieved through donations or non-provenance system business.
- Funding not achieved, Digital Commons organizations make do with fewer resources.
- Provenance and related efforts are cost centers, funded through donations and non-provenance system business in addition to any required in absence of provenance efforts.

Downsides of realization for Digital Commons

- A provenance system developed with the primary aim of generating profit may be optimized for that purpose. The potential service to the commons might lessen, or even result in negative consequences such as those mentioned in the downsides.

Requirements (scale)

- In 2008, a registry was proposed[15] as a means of such support for Creative Commons (approximately US$3.5m annual spending[16] at that time). If profits were to support an organization the size of Creative Commons, such a system would need to be many times the size of such an organization, assuming normal profitability.

Requirements (Partnerships)

- Partnerships with entities that wish to pay for provenance system services or with the ability to help acquire such customers will be necessary.

Requirements (Features)

- One or more features (inclusive of services) that provide a direct, significant, and unique value to potential paying customers.

[15] http://creativecommons.org/weblog/entry/10043
[16] http://ibiblio.org/cccr/docs/990-2008.pdf

2. Demands from provenance systems from various stakeholders

Various stakeholders have posited that they would benefit if there were a better provenance system available. This section collects these demands as each group sees them, tease out what solutions are implicated by those demands, and any other requirements that might constrain or open up design space.

2.1 Authors / Creators

Demands

- Authors / Creators want their works to be found, consumed and used.
- Authors / Creators want to be properly attributed. In the absence of a system that provides provenance information, reusers have a harder time tracing provenance information and thereby attributing correctly.
- Authors / Creators want to be able to track and identify downstream uses of their works.
- Authors / Creators want to be traceable/findable by third parties interested in using their works.
- Authors / Creators who make works available under open licenses want those licenses to be reliable for others who want to use their work.

Solutions implicated

- Lookup, to the extent it facilitates better attribution and tracking
- Reliability to the extent other features rely on it
- Whitelist in order to prevent creators' works in Digital Commons from being inappropriately hidden from audiences
- A metrics feature that shows where work is used on the internet allows Authors / Creators to find infringing use/derivatives.
- A discovery feature will help creators to be found on the web and to avoid obscurity.

2.2 Right holders (publishers)

Demands

- Right holders want want their works to be found, consumed and used.
- Right holders want to be able to verify licensing information.
- Right holders want to be able to track downstream uses of their works and identify downstream users (possibly in order to enforce compliance with license terms).
- Right holders want to be traceable/findable by third parties interested in using their works.

Solutions implicated

- It is assumed that Right holders represent large collections of works. They will want to have an API that structurally returns information about the works they hold. This will enable them to compare their own systems to the information in a provenance store.
- Right holders would benefit from a metrics feature to find information about the use/reuse of the works they hold.

2.3 Users

Demands

- Users want to know if they have permission to share and/or reuse a work or not, to have clear rights statements on those works.
- Users want to discover works that they can share and/or reuse.
- Users want certainty that works that appear to be part of the Digital Commons are indeed part of the Commons.

Solutions implicated

- A discovery functionality allows users to identify high quality content that they can use (under open licenses or otherwise).
- Users want to have a lookup function that returns reliable provenance information about the work they are interested in.

2.4 GLAMs

Demands

- GLAMs want to ensure that works that are made available online can be discovered by third parties.
- GLAMs want to be able to ensure that they get attribution when works that they make available online are reused or shared by third parties.
- GLAMs want to deliver users reliable information about the rights status of works and make it easy for them to observe licensing conditions.
- GLAMs want to be able to find out how often and where their works are reused to be able to measure the impact of their online activities.[17]
- GLAMs profit from publicly available reliable provenance information as this can help them to establish the rights status of works in their collections.

Solutions implicated

- In order to measure impact, GLAMs require metrics functionality. A metrics feature can assess popularity of works to provide feedback to funders/boards.
- GLAMs benefit from reliable provenance information that is publicly available.
- As contributors to the Digital Commons, GLAMs benefit from a discovery functionality.
- A lookup functionality helps to fulfill the desire to be credited for their contributions to the Digital Commons.

2.5 Consumers

Demands

- Consumers want to discover content.
- Consumers want provenance information about the content they have discovered to be reliable.

[17] Illustrated by a British Library challenge “to encourage and establish the necessary feedback to measure the use and impact of public-domain content available through existing online platforms”.

- Consumers want to be able to discover content based on provenance information (e.g. more works from the same author/source).

Solutions implicated

- A discovery feature helps consumers to discover new works.
- Provenance information has to be reliable and easy to obtain.

2.6 Platforms / Intermediaries

Demands

- Platforms / Intermediaries want to have less notices and takedown requests and want ways to deal with them automatically. Reliable provenance information helps to determine the validity of claims and takedown requests more easily.
- Platforms / Intermediaries want reliable information about the copyright status of works that are hosted on their platforms.[18]
- Platforms / Intermediaries want to have metrics about downstream uses of works that they made online available.

Solutions implicated

- Platforms / Intermediaries want to have a whitelist feature that can be accessed via an API.
- Platforms / Intermediaries want a content derived lookup of reliable information provided by a provenance system or an API.
- Platforms / Intermediaries want a metrics functionality to complement their platform specific metrics in order to obtain more information about (the popularity of) their works.

2.7 Commons advocates

Demands

- Commons advocates want reliable information to assess if works are in the commons or not, both on a per work basis and in an aggregate form.
- Commons advocates want to increase the popularity of works that are part of the Digital Commons and encourage their reuse.
- Commons advocates want to increase the legal certainty of open licenses both to licensors and licensees.
- Commons advocates want to protect works in the commons from unjustified take-down requests.
- Commons advocates want to build arguments why a thriving public domain leads to more creativity.
- Commons advocates want to make it more attractive to Authors/Creators, Right holders and GLAMs to make a contribution to the Digital Commons.

[18] Note that the limitations of intermediary liability that are part of the DMCA and the E-commerce directive do not require hosting providers or platforms to check the copyright status of Digital Objects on their services/platforms. So this demand is only relevant for platforms/service providers that determine the rights status of works for other reasons (for example because they limit themselves to hosting Digital Objects that are part of the Digital Commons.

Solutions implicated

- A reliable lookup of works that are in the DIgital Commons aimed at encouraging confidence in the Digital Commons and associated tools like open licenses (this includes a possible white list feature).
- A discovery features that is explicitly targeted at driving Users and Consumers towards works that are in the Commons.
- A metrics functionality can measure the health and status of the commons, and identify areas where the commons can be expanded/reinforced.

3. Provenance systems as unintentional or (in)direct side effects of online Services

This section examines online services (relevant to the Digital Commons) that indirectly provide some of the functionalities of a provenance system. Most — perhaps all — of these serve as provenance systems for the commons through collecting digital objects in the commons, or metadata about works, or digital objects in commons with some particular domain/media/other concentration. In theory any site publishing digital objects or information about works or digital objects could be a provenance system.[19] In practice only select services have become useful as provenance systems, while many have not, such as would-be aggregators[20] that haven't aggregated enough to be actively used as provenance systems.

Each service is briefly described and then examined along the 7 functionalities that have been identified in section 1 above.

3.1 Europeana

Europeana.eu is an internet portal that acts as an interface of millions of books, paintings, films, museum objects and archival records that have been digitised throughout Europe.[21] Launched in 2008, it forms a metadata library collection of information of over 2.500 cultural institutions in Europe (called "data providers").

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided
- **Registration:** All Digital Objects made available via Europeana are associated with the contributing project or institution (data provider). It is possible to indicate Author/Creators but this information is not mandatory.
- **Reliability:** There is no consistent quality of the metadata in Europeana. The quality fluctuates greatly. Europeana does some high level checking of rights statements contained in the metadata.

[19] Enough such sites, publishing machine-readable provenance information, could constitute an emergent distributed provenance system. This is the intentional goal of the CC RDF scheme, covered in the next section.

[20] Some others are listed at http://wiki.creativecommons.org/Content_Directories or https://en.wikipedia.org/wiki/Creative_Commons-licensed_content_directories each badly in need of curation.

[21] See https://en.wikipedia.org/wiki/Europeana

- **Whitelist:** All records (over 38 million) have a rights statement, including the public domain mark (~8.4 Million works).
- **Metrics:** Europeana publishes a metrics dashboard (in alpha).
- **Discovery:** Europeana's prime focus is discoverability. With an extensive search platform and API that enables search, users can discover Europe's cultural heritage.
- **Sustainability:** Europeana is funded by the European Commission until 2018. Their business model relies on funding from the European Commission and some of the EU Member States.

3.2 Flickr

Flickr is an image/photo sharing platform launched in 2004. Users can upload photos and add specific information, such as location and license.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.
- **Registration:** All images uploaded are associated with a user account. User accounts can be anonymous.
- **Reliability:** Most content uploaded is managed directly by users. Content uploaded without rightsholder permission can be removed.
- **Whitelist:** Not provided.
- **Metrics:** These are available for users. There are also some specific functionalities, such as how many photos are taken with a specific camera.
- **Discovery:** Users can use a search engine and search according to different types of licenses.
- **Sustainability:** Flickr is owned by Yahoo. Users can pay for accounts with additional features.

3.3 Google search

Google search is the most popular general purpose search engine, providing multiple lookup methods and filtering options, e.g. showing images only, location or a specific timeframe.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Almost no support. The "similar images search" is content-derived in the sense that the searcher provides an image. However, no attempt to directly identify images using the same work or any provenance information is made. For text, one can manually search for phrases from a text and possibly find the same work.
- **Registration:** Not provided.
- **Reliability:** Web pages with false copyright information can be removed from the index.
- **Whitelist:** Not provided.
- **Metrics:** None directly, though for works closely associated with web pages, the "page rank" of those web pages is important.
- **Discovery:** Google search forms many people's default expectation of what discovery is. Google web and image search offer license filters.
- **Sustainability:** The business model is based on advertisement.

3.4 Internet Archive

The Internet archive was originally founded in 1996 to build an “Internet library”. It is one of the biggest content repositories online and holds a large collection of works that are part of the Commons.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** There is almost no support. SHA1 and MD5 hashes of media collection files are published,[22] but no facility provided to look up items given a hash.
- **Registration:** Digital objects that are uploaded are associated with a user account. User accounts can be anonymous. It is possible to provide additional provenance information.
- **Reliability:** Unauthorized content can be taken down, but there is no facility for non-uploaders to correct or update provenance information associated with uploads.
- **Whitelist:** Not provided.
- **Metrics:** Not provided.
- **Discovery:** Provided through the search option that allows people to look for different types of content, e.g. video, text, all media
- **Sustainability:** U.S. 501(c)(3) non-profit based on donations.

3.5 Learning Registry

The Learning Repository is a database and publishing platform for learning materials, launched in 2011 as joint project by the Office of Defense and the Office of Education. It is set up as a “resource distribution network with open APIs that anyone can use to expose or consume learning resources and information about how they are used.”

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.
- **Registration:** Information published into the Learning Registry is digitally signed by the publisher.
- **Reliability:** See registration. Unclear what other mechanisms are available.
- **Whitelist:** Not provided.
- **Metrics:** Allows publishing of use and effectiveness information about educational resources.
- **Discovery:** Allows users to search, browse by subject or browse by standard. Learning Registry Nodes provide a json API to get lists of documents matching metadata.
- **Sustainability:** Funded by the US government.

3.6 Libre.fm

Libre.fm is a clone of last.fm focused on free culture. The server software is free software, and hosted and recommended music is free culture.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.

[22] Example https://archive.org/download/HR053/HR053_files.xml publishes hashes for files in https://archive.org/details/HR053

- **Registration:** With an “artist account” one can supply their own music to the system.
- **Reliability:** Most free music in system is ingested from Jamendo, so piggybacks on whatever measures taken and survival-in-face-of-exposure depend on that host.
- **Whitelist:** Not provided.
- **Metrics:** “Loves,” listens for artists and tracks.
- **Discovery:** Uses counts of how often a song is listened to, tags, and collaborative filtering to show popular and related artists, recommend artists to particular users.
- **Sustainability:** None known other than very low costs, donated servers, and small donations.[23]

3.7 MusicBrainz

MusicBrainz hosts structured data about musical artists, works, recordings, releases. The service aims to be comprehensive. Mostly open source, open data (public domain, but limited amount is CC-BY-NC-SA), and it is run by an unusually transparent nonprofit.[24]

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Uses acoustid.org, an open source acoustic fingerprint implementation and service.[25]
- **Registration:** Not provided directly, through a rightsholder could provide information about their works as a regular MusicBrainz user.
- **Reliability:** Mass collaboration, can be thought of as a structured data wiki; processes include voting on submitted information. It follows strict information guidelines.
- **Whitelist:** Not provided. It is possible to associate license information with works, but this feature is rarely used. It would be possible to add a whitelist feature on top of existing capabilities.
- **Metrics:** Not provided.
- **Discovery:** One can search and browse the MusicBrainz database, but these features feel more oriented towards facilitating work on the database rather than music discovery, and there is no licensed-filtered search.
- **Sustainability:** A very low but not zero cost service. Open finances.[26]

3.8 Wikimedia Commons

Wikimedia Commons (2004) is a database of more than 24 million freely usable media files to which anyone can contribute. Wikimedia Commons is the image bank for all Wikimedia projects. Most images on Wikipedia are hosted on Wikimedia Commons[27].

Provenance system features provided, how, at what scale

[23] https://libre.fm/donate.html

[24] MusicBrainz community and relationship to music industry and standards have recently been written about in depth in *Jess Hemerly (2011/05/05) Making Metadata: The Case of MusicBrainz* and *Tony Brooke (2014) Descriptive Metadata in the Music Industry: Why It Is Broken And How to Fix It. Journal of Digital Media Management* respectively.

[25] https://acoustid.org/webservice#lookup specifically for the lookup API, https://musicbrainz.org/doc/AcoustID for MusicBrainz client use and https://musicbrainz.org/doc/Fingerprinting for further background, including two proprietary acoustic fingerprinting services previously used by MusicBrainz.

[26] http://wiki.musicbrainz.org/MetaBrainz:Finances

[27] There are some exceptions where a specific language wiki project hosts their own media files due to local copyright laws.

- **Content-derived Lookup:** No support.
- **Registration:** Media file uploads can be linked to a user account or done anonymously. Provenance information can be contributed by any user of the system.
- **Reliability:** Wikimedia Commons is generally very trustworthy due to a community that checks the rights status of a media file. Each file needs to have a field indicating an author and a license. The metadata at Wikimedia Commons it often not complete.
- **Whitelist:** No whitelist is available, although all available files on Wikimedia Commons are either openly licensed or public domain.
- **Metrics:** Metrics are available, but relatively unexposed.[28]
- **Discovery:** Files can be found on all Wikipedia projects or searched on Wikimedia Commons. Wikimedia Commons can be used to discover new media files.[29]
- **Sustainability:** This service is currently sustainable as they are recognized as a standard project by the Wikimedia Foundation.

[28] See https://tools.wmflabs.org/glamtools/glamorous.php and https://tools.wmflabs.org/glamtools/baglama.php

[29] Using categories you can discover images about certain topics, for example: https://commons.wikimedia.org/wiki/Category:Yawning_cats

4. Intentional or direct provenance-first systems

This section examines services that have been built with the explicit intention to serve as a provenance system. Many of the services listed in this section are no longer operational and perhaps none have fulfilled the wishes for a provenance system aimed at supporting the Digital Commons. Two of the services listed below are not yet operational.[30]

As in the previous section each service is briefly described and then examined along the 7 functionalities that have been identified in section 1 above. We then provide a (where applicable) the reasons for not being successful in developing a provenance system.

4.1 Attributor

Attributor (2005 - 2012[31]) was a startup intended to crawl the web and find reuses of registered texts, with the desire to facilitate soft enforcement and create revenue streams for the commons. It was pivoted to hard enforcement, acquired by DigiMarc.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided. Systems did rely on ingesting content and matching other content based on text similarity, but no direct content-derived lookup mechanism were provided.
- **Registration:** Users could upload texts or provide a feed from which Attributor would ingest new texts.
- **Reliability:** Unclear.
- **Whitelist:** Not provided.
- **Metrics:** The major feature of the service (before it became enforcement-only) was tracking of text-uses on the web. Each user had a metrics page from which they could see aggregate statistics on what Attributor had found and drill down to specific instances.
- **Discovery:** Not provided.
- **Sustainability:** Floated idea of a Creative Commons-like RevShare license. Idea was not taken up.

Reasons for lack of total success

- Apparently there was little demand for a soft enforcement mechanism (pivoted to hard enforcement).
- Skepticism/disinterest from the commons community in spite of press release from CC.
- Non-implementation of RevShare license.

4.2 Bitzi (2001-2013; actively developed 2001-2003)

Bitzi (2001-2013) was a file metadata service/mass collaboration community. It was actively developed from 2001 to 2013 and shut down at the end of 2013.

---

[30] Some of the following project are also analyzed in the Survey of Private Copyright Documentation Systems and Practices. WIPO also published a short piece on Copyright Registration and Documentation.

[31] Acquired by digimarc in 2012, but had pivoted entirely in direction of enforcement earlier. Only interesting as potential provenance system for commons in first few years.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** By use of an associate file metadata with SHA1 hash of the file. Any change creates a different hash, but very limited resilience could be obtained through referencing files with similar metadata. There is a lookup function of any file except the most wildly popular on P2P filesharing networks (lookup/contribution was built into some Gnutella clients) are nearly certain to obtain no information. There is limited utility for looking up quality (e.g. resolution, malware) of widely shared files on Gnutella in the early 2000s. Vastly greater scale and additional resilience to modified files would be needed for any utility of arbitrary content shared on the web in 2014.
- **Registration:** A username was associated with every report of a file and of metadata of a file. The first reporter established a high probability that they had access to the digital object at the time of the report or before (the other possibility is that the digital object does not exist; the report was of a random hash, or that they obtained the hash from elsewhere without directly accessing the file). A “privileged listing” feature allowed the agent with creator or right holder status with claiming the relationship to that digital object.
- **Reliability**: Accuracy is in theory obtained through community curation. Its effectiveness has never been studied. The scale is probably inadequate. Adding license information was facilitates, though very little used.
- **Whitelist:** Not provided.
- **Metrics:** Number of unique user reports of a file. Files that are widely shared might be expected to have many reports.
- **Discovery:** Lists of the most reported and highest “quality.” Users could rate files, but with respect to technical quality (ranging from deemed to be best available encoding to malware) rather than quality of the work manifested in a file/digital object. Due to specific implementation, most reports included many widely shared low quality files (e.g. zero-length file) and the highest “quality” included files both widely shared and deemed to be technically good/useful relative to other digital object manifestations of a work. Text search of metadata was also supported. No filtering of lists nor search for material in commons was available, thus there was almost no utility for discovery of works in Digital Commons.
- **Sustainability:** There was neither a profitable business nor one primarily directed toward support of the Digital Commons.

Reasons for lack of total success

- Bad timing (started late 2000 in wake of the .com crash) and a poor market positioning (P2P file sharing turned out to be doomed as a business model; though Bitzi not a file sharing platform, its most immediate adoption was by filesharers and integration into filesharing programs).
- Lack of total focus on features oriented toward advertising (potentially a digital object useful context for highly-targeted but not user-specific ads, much like keyword search).
- From a Digital Commons perspective, there was no intentional exploitation of the platform to promote works in commons.

4.3 CC-REL

CC-REL is a recommendation to publish provenance information for works and digital objects on the web (in RDF). CC-REL is maintained by Creative Commons and was actively worked on from 2003 to 2010. Based on the CC-REL specification Creative Commons aimed to encourage tools and services to make use of provenance information and more specifically the Creative Commons licenses.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Almost not provided. Conceptually there is some encouragement to publish hashes for licensed digital objects, but no inherent mechanism to perform lookup.
- **Registration:** An entity could assert a relationship between itself and a digital object or work by publishing metadata describing claimed relationships on entity's own web pages.
- **Reliability:** There was no direct mechanism for improving metadata published on random web pages. Indirectly, low-quality/non-curated web pages might have a shorter lifespan than high-quality/curated web pages, and specifically web pages with false assertions about copyright and right holder statuses could be subject to take-down.
- **Whitelist:** None provided, conceivably could be constructed (see below).
- **Metrics:** None provided, conceivably could be constructed (see below).
- **Discovery:** A primary intent of CC-REL was for others to construct (see below) this feature, e.g. search engines ingesting distributed CC-REL metadata and using it to enhance search. Implemented in extremely limited fashion (filter-only, probably taking an ad-hoc approach to consuming CC-REL, e.g. regular expressions, and largely not requiring CC-REL at all - known platforms such as Flickr and known license URLs are wholly adequate for realized implementations) for a time by Yahoo! web search, and by Google web and image search.
- **Sustainability:** Indirectly via helping Creative Commons do traditional fundraising; CC-REL was a part of fundraising pitches.

Reasons for lack of total success

- Several of the features above conceivably could be constructed on top of the distributed scheme by crawling the web and ingesting and curating found CC-REL assertions, but cost/benefit for doing were limited, thus almost completely not done:
    - Low information quality of published distributed CC-REL
    - Low technical quality of published distributed CC-REL
    - Identification of digital works and objects as actually published very fuzzy
    - Identification of agents as actually published very fuzzy
    - Pushed techno-political POV rather than working eagerly with all wanting to publish metadata about CC licensed/public domain works
    - No actual value provided above unannotated license url; CC-REL made CC adoption/implementation more confusing and costly for no benefit (other than being part of CC story)
    - Underestimated role of dominant platforms for publishing in the Digital Commons, severely overestimated role of licensor-controlled websites and of distributed metadata

4.4 creativecommons.net

Creativecommons.net (2009 - 2010) was a registry built by Creative Commons serving as both an affiliation mechanism/fundraising benefit and exploration of adding a form of distributed trust on top of CC-REL.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.
- **Registration:** Primary feature. Users could assert that they had licensed specific works at specific URLs, as well as collections of works under a URL prefix.
- **Reliability:** Users paid money to Creative Commons for accounts, creating a cost barrier to false information (not to mention true information) and a form of identification of users with real people or organizations. Claims could also be published on user-controlled website, vocabulary specified and used to denote symmetric claims on the user-controlled website and creativecommons.net. Similar to CC-REL alone above, false information/low quality websites expected to have a relatively short lifespan and be subject to take-down.
- **Whitelist:** Not provided.
- **Metrics:** Not provided.
- **Discovery:** Not provided.
- **Sustainability:** Primary aim of the project was to encourage donation to Creative Commons: accounts were based on donation benefits. Probably cost were higher to implement than any increase in donation due to offered benefits. Only a small fraction of the donors used benefits. However, exploration of the registry business was part of the large grant pitch and stipulation, thus in short the term concept helped to sustain Creative Commons.

Reasons for lack of total success

- In theory this could have been the foundation for constructing features built on top of CC-REL (see above) but as an extension of CC-REL. It suffered from many of that scheme's weaknesses.
- After the donation campaign at the launch of CC-REL, Creative Commons did not make a consistent effort to gain individual donors nor to further build technically on creativecommons.net and CC-REL, rendering the project doubly irrelevant.

4.5 Elog.io

Elog.io developed by Commons Machinery. Elog.io aims to repair broken attribution trails and ensure credit is given where credit is due. It currently exists as a browser plugin that gives users an opportunity to explore the photos that they encounter while browsing. For any of the photographs that are part of the Elog.io catalog it can tell information about the author of the photo, where it is from, and what permissions are attached to it. At the moment it has indexed 22 million images from Wikimedia Commons.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Provides content-derived lookup, and allows metadata lookup copied from other systems which might have information about a digital object.
- **Registration:** Unclear.
- **Reliability:** Intends to support user curation. Also intends to to use a crowdsourcing mechanism, would probably have some effect on reliability.

- **Whitelist:** Unclear.
- **Metrics:** Unclear.
- **Discovery:** Unclear.
- **Sustainability:** Unclear.

Reasons for lack of total success

- Incomplete, what exists is in startup phase, could become be a total success. Some key risks:
    - May need a much larger dataset and more sophisticated matching in order for lookups to not almost always find no results.[32]
    - Focus on attribution might constrain benefits, raise specific expectations that are hard to meet.

4.6 NoAnk

No Ank media (2007) was a proposed service offering a blanket licensing model for works distributed by peer to peer technologies. It was a spin-off of the Berkman Center at Harvard University but never became operational.[33]

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Was to use file fingerprints to facilitate tracking/metrics, unclear whether public/end user ability to do content-derived lookup contemplated.
- **Registration:** Licensing/rights-focused registry supporting registration,[34] but unclear to what extent direct registration (as opposed to ingesting records e.g. from labels and CMOs) would have been important.
- **Reliability:** Unclear, but probably was to depend on reliability of data from labels, CMOs, etc.
- **Whitelist:** Unclear if contemplated.
- **Metrics:** Tracking use was the core of the idea, but whether for any purpose other than apportioning revenue remains unclear.
- **Discovery:** Intended to be provided through integration with music consuming applications.
- **Sustainability:** Business model was supposed to be built upon a content fee to be collected from ISPs, device manufacturers etc (in the line of other proposals for alternative compensation systems).

Reasons for lack of total success

- The first intended large customers were universities in China; sharing music would be for free on-campus, use was tracked, right holders paid from deal. Unclear how realistic this ever was, but deals were not completed.
- Predicated on what has stopped vast majority of net music startups: deals with right holders. Unclear whether this had come into play by time university deals fell through.

[32] https://www.plagiarismtoday.com/2015/01/07/elog-io-fixing-image-attribution/
[33] See also this project presentation
[34] http://web.archive.org/web/20070606192208/http://www.noankmedia.com/howitworks.html

4.7 The Digital Registry / The Public Domain Registry (Berkman Center)

The goal of the Digital Registry was to create a legally defended digital registry for all copyrighted works, all orphan works and all works in the public domain[35]. The idea was born in the context of the Digital Public Library of America and submitted as one possible project to be realized within the Library, but it was rejected and abandoned.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Provided through the Registrar Identification Code (RIC) and a SHA-512 hash[36].
- **Registration:** Provided.
- **Reliability:** Unclear.
- **Whitelist:** Unclear. However, given that “defending” the public domain was one of the main goals for this registry project, it can be assumed that a whitelist was supposed to be provided.
- **Metrics:** Unclear.
- **Discovery:** Provided. The goal was to provide a registry with browsing and searching capabilities to the public at large.

Reasons for lack of total success

- DPLA was not interested to take over.
- Lack of other funding sources.

4.8 Numly

Numly was founded in 2006 in as a "digital copyrights & license management” service. The service assigns ESNs (Electronic Serial Numbers) to content on request by authors. These are connected to 'rights statements' (CC licenses or All Rights Reserved). Users can access these rights statements with the ESN. They receive a link that includes the number or a barcode.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.
- **Registration:** Provided via ESN.
- **Reliability:** The information that could be accessed was accurate, but it seems there is not a lot of information/content registered.
- **Whitelist:** Not provided without a user account.
- **Metrics:** Not provided without a user account (see payment package below).
- **Discovery:** Unclear. It seems that ESN is supposed to give users the option to search content via a number and/or the barcode of that number.
- **Sustainability:** Numly charges money to generate ESNs (9.95/month for 100 ESNs, which seems to be quite substantial given that they encourage creating ESNs f.e. for every single piece of content (such as individual blog posts)).

Reasons for lack of total success

[35] See project presentaion/overview
[36] More details can be found in this API specification

- No discovery function. The author doesn't want to pay to register every (non)-creation and it seems to have stopped development in 2006 (though continues to exist).

4.9 Ozmo

Ozmo was initiated in 2008 by Copyright Clearance Center (USA). It Supports the CC+ Protocol to allow CC licenses plus the possibility to obtain a private license via Ozmo's licensing system to purchase the rights which are not licensed under CC's public licensing suite. Ozmo was supposed to be an easy and convenient way to license content on the web. Ozmo made the claim that it gives legal certainty to buyers, though it was unclear if and how they check if the uploader is a creator (the user adds a license by identifying work (manually), setting up a 'license' and then publishing that).

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Not provided.
- **Registration:** Provided.
- **Reliability:** Unclear.
- **Whitelist:** Not provided.
- **Metrics:** Unclear.
- **Discovery:** Unclear.
- **Sustainability:** The Business Model includes no fees for signing up or creating a license, but Ozmo keeps 30% when someone licenses.

Reasons for lack of total success

- It was set up and launched by Copyright Clearance Center, which has a questionable reputation among Commons advocates.
- Development seems to have been abandoned at a very early and immature stage.

4.10 Registered Commons

Registered Commons is a service that promises to provide "secure registration of authorship of creative workings, no matter if it's photography, poetry, a series of mp3 files or an open source software project." It positions itself as a registration service that operates on top of (or in addition to) the Creative Commons licenses. Its primary selling point is that it helps authors/creators when licenses are infringed.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Registered Commons does support lookup by use of a hash but only if the hash is manually copied in the URL. It has no obvious interface.[37]
- **Registration:** Provided with user account.
- **Reliability**: Information is added by users themselves. They upload a file to then receive a certificate.[38] Files are hashed using SHA-1 and MD5. Any edit of the content of the file will render these hashes useless. Anyone can upload a file and receive a certificate for that file.
- **Whitelist**: Not provided.
- **Metrics**: Not provided.
- **Discovery**: There is a limited possibility to discover "recent works," and there is a search tool for keywords, though this is not very useful.

[37] See https://www.registeredcommons.org/faq#f11
[38] See a test upload here: https://www.registeredcommons.org/document/9022259474.pdf

- **Sustainability**: Registered commons provides basic services for free but maintains a Freemium business model: Payment for more registrations or to 'gain trust' is €150 per company; €50 per individual; €20 per developing country.

Reasons for lack of total success

- User Interface is unfriendly and confusing.
- There was a lack of momentum: only about ~ 60 public registrations happened in 2014 to date (the “If you build it, they will come” fallacy)
- There are cost for registering more than 3 works. The service seems too expensive given no demonstrable benefits.

4.11 Rights Data Integration

The Rights Data Integration project is an EU funded project that aims to realize a number of pilot implementation of the Linked Content Coalition concept. The project has been initiated by the LCC (European Publishers Council) and includes a number of content partners. The objective is to demonstrate the viability of the LCC rights reference model for internal and cross sector flows of rights information. The underlying LCC ambition is to provide a distributed rights information network.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Supported. The description of work calls for an “ability to plug into visual search and fingerprinting technology.” Some pilot projects such as the CEPIC image finder demonstrate content-derived lookup.
- **Registration**: This is its primary feature. The creator is intended to provide information to the system at the earliest possible opportunity.
- **Reliability**: Unclear, but possibly relatively good as they work directly with large content holders.
- **Whitelist**: Unclear.
- **Metrics**: Unclear.
- **Discovery**: Some implementations such as CEPIC image finder or a proposed collaboration with the UK copyright hub offer discovery features.
- **Sustainability**: Unclear. The RDI project is 50% funded by an EU grant and 50% funded by (some of) the project participants.

Reasons for lack of total success

- Too early to judge, most of the pilot implementations are in early conceptual stages. A total success would be limited as the RDI only intends to be a pilot. Some risks to any form of success are:
    - Pushes for bearing costs early (to creators, tool builders) for unclear/eventual benefits. This is a recipe for no or slow implementation.
    - Eschews centralization, which might increase complexity and cost.
- For Digital Commons: mentions supporting CC licenses and public domain status, which is unlikely to be a priority. If a post-pilot system was wildly successful it might lead to lower costs for paid licensing to a point that it makes the commons irrelevant.

4.12 SafeCreative

SafeCreative claims to be "the first copyright registry that informs about copyrights allowing Right holders to manage their rights in the Digital Era." Their main offering are certificates

that authenticate licenses and copyright claims. According to them this benefits creators (easier enforcement) as well as users (higher certainty of licenses). They also created a spinoff called semantic copyright which supposedly allows for querying 'the semantic copyright net' by content and then through hash identification receiving provenance/license information.

Provenance system features provided, how, at what scale

- **Content-derived Lookup:** Unclear, but likely through: uses MD5+SHA-1+SHA-512 to 'check integrity of the work'.
- **Registration:** Supported. It is the core of the service. The service also allows for automatic registration of works which are published on facebook via a simple tag based system.[39]
- **Reliability:** Unclear what mechanisms are in place for information other than timestamps.
- **Whitelist:** Not provided.
- **Metrics:** There are statistics about sales, and it is possible to generate a list of digital content that is licensed under specific licenses, e.g. CC etc. The result is given in percentage per month.
- **Discovery:** There is a search page for keyword search, and advanced search offers the option to narrow the search based on type of content, so that is the only useful one as otherwise you will get lots of text or music with lyrics even if you want an image. Every work can be accessed by a specific number.
- **Sustainability:** Safe Creative operates on a Freemium business model: they still have free service, but for more than 10 registrations per month users have to pay €72.60/year (personal) or €195 (corporate) per year. No specific information on funding so far although they appear to have had some government funding in the past.

[39] See this presentation

5. Design space of a provenance system

This section outlines the design space for a provenance system. Given the wishes, partial and attempted solutions that are outlined above, what can be summed up about the design space for a new and vastly superior provenance system?

5.1 Competitive advantage/untapped potential

There is a huge amount of provenance information on the web. Most of it has some credibility due to the publisher and/or survival online. Content hosts also collect and publish data about the use of digital objects and works, or are in a position to improve such collection and publication.

We desire to design a system that achieves a relatively comprehensive dataset that is reliable, which when coupled with metrics and discovery has immediate utility. Features either provided in isolation by unsuccessful systems or not provided at all (registration, content derived lookup, and whitelist) gain value and feasibility as the system scales. These are follow-on features that can also in the future be a revenue plan.

5.2 General principles

Whatever we are going to build, we will adhere to some principle that we believe to be important in the development of the Internet and address the underlying problems of provenance systems.

Openness

Our technology stack will be open. This means that we will use open source, open standards and open data, thereby limiting some of our technology choices. In case that this project fails to take off or is successful we as a group will have openly contributed to the narrative of provenance stores. If other users want to adopt, enhance our ideas they should be able to do so.

Any solution that we will think of will have an open governance model. Most solutions that we've encountered can be considered as silos of provenance information. We want to break through that barrier.

Global scale

Most researched solutions are intended for a global market. This is also the market we want to address. We see the lack of good provenance on the Internet as a fundamental problem of the medium and we intend to offer a direction to contribute to that issue. Any service or solution provided will therefore also not be jurisdiction specific.

[de]centralization

Should the provenance system be "centralized" - in the extreme, the one and only place one publishes (inclusive of "registration") and finds provenance information about any work, any digital file, of any sort - or some variety of decentralized/distributed/federated?

An extreme centralized version is unlikely to be obtained, even if it would be the explicit goal, for political and market reasons. LCC/RDI explicitly eschews a centralized registry, opting for

“hub and spoke.” Commons Machinery probably has something similar conceptually planned (“DNS for content”, “lookup referral”). It is unclear how such schemes will work in practice (consider that there is no natural hierarchy of content hash values that maps to authority). Further, decentralization greatly increases technical and market challenges.

Another way is to design a system that is centralized, with a sound market niche (perhaps the biggest challenge), but can publish and ingest linked open data (or really any data) and thus can “join” federation schemes in an ad hoc manner as warranted, and contribute to the web overall irrespective of explicit provenance systems.

5.3 Summary evaluation of existing and planned systems

Below is an overview of a scoring of different solutions. This scoring is not rigorous, but it serves the purpose of getting a better — if extremely coarse — feel for what existing/expired/planned provenance systems for the commons have accomplished. Each individual score is highly debatable, skip to the solutions review section for a narrative on each debated functionality.

Key: 0: solution not provided, 1: solution provided, but not robustly, or latent, 2: solution provided directly or indirectly and usefully, 3: solution provided, current best practice, *: proposed or never built systems, evaluations projected/what if

Intentional Provenance Systems

	Content-derived Lookup	Registration	Reliability	Whitelist	Metrics	Discovery	Sum
Attributor	1	2	0	0	2	0	5
Bitzi	2	2	1	0	1	1	7
CC-REL	0	1	0	0	0	1	2
CommonsMachinery*	2	1	1	0	1	1	6
creativecommons.net	0	3	0	0	0	0	3
NoAnk*	1	1	1	0	2	1	6
The Digital Registry*	1	3	1	0	0	1	6
Numly	0	3	0	0	0	0	3
Ozmo	0	3	0	0	0	0	3
RegisteredCommons	1	3	0	0	0	0	4
RDI*	1	3	1	0	2	0	7
SafeCreative	1	3	0	0	0	0	4
AVERAGE	**0.83**	**2.33**	**0.42**	**0.00**	**0.67**	**0.42**	

Side Effect Provenance Systems

	Content-derived Lookup	Registration	Reliability	Whitelist	Metrics	Discovery	Sum
Europeana	0	0	1	0	0	2	3
Flickr	0	0	1	0	2	2	5
Google search	1	0	1	0	0	1	3
Internet Archive	0	0	1	0	1	1	3

Learning Registry	0	1	1	0	1	1	4
Libre.fm	0	0	1	0	1	2	4
MusicBrainz	2	0	3	0	0	1	6
Wikimedia Commons	0	1	3	0	2	2	8
AVERAGE	**0.38**	**0.25**	**1.5**	**0**	**0.88**	**1.5**	

5.4 Solutions review

This is a summary of lessons derived from our survey of solutions. Some solutions are intentional, others are the side-effect of provenance systems and stakeholder demands. Possible implications for a provenance store are described below.

Content-derived lookup

There are roughly three kinds of content-derived lookup: 1) cryptographic hash, which securely identifies an exact file, but cannot tolerate even one bit of alteration; 2) perceptual hashes which attempt to map media-specific perceivable features to an identifier which other files with similar perceivable features should also map to;[40] and 3) other methods, for example if the content itself serves as the lookup query and one has a huge and relevant corpus to match against.

Scale is an important aspect of general content-derived lookup. A useful content-derived lookup service must have access to a database containing records of digital object with hash values of such scale that it is fairly likely that a typical lookup will obtain matches - otherwise performing a lookup is a futile exercise. If the “typical lookup” expected is constrained to some domain, the necessary scale of available records is lowered.

Implementation of content-derived lookup based on cryptographic hash is trivial: in a database or other store, a digital object record has a hash field. A service accepts a hash as a query and returns record(s)[41] containing matching value in hash field, or none. The interface for a content-derived lookup based on a perceptual hash is the same, but implementation is non-trivial as matches are non-exact.

Among expired and existing systems surveyed in this document, only Bitzi and MusicBrainz have realized a somewhat useful content-derived lookup. Bitzi’s database largely consisted of metadata for files widely shared on the gnutella network, making it somewhat likely that a lookup of a file identifier found through a gnutella search would obtain a match. MusicBrainz is focused entirely on music and uses a perceptual hash, making it very likely that a typical lookup of popular music will obtain matches. Obtaining provenance information for arbitrary digital objects without context seems to be a common idea, however direct demand for a non-omniscient version appears weak.

Implementing a hash lookup service which queries relevant systems based on the type of hash, e.g. AcoustID or CommonsMachineryStillImagePerceptualHash seems obvious for

[40] Commons Machinery is working on a perceptual hash for images, MusicBrainz uses AcoustID for audio.

[41] For a cryptographic hash lookup, there should be only one digital object record, but erroneous associations of hashes with records are inevitable.

providing provenance information.[42] Especially when we keep in mind that content-derived lookup may be demanded indirectly if it is a mechanism for realizing other solutions, e.g. whitelist and metrics.

A remaining question is whether there is some domain in which Project Octopus should build a new database of records which is both feasible and valuable — requiring a database of substantially fewer than billions of records and an eager audience for content-derived lookup that will use the service to lookup digital objects that appear to be in the targeted domain.

Registration

Most of the intentional provenance systems surveyed in this document offer registration; as a group it is their *only* strong feature. As none of those systems have met with much success, it might be concluded that registration is not in isolation a useful feature. Policy incentives or requirement for obtaining the benefits of some other useful feature and critical mass are needed for registration to have utility. This seems to be further borne out by lack of direct demand for registration by any stakeholder group and the failure to attract sizable user bases by projects that tote registration as their main feature (Ozmo, Numly, RegisteredCommons, and Safe Creative).

Much like content-derived lookup, implementation of unvetted registration is trivial, but should be approached with caution — a system offering registration not coupled to other compelling features will be yet another built-it-and-they-will-not-come nearly empty database on the web years later, assuming anyone bothers to keep the power on.

People do publish assertions of relationship between themselves and works and digital objects (recall these assertions are what “registration” is, abstracted from registration solely for copyright purposes) on the web, but largely as a side effect of publishing digital objects on the web, often in an ad hoc manner, and often without strong relationship semantics — the uploader of a still or moving image to Flickr or YouTube or the Internet Archive does not necessarily have any relationship with the image other than that as uploader to the service in question — but very often the uploader and creator are the same entity, and further the creator and other relationships may be indicated in free-form description.

MusicBrainz and Wikimedia Commons are notable for separately identifying contributor/uploader to the service and creator of the work/digital object in question, and both are about as far from the registration-by-copyright-holder model as one can get — both are mass collaborations dedicated to accurate information, with references rather than a direct assertion by authoritative agents to the system as the standard (though Wikimedia Commons does accept direct assertions through its OTRS system).

Some design possibilities regarding registration:

1. One path forward which avoids creating a permanently empty registration database and instead adds value to the existing latent provenance ecosystem for the commons could be to ingest creator and other relationships explicitly annotated by Wikimedia Commons and MusicBrainz (and also encourage other sites to make such annotations explicit). Those annotations could be presented as “registrations” within a Project Octopus interface — essentially making the existing latent provenance

[42] This can be both live queries (ie. a meta search engine) on other service or based on ingested data.

system explicit, thus encouraging further investment in the system and contributions to the commons.

2. Another and complementary path forward is the one trialed — with near zero uptake — by the CC Network: registration as claim/reinforcement of relationships to works published as digital objects elsewhere on the web. One way to envision this is as complementing (1). A service where you can “claim this” or claim “this is me” function in the Project Octopus interface whereby a user could assert that they are the agent (e.g. creator) in question, with whatever vetting of users Project Octopus does adding to the reliability of published relationships.

3. Yet another path is to go ahead and create yet another empty registration database, but with some compelling value for potential registrants. Is there some domain (note similarity to content-derived lookup) which Project Octopus participants have a solid relationship with and for which registration is really, truly, non-delusionally (i.e. no magic results expected from mere registration) demanded as a feature independent of any other and that Project Octopus might specifically target? If so the compelling value and demanding group needs to be made very explicit.

Other potential provenance system features do not necessarily depend on registration, but registration could be one mechanism for increasing reliability (and second-order, increasing credibility of a whitelist feature). An actual or planned registration feature in Project Octopus should be subservient to reliability, unless (3) can be established.

Reliability

Formerly existing and expired intentional provenance systems have made little effort toward publishing reliable provenance information. They have relied on assertions made by individual customers or entities with known poor quality data with no clear mechanism for correction nor massive exposure to takedown (which comes with hosting digital objects, not only metadata).

Provenance systems that are side-effects of hosting content potentially provide a great deal of reliability, as they are subjected to massive exposure to requests of takedown, and their users have some interest in accurate metadata. Some, most obviously Wikimedia Commons, include mechanisms for user curation of metadata.

Although on its face not as relevant to implement as content-derived lookup or registration, adding a mechanism for correcting errors in a provenance database is essential. This hopefully increases the reliability of provenance information in a database, and is conceptually simple. Implementation itself could be fairly trivial if an existing technology platform that has some of the required features (e.g. mechanism for making corrections, referencing and discussion of assertions, fully visible and retained audit trail) is adopted as the basis for Project Octopus. WikiBase being an obvious candidate.

A demand for reliability seems integral to a demand for content-derived lookup from the perspective of various stakeholders. But still, mechanisms for obtaining reliable provenance information are not inherently useful, except in case some community is so passionate about accurate information that they contribute to the accuracy of information even if it is not largely utilized by other services. MusicBrainz is probably the best example of such a

community — though MusicBrainz data is now increasingly utilized by other services, MusicBrainz contributors do not necessarily use those services. Compared to Wikimedia Commons, for which curation has the immediate purpose of facilitating the use of images in other Wikimedia projects.

1. A path forward, much like that for registration, is to make reliability a core aim and not merely present it as a side-effect of provenance systems. This could be done by ingesting metadata from comparable (reliable) systems and presenting this data explicitly as provenance information for works and digital objects within a Project Octopus interface, and referencing other “upstream” provenance systems as sources.

    a. Further, Project Octopus might expose reliability-relevant information that is hard to obtain but could be crowdsourced or automated, for example for how long a digital object has been published uninterrupted (not taken down) at a particular source, and the longevity of specific assertions particularly at sources that facilitate corrections.

    b. Further still, Project Octopus might facilitate making corrections upstream — e.g. if a Project Octopus user notices an erroneous statement sourced from MusicBrainz, there should be a mechanism that makes a correction upstream in MusicBrainz, where the correction will be vetted by MusicBrainz editors, in addition to any local mechanism Project Octopus has.

    c. Upstream sources that do not have a mechanism for correcting their provenance information might work with Project Octopus to ingest corrections vetted within Project Octopus, essentially outsourcing metadata curation, or a portion of it, to Project Octopus.

2. Another path, again concentrating on a new database not ingested from elsewhere, is to focus on correction/curation mechanisms within Project Octopus (obviously this is complementary to 1, and mandatory for 1c). Vetted registration could be one mechanism, wiki-like mechanisms another.[43]

There seems to be some direct demand for reliability, and reliability is crucial for a credible whitelist feature. Further, establishing the reliability of peer produced provenance information is crucial for companion policy positions. It seems that reliability mechanisms could be the highest priority for Project Octopus.

Whitelist

There seem to be a direct demand for a whitelist solution. For example from creators of commons-licensed works who would prefer their works not be illegitimately taken down as well as of commons advocates who want to protect the commons from illegitimate takedown and platforms that want to decrease both the risk and the cost of hosting UGC.

[43] This makes WikiBase an attractive technology platform, at least for a pilot project, but note that content-derived lookup based on perceptual hashes would require additional development as built-in search is based on exact values.

However, no existing or expired provenance system that we have surveyed provides a whitelist feature for works or digital objects in the commons. MusicBrainz *could* be used as such for audio — content-derived lookup based on AccoustID coupled with reliable provenance information, including a field for license — but the license field is largely unused and there has been no organized effort to ensure MusicBrainz coverage of musical works in the commons is comprehensive and exhaustive, including having license field populated.

Perhaps there is no existing solution because there is a high bar for realizing a useful whitelist service, including (a) content-derived lookup, at great scale; (b) highly reliable provenance information, both in actuality and perception; and hardest (c) willingness of significant platforms to utilize whitelist service, likely requiring years of business development effort and credibility building. Finally (d) there is an especially great challenge for whitelists and the commons — entire works and their digital object manifestations can be whitelisted, but if those works are detected in a derivative, that derivative cannot be whitelisted without further investigation — it could include elements which are subject to unmitigated copyright or be out of compliance with commons licenses. The benefits to be realized from a whitelist service may be too small or too diffuse to have been pursued until now.

1. If a commons whitelist service is to be developed by Project Octopus, it will be through a combination of content-derived lookup and reliability features (above) with implementation choices for those features informed by the requirements of the whitelist, coupled with a business development effort, which starts by gauging demands and gathering requirements from potential significant adopters.

2. Another path is to question whether an automated whitelist is actually of urgent priority for the commons (additionally weighing whether its promotion of algorithmic regulation is a net win for the commons), or whether the more effective “whitelist” is cultural — more popular, widely recognizable works in the commons, which people react against having taken down. The case of YouTube, Sony, and Sintel could be taken as a success for this kind of cultural whitelist. This path would call for not prioritizing an explicit content-derived lookup-based whitelist (even if one could be constructed by an interested user, as one could currently from MusicBrainz) but instead focusing on what a commons provenance system can do for discovery (below).

As (1) is a long-term but potentially attractive and problematic objective, it should probably be left to a second stage of development.

Metrics

There seems to be a strong demand for metrics concerning the use of works and digital objects in the commons, and hope that a provenance system for the commons might provide such metrics. However, metrics are essentially not provided in the surveyed intentional provenance systems. Some surveyed content hosting systems with provenance as a side effect provide some metrics (e.g. views, listens, referrers, reuse in Wikimedia projects).

Below are a number of approaches on how to do metrics within Project Octopus:

1. Ingest metrics along with other provenance information from the content hosting as a side effect system. The Project Octopus interface could expose metrics which

otherwise would require individual research. For example by aggregating views for a work across multiple platforms.[44] A further development of this path would (a) be working with platforms (both services and software used for self-hosting of content) to improve their metrics collection practices — they have a prime vantage for collecting views and referrers, and often don't do this adequately — and (b) to push those metrics to Project Octopus.

2. Another path is crawl-based, scouring web for uses of works (as Attributor did). This could be an independent crawl,[45] or (with delay of months) analysis of third party crawl (e.g. Common Crawl[46]). This would be a large and ongoing project which would require much experimentation to optimize for finding reuses, utilizing diverse methods to find matches.

3. Project Octopus could develop and host a pingback service,[47] which users or others with knowledge of use would call, indicating that "Work X has been used on URI Y" with possible further qualification of how the work has been used. This could be seen as a highly specific version of (1b) and/or (2) depending on implementation. Some mechanisms for client authentication and spam fighting are probably needed, evidenced by (ab)use of blog pingback.

4. Another strategy would be to develop tools for crowdsourcing reuse information. In essence this is a variant of option (2) discussed under registration, allowing users of the service to claim or identify reuse of digital objects. In the long run these would likely require a significant number of active curator-participants on the Project Octopus site. Precedents for crowd- and self-sourcing reuse information include Wikimedia Commons reuse categories,[48] ccMixter "trackback sitings",[49] and one of this paper's authors "reused" Flickr album.[50] In-depth investigation of reuse tracking by communities and individual creators would be a good topic for future research.

5. Finally, a Project Octopus service itself will generate metrics, based on automated and human views and edits for work and digital object provenance records. It is not clear these will be useful metrics; it is probably not the beneficial to the project to make them a major focus.

Ingestion (1) seems like the obvious low hanging fruit. Crawl (2) is complementary and also obvious, but requires significant resources or partnerships and should be part of a second

---

[44] For example views and reuse by other Wikimedia projects published by Wikimedia Commons. Some sites that publish view information may require scraping if such information not provided via an API.

[45] There are many strategies for what web resources to retrieve, e.g. http://commonsmachinery.se/2014/08/wikimania-part-2-what-will-elog-io-do/ retrieves only resources indicated by consumer user agent plugin.

[46] See http://commoncrawl.org/

[47] See http://www.w3.org/wiki/Pingback

[48] See https://commons.wikimedia.org/wiki/Category:Commons_as_a_media_source listing "images on Commons that have been used in a media publication (in the broadest sense) outside of Wikimedia projects" and https://commons.wikimedia.org/wiki/Category:Files_reused_by_external_parties_out_of_compliance_with_licensing_terms

[49] See http://ccmixter.org/pools/pool/9

[50] See https://www.flickr.com/photos/mlinksva/sets/72157613531054166/

development stage. Ping (3) requires third party software integration, most attractive version may be (1b). For purposes of a demonstrator (4) may be most realistic to implement.

Discovery

Discovery is provided to an extent by content hosts within their confines, both search (also provided across sites by Google web and image search with CC license filters) but more importantly various mechanisms for surfacing and recommending superior or relevant content, for example, featured images on Wikimedia Commons. However there are no existing mechanisms leveraging metrics across content hosts to expose the “most popular” works of a category in the commons or to make recommendations.

Direct demands for discovery exist. It may be useful for positioning the utility of provenance systems. Most intentional ones are rights-focused, to a near total extent when considering non-commons-oriented intentional provenance systems; cf. “obscurity is a greater threat than piracy”; and it is indeed for Digital Commons, which have largely failed to compete directly with proprietary channels — meaning competing on marketing and distribution.

A provenance system with reliable information about what is in the commons, and use and other popularity data, could be in a good position to attempt to take a first stab at filling this fundamental gap.

1. The obvious path is to develop aggregate metrics and rankings based on metrics (primarily ingested, see above), across works of various categories. Further (a) would be to work with upstreams to improve their metrics and ensure they are incorporated into Project Octopus discovery mechanisms.

2. Rating and curation mechanisms for use directly on the Project Octopus site could be developed; these would likely require a significant number of active curator-participants on the Project Octopus site to work well.

3. Project Octopus could expose browse and search mechanisms across provenance information in its database. This may be of more utility to curators/editors/other Project Octopus contributors than as a discovery mechanism.

4. Search across content, not only provenance information, could be developed, requiring ingesting actual content. This would require the most substantial resources and would need to be better for a set of use cases than for license filtered Google web and image search, which is probably a tall order.

Any of these will require substantial development and tuning to get beyond the demonstration level, but may nonetheless be useful to demonstrate concretely.

Sustainability

Arguably none of the existing systems have achieved sustainability based on their provenance features, with the possible exception of MusicBrainz (which operates on an extremely lean budget). As metadata in general and provenance information specifically is a cost center, it is probably wise to consider it extremely unlikely a new provenance system would be self-sustaining let alone throw off large profits to fund other commons activity, as had been imagined previously by Creative Commons.

Four paths forward, each complementing the others, but appealing to different constituencies:

1. **Thematic focus:** Develop policy and general benefit to culture/economy/internet story for Project Octopus. Could include building a new provenance system as in (2) below, but not focused only on building and scaling the system as a product. Effectively this path is a catchall, we promise to do some of all of below. Perhaps it'd be best to look for more resources for a stage 3 to continue and scale any that to show additional promise.

2. **Product focus:** Scale the provenance system on the likelihood that with a giant userbase and unlike previous systems influential position (perhaps both due to non-traditional features of provenance systems such as metrics and discovery) sustainability will emerge, through either user support (a la Wikimedia) or one or more of usual advertising/paid services/consulting businesses.

3. **Research focus:** Develop a research program that theorizes and tests value and effectiveness of provenance systems, including feasibility, reliability, policy relevance of commons mechanisms for provenance (e.g. open data, open source, open governance, peer production). This could complement (1) and possibly (2) or (3), but most likely the main output would be papers.

4. **Provenance ecosystem development focus:** In contrast with (2), Project Octopus could focus on the development and promotion of practices which improve how other systems act as provenance systems, e.g. how can already valuable side effect systems make providing provenance solutions first class? This would involve more convening and not require product development, although it would complement product development, especially any product that ingests data from elsewhere. A variant on this would be to focus on making digital commons related provenance information available via existing systems aimed at aggregating provenance information (such as the systems proposed by LCC/RDI).

6. Way forward/Conclusion

There is an unfulfilled desire from a number of stakeholders for a provenance system that benefits the digital ecosystem — most prominently on the side of users, GLAMs and commons advocates but also on the side of content publishers and rightsholders. Even though there are numerous examples of intentional and direct provenance systems as well as unintentional systems and systems resulting as indirect side effects from prominent online services, such as the Internet Archive, Google Search, Europeana and others, all previous attempts to build a functioning and widely used provenance system have proven to be unsuccessful and there are reasons to believe that ongoing efforts will meet the same fate.

However, based on our analysis, and especially the review of the different stakeholders' demands, we think that it will still be useful to develop a new provenance system, which is dedicated to tracking use of works and sourcing provenance information from across the web: When created as the first whole web observatory[51] of content flows, the envisaged new provenance system would allow everyone to identify and track digital content and related licensing information while avoiding building another locked-in database.

As a starting point, we propose to develop a prototype,[52] which includes interfaces for maintaining a basic set of provenance information concerning digital works and publication, and the kernel of a community with a passion for accurately documenting uses of their works on the web.

Based on this prototype, we will continue to analyze stakeholder demands and build partnerships with entities interested in engaging communities around provenance and use tracking, e.g., GLAMs, publishers, and software platforms. We will also begin the development of algorithmic/automated enrichment mechanisms utilizing content-derived and perceptual hashing, content analysis, and guided by the demands of users.

It is important to note that we are not planning to build yet another new copyright registry, but an open web observatory of content flows, which can be equipped with additional features and services.

6.1 Future research

This paper only scratches the surface of "commons provenance", focusing on a design space for a new provenance system for the cultural digital commons. That system, Project Octopus, will if successful create many new lines of inquiry. We have mentioned three complementary areas for research: (1) commons mechanisms (open source, data, governance) for provenance systems and provenance policy and attendant legal questions; (2) provenance practices for the software and data digital commons; and (3) practices of tracking reuse and other cultural flow metrics by digital commons communities and individual creators.

[51] For a definition of web observatory, see http://webscience.org/web-observatory/ "A Web Observatory is a system which gathers and links to data on the Web in order to answer questions about the Web, the users of the Web and the way that each affects the other" and complementary definitions at http://tw.rpi.edu/web/web_observatory and http://www.w3.org/community/webobservatory/

[52] Development is visible at http://project-octopus.org