Current methods to FAIR data in LAM

Angela Hurshman

Final Assignment
INFO 5841: Digital Curation Fundamentals

Current methods to FAIR data in LAM

Introduction

Libraries, Archives and Museums (LAM), are cultural heritage institutions that curate information about society’s identity, memory, knowledge, history, art, rituals, language, and cultural objects for current and future generations. In the better part of the last century, and in the analog world, their core purpose has been the access and use of cultural information and heritage. Museums are a bit unique in the traditional mode of access and use, where access can be a visit to the museum, and use is normally in the form of observing and recording (with pencil and paper) information, or to consult with the collections or the information professionals within (Hamilton & Saunderson, 2017.) With the rise of the world wide web, and the effects of digitization across all sectors of society, many LAM have developed and expanded their digital collections over the past few decades. Much of the collections have been developed through digitization of physical cultural objects, the results of which are referred to as “digital surrogates” (Sugimoto, et al, 2021, p. 27)

Digitization efforts began about a few decades ago, before the solid establishment of the digital curation field and consequent development of shared standards. Additionally, the rapid rise of Digital Humanities (DH) scholarship promulgated digital collections development by LAM to meet the growing demand of DH researchers. This has led to unique and separate modelling, databases, and standards across LAM, with digital collections often kept in their own
Current methods to FAIR data in LAM institutions’ repositories. These circumstances have led to lack of interoperability and the “siloed data” effect (Wang, et al., 2021.) This is despite most digital cultural heritage collections in LAM are “open” in theory by means of open licensing and digital publishing to institutional websites, but because of the “silo effect” these data are not findable or interoperable across systems, and therefore do not meet current standards for open data. Current standards in the open data movement follow the FAIR principles, which simply put, are about making open data Findable, Accessible, Interoperable, and Re-Usable in a way that’s human and machine readable (Go-FAIR.org.) In addition to the immense value that the move toward FAIR open data in the cultural heritage domain of LAM will have to the research infrastructure of DH scholars, demonstrated by Wang, et al.’s literature (2021), Hamilton and Saunderson (2017) also advocate for the movement within their analysis of the benefits of openly licensed, accessible, and reusable cultural heritage data. They cite return on public investment in the form of increased impact, creativity and ingenuity, public cultural knowledge and engagement, understanding of intellectual property, among others. In order for the scholarly and public sectors of society to reap these benefits, LAM require innovative and highly semantic digital curation solutions to overcome the roadblocks to making their currently siloed cultural heritage data FAIR. Before analysis of the trend to Semantic Web methods in responding to the issue of siloed cultural heritage data, I will provide a brief prelude contextualizing the intellectual property considerations omnipresent to the processions and proceedings of any movement in the cultural heritage sector, and relevant to appraisal and selection of what can be made FAIR open data from digital cultural heritage collections. I will conclude with examples of contemporary Semantic Web and Linked Data solutions that have been employed with the typical low resources of LAM, and then a summary of findings.
Current methods to FAIR data in LAM

What Data Should Be “Open” Data In LAM?

Prior to discussing methods addressing the silo effect and going FAIR in the cultural heritage sector, it is important to first describe what can or should constitute as “open” data in LAM. Open data, in this writing, refers to data that can be made accessible, and to some degree, re-usable, usually through the application of a Creative Commons license (CreativeCommons) or Public Domain attribution, or similar. As cultural heritage (CH) institutions, LAM manage parts of or whole collections that have legal or ethical obligations in regards to copyright and intellectual property. It would be negligent not to at least attempt to summarize this complexity to the open data movement within the CH sector. Both Chowdhury (2015) and Hamilton and Saunderson (2017) discuss the need for CH institutions to first consider their original reasons for digitizing collections when planning access and use policies and practices. There are valid reasons that certain cultural heritage data should be kept as “closed” data. The idea of “closed” data is exactly as it sounds, which is data that is not under an “open” license or not open access. (Hamilton & Saunderson, 2017.) There are three scenarios for logically maintaining closed data in the CH sector. In the first scenario are digital surrogates created from physical objects that are currently being kept for preservation purposes only, and therefore are created purely for the purpose of preservation efforts of such physical items that are damaged or deteriorating to the point that they will not outlast the digital surrogates (Chowdhury, 2015.) The second and most common scenario, involves either born-digital works or digital surrogates of works that are under third party copyright (usually creators of the works) that have either chosen to place a license on the work that is too restrictive to be considered open access and published as FAIR data, or works that may have been intended to be open access, but the copyright holder failed to choose a
Current methods to FAIR data in LAM

license for the work and so open accessibility is hampered by legal frameworks concerning intellectual property (Hamilton & Saunderson, 2017.) The third scenario concerns items in CH digital collections that require sensitivity to and protection of intangible cultural heritage (ICH) or Indigenous cultural heritage. ICH, refers to traditional heritage that is not tangible; such as, traditional performances, oral traditions, traditional medicinal practices, etc., which can oftentimes be associated with a tangible or physical cultural object also in holdings. Digital ICH collections may typically be in the form of recordings and documentation of the ICH. (Bonn, et al., 2016; Chowdhury, 2015; Sugimoto, et al., 2021; and Talakai, 2007.) The World Intellectual Property Organization (WIPO) warns “[R]ecording, digitizing and disseminating a traditional song or design, while valuable for preservation and promotional purposes, can also make [indigenous and traditional communities] vulnerable to misappropriation and misuse, especially in a digital world” (Talakai, 2007, p 10.) Ethical consideration in this scenario typically warrants restrictions to access and re-use regarding ICH, and does not fit into the current open data movement.

In Chapter 4 of Open licensing for Cultural Heritage by Hamilton and Saunderson, (2017) they criticize cultural heritage institutions for frequently claiming copyright of the digital surrogates they create in the digitization process of public domain works and asserting that the digital surrogate is a “new creative work” from the underlying work. Beyond pointing out the debatable legality of making such claims, they have this to say: “[seeking] to restrict the reuse of openly accessible and available copies of works that are intellectually in the public domain, and often retained and managed at public expense, generates new and unwarranted layers of copyright complexity” (p 80.) Although, they digress to acknowledge the potential risk of “loss,” the source of which could be loss of income, control, visits, (competitive) position, and “aura,”
they contend that the benefits to open licensing outweigh the potential for loss. Later in this writing, I will write about one of the case examples they use to back-up that contention. But already, they shine a light on the confusing legal complexity created by the practice of some LAM copyrighting digital surrogates, creating friction with the trending push for FAIRness of open data in the CH sector (Hamilton & Saunderson, 2017.) According to FAIR principles (Go-FAIR.org) and the logic Hamilton and Saunderson present in their literature, open licensing of digitized public domain works is only rational, and that clear communication and documentation of licenses placed on third party copyrighted work in open data are the best current practices for achieving FAIRness in LAM. Each institution should evaluate their collections and determine based on each of these prior scenarios, what data should be published as open data, and made current to FAIR data standards.

**Semantic Web Methods for Siloed Data**

So, we have large “silos” of data in LAM and they’re separated by unique databases, software applications, and standards (Wang, et al., 2021.) Since they are unique, they don’t understand each other or talk to each other, so how can machine readability and interoperability between all this “open” data be achieved? Finding an answer to this question is what has digital curators talking about Semantic Web methods. As described by W3C, to understand what the Semantic Web is, is to first understand that it is not ‘the’ web, it is “a” web. It is “a common framework that allows data to be shared and reused across application, enterprise and community boundaries” (W3C Semantic Web Activity.) It is also a “semantic” framework, meaning it focuses on language and recording how data relates to the real world concepts the data represent. The structure of this framework is based on the Resource Description Framework (RDF), which
Current methods to FAIR data in LAM consists of a subject (ex: Leonardo), a predicate (ex: PainterOf), and an object (ex: MonaLisa), this relational structure is referred to as “triples” and represented in Figure 1 (W3C.) If the Semantic Web is the structure upon which data is shared, Linked Open Data (LOD) is a method of publishing data against the RDF-based structure using HTTP Uniform Resource Identifiers (URIs) (Hausenblas & Karnstedt, 2010.).

**Figure 1**

*An RDF graph.*

“LOD Principles:

1. All Items in a dataset should be identified using URIs;
2. All URIs should be dereferenceable: using HTTP URIs allows looking up an item identified through an URI;
3. When looking up an URL, it leads to more data (typically represented in RDF);
4. Links to URIs in other datasets should be included in order to enable the discovery of more data” (Berners-Lee, 2009, as cited by Hausenblas & Karnstedt, 2010.)

By Following the above principles of LOD when mapping data to an RDF graph in the Semantic Web, that means each concept in the triple will have a URI, or link, which in turn increases machine-readability and interoperability across the data silos (W3C; and Go-Fair.org) This
Current methods to FAIR data in LAM

potential for interconnectivity across LOD is why this method has gained traction in the effort to make LAM data FAIR.

In addition to better interoperability on the machine side, publishing LOD adds further value to researchers in the DH discipline, for which the digitization of CH collections forms the basis of their research, by using standard ontologies, such as the CIDOC-CRM and publishing them as LOD terms to this semantic framework (Vlachidis, et al., 2021.) CIDOC-CRM is the ISO standard formal ontology in the CH domain (ISO 21127:2014.), and in the CIDOC-CRM Tutorial video-recording, Steven Stead (CIDOC) explains that it was made to be and “should be the language for semantic interoperability in the CH domain.”

Aggregation for Modest Cost

While the semantic enrichment and LOD technologies are trending methods to achieving FAIR data in the CH sector, mapping the data to these Semantic Web frameworks is a concerted, costly, and technical effort, and LAM are typically low-resource institutions with limited technical knowledge by staff (Smith, 2021.) In response, digital curation efforts within the CH sector have put a focus on aggregation platforms and models. Aggregation refers to the harvesting of metadata from data providers and maps the data to a common data model using shared vocabulary terms embedded in the metadata (Smith, 2021; and Sugimoto, et al., 2021.) As Sugimoto et al. (2021) point out, “[metadata] is key to organizing and managing digital collections of cultural and historical objects,” due to the domain common catalog structures and metadata standards widely used in LAM. The most widely known and successful data aggregation model in the CH digital curation world, frequently cited in the literature, is the Europeana Data Model (EDM) (Smith, 2021; Sugimoto, et al., 2021; and Wang, et al., 2021.)
Current methods to FAIR data in LAM

Upon ingestion of the metadata received from multiple sources, Europeana analyzes the metadata fields for matching terms and concepts from shared ontologies, mapping the data using RDF graphs and through linking data, semantically enriches the data (Europeana, 2013; and Sugimoto, et al., 2021.) The rationale presented in the Europeana Data Model Primer (2013), is that the EDM ascribes to a Semantic Web approach, “there is no such thing as a fixed schema [...] EDM can be seen instead as an anchor to which various finer-grained models can be attached, making them at least partly interoperable at the semantic level, while the data retain their original expressivity and richness” (Europeana, 2013.) Other similar data aggregation platforms have been deployed as well, such as the Swedish Open Cultural Heritage (SOCH) linked data platform case, studied by Smith (2021), and the Open Archive Initiative Object Exchange and Reuse (OAI-ORE) standard (Open Archives Initiative; and Sugimoto, et al. 2021.)

Semantic Enrichment by Deposit to Wikimedia Commons

Although aggregation is certainly the leading method for semantic enrichment and making open data FAIR among LAM, aggregation platforms and models are not the only low-cost option or current method to achieve semantic enrichment and FAIR data in the CH domain, cited by Hamilton and Saunderson (2017.) They discuss another current method to semantic enrichment is through the deposit of open digital collections to a 3rd party’s commons, namely Wikimedia Commons. It’s not the most technically ideal, but perhaps is the most frugal method for LAM who want to semantically enrich and increase FAIRness of their open data with very little budget afforded to the project and therefore is worth consideration (Wikimedia; and Hamilton & Saunderson, 2017.) Wikimedia Commons is an open repository of images and media, and these files can be re-used across all Wikimedia projects without re-uploading the
Current methods to FAIR data in LAM files; they can be linked and embedded directly (Wikimedia.) It may not be an aggregation platform, but as Smith (2021) points out, the sheer scope of the platform captures large quantities of triples and “[…]functions much like a directory of the identities within the world of Linked Data.” An example of a LAM institution benefiting from this method, provided by Hamilton and Saunderson (2017), is the case of the National Library of Wales’ effort to publish copies of their open licensed and public domain images from their digital collections onto Wikimedia Commons in 2015. The effort and decision to publish the images there was made to follow through on a prior commitment made by the library to make their data open, before a fire at the library with, fortunately no injuries, but major damage loss incurred to the library and its holdings. Over a six month period in the first year, they reached an average of almost 180k visits, and from that, for the same period in the next year, 2016, it reached a 76-fold increase (Hamilton & Saunderson, 2017.) Hamilton and Saunderson concede that this type of exposure is not reflective of direct impact, but it can certainly be used as tangible proof to demonstrate significant potential value to contracting stakeholder interests in generating return on public investment. A couple other LAM and CH entities have joined Wikimedia Commons and published their openly licensed and public domain digital CH collections; the Digital Public Library of America (DPLA), and Europeana; yes, the aggregation platform we were just discussing before (Wikipedia.)

Summary

We have reviewed examples of current Semantic Web and Linked Data methods that have been employed as digital curation solutions to address the accumulation of siloed data across LAM; CH institutions that typically face low resources for big projects. Semantic Web technologies such as LOD and RDF have afforded interoperability to a looming barrier of unique
Current methods to FAIR data in LAM

and separate systems and standards that occurred as a result of en mass digitization of CH
collections starting with the rise of the world wide web three decades ago. These innovative
solutions mark the beginning of a transformation of LAM from century old silos of cultural
heritage objects, to stewardship over the provenance and curation of cultural heritage
information, and the beginning of FAIR open CH data, which will add value to DH scholarship
infrastructure and public investments. Though issues of intellectual property persist inherent in
the collections of LAM, digital curation practices can help to address appropriate appraisal and
selection decisions in regards to what can constitute as and be published as FAIR CH data, and
what should continue to be closed, but safeguarded for future generation and knowledge. These
findings are evident of the growing importance to digital curation professionals within the CH
domain, and can help with advocating for funding towards the investment of digital curation
resources and solutions in LAM.
Current methods to FAIR data in LAM

References


CreativeCommons (2022, May 08). *About*. https://creativecommons.org/about/


Hausenblas, M. & Karnstedt, M. (2010). Understanding Linked Open Data as a web-scale
Current methods to FAIR data in LAM database. *2010 Second international conference on advances in databases, knowledge, and data applications*. pp 56-61. DOI: 10.1109/DBKDA.2010.23


Current methods to FAIR data in LAM

*Practices and Protocols in the South Pacific.* WIPO.


Wikimedia Commons. (2022, May 07). Commons: *Structured Data/GLAM.*

https://commons.wikimedia.org/wiki/Commons:Structured_data/GLAM

Wikimedia Commons. (2022, May 08). *File: Basic RDF Graph.svg.* [Graph]

https://commons.wikimedia.org/wiki/File:Basic_RDF_Graph.svg

Wikipedia. (2022, May 07). *Wikimedia Commons.*

https://en.wikipedia.org/wiki/Wikimedia_Commons

