My lecture today is about digital archives. I’ll start out with a riddle to get us thinking about different kinds of archives. Then I’ll talk about digital archives more broadly: how they come into being, what infrastructures underlie them. Then we discuss what risks they face, what ethical and analytical decision-making they bring to the fore, and what kinds of scholarship they enable.

I want to begin with a riddle. It’s from the Old English Exeter Book, a tenth-century manuscript that includes a collection of riddle poems, like this one.

The most frequent solution is Moth or Bookworm: a creature that devours books and so swallows up human knowledge and its “strong foundation” without understanding.

Here’s an example of bookworm damage in a medieval manuscript. Erik Kwakkel writes: “The most peculiar thing about the holes [bookworms] left behind is that they usually stop halfway the book. It is as if the worm, finished with his meal, turned around and traced his steps back to the restaurant’s exit. Like a satisfied customer.”

An easy solution to our riddle, right? But perhaps… too easy. If you look at the other Exeter Book riddles, no other riddle names the solution in the poem itself.

So if not “bookworm,” what else could the answer be? Let’s take a closer look at the riddle to find out.

What exactly is it that the worm devours? It eats words, a man’s song, a speech and its foundation. The foundation could be the physical container, the manuscript page itself. But in Old English, poetry may be written or spoken. Old English poetry survives in manuscripts, but it is ‘deeply influenced by its situation in a residually oral culture’ (O’Brien O’Keeffe 1994: 148), a culture where poetry may have been written down but still would have been heard rather than read by most people.When such poems were composed, their poets drew on stock phrases, images, and episodes; when such poems were recited, their listeners or readers recognized these traditional phrases and reacted to the phrases’ cargo of traditional affective associations.8 Listeners in Anglo-Saxon England, for example, would hear a poem’s reference to “wyrd” (fate) and remember what the poets associate with this fate—vanished communities; splendid buildings in ruins; the death of kings, warriors, civilization; the destructiveness of time gnawing away at all the works of humankind. Or they would hear “wyrm” (worm) and think not only of the insect, but of the word’s other meaning—which is dragon. I argue, along with Nicolas Jacobs and Martin Foys, that the bookworm riddle’s answer is at least twofold.9 One of the archives at risk is written and material: books, the foundation of human knowledge, are at risk from the insects that devour their pages. The other archive is oral and memorial: the vocabulary of traditional poetry, words and images and their deep-rooted affective resonances.10 If this shared poetic treasury, this collective memory, is the ‘strong foundation’ for speech and song, then what is the heedless worm that devours it? Forgetfulness? History? Time? The discussion of our riddle serves to focus students’ attention on technologies of text and on the way that complementary technologies of text—here, oral poetry and manuscript production—coexist and interact. The riddle also evokes two kinds of at-risk archives: archives endangered because their physical form is endangered and archives endangered because, for reasons technical or cultural, they become illegible, inaccessible to changing reader communities.

So I argue, with Megan Cavell and Martin Foys and other scholars, that the bookworm riddle’s answer is at least twofold.

One of the archives at risk is written: books, the foundation of human knowledge, are at risk from the insects that devour their pages. The other archive is oral and memorial: the vocabulary of traditional poetry, words and images and their deep-rooted affective resonances.10 If this shared poetic treasury, this collective memory, is the ‘strong foundation’ for speech and song, then what is the heedless worm that devours it? Forgetfulness? History? Time? The discussion of our riddle serves to focus students’ attention on technologies of text and on the way that complementary technologies of text—here, oral poetry and manuscript production—coexist and interact. The riddle also evokes two kinds of at-risk archives: archives endangered because their physical form is endangered and archives endangered because, for reasons technical or cultural, they become illegible, inaccessible to changing reader communities.

Digital archives live somewhere between the physical and the virtual: they are abstracted versions of physical objects, representations of physical, touchable objects as 0s and 1s in the collective memory of server farms. First, we look at the making and preservation of digital archives; then we turn to some scholarly possibilities that these archives enable.

Archives are collections of documents. They can contain digitized manuscripts, books, newspapers, legal documents, video footage, oral histories, sounds, and social media interactions.

Digital archives can originate with print media. Here is the Toronto Public Library Digital Archive: originating from a collection of print media, digitized and made available online.

Other archives are born digital. Here is an archive of social media activity during the Egyptian revolution of 2011: tweets, blogs, Facebook posts, and so on.

Archives consist of records: digitized manuscripts, books, newspapers, legal documents, video footage, oral histories, sounds, tweets.

Each record of an archive is described by metadata: structured data about data. For example, think about your library’s online catalogue: it contains the records of books, with fields like Author, Title, Publisher, Editor, Date. Or take a look at the Omeka Item for this print: its metadata includes title, creator, description, source, rights, etc: the elements of the Dublin Core metadata schema. A metadata schema is a set of categories to describe the stuff in your archive. If you use standard metadata schemas, like Dublin Core, MARC, or MODS, your data will be discoverable and interoperable with library catalogues, other scholars’ projects, or other data sets.

How are digital archives created? For my first example—the digitization of physical collections--I want to draw on work at the Internet Archive—work I had the privilege of witnessing thanks to Andrea Mills, at the Internet Archive scanning centre here in Robarts Library. The Internet Archive, a [San Francisco](https://en.wikipedia.org/wiki/San_Francisco)–based nonprofit [digital library](https://en.wikipedia.org/wiki/Digital_library) whose stated mission is "universal access to all knowledge". The Internet Archive provides free public access to collections of digitized materials, including books, movies, software, music, and websites. On the 7th floor of Robarts, the Internet Archive digitizes books—from Fisher Rare Books Library; from other UofT libraries; and from all over Canada. In the slides that follow, I want to trace the journey of these materials from physical artifact to digital surrogate.

Digitization requires very specialized equipment and expertise. Before digitization can begin, materials travel to the scanning centre. Some books travel in humble cardboard boxes, like this stack of yearbooks from Vancouver College.

Other materials travel in style—in this case, in enclosed archival boxes that themselves travel in padded wooden crates very much like vampire coffins (because, like vampires, these books are fragile and long-lived, travellers from an earlier time).

Here’s the vampire coffin.

Here is a human being, for scale. My disembodied hand indicates the thickness of the padding of the box in which the boxes of archival materials were stored.

Once you have assembled your material, digitization work can begin. You’ll need a specialized camera that can capture the colours of your manuscript, the tiny details (some not visible to the naked eye), erasures, scribal doodles, the marginal notes written by scribes and readers over the years, even bookworm damage. You’ll need a cradle to support the manuscript. You’ll need a conservation-friendly room, where you can control all the light sources, to make sure you take the clearest, most accurate photograph.

Here is a specialized scanner as used by the Fisher Library. Notice the black book cradle under the book, and colour calibration strip on the right-hand side of the book. These help ensure archival-quality images.

Here is a specialized scanning station from the Internet Archive’s outpost on the seventh floor of Robarts. Notice the curtains around the station, ensuring a minimum of light interference; the bright light on the scanner itself; the computer screen, to check the quality of the scan; and the book cradle, holding the book open at a 100-degree angle so as not to injure its spine. On the seventh floor of Robarts, in the Internet Archive’s quarters, there are seven or eight such digitization stations, each in its own cocoon of drapery and with its own overhead light.

The scanning bed is angled, forming a book cradle that allows a book to sit partially open. This supports fragile spines and bindings.

Colour reference sheets help calibrate the scanner, ensuring the colours it records in its images are as accurate as possible.

Show the video.

Not from the Internet Archive, but from Google Books, here is an example of digitization error: the hand of a scanner operator appear in digitized Google Books.

Amusing and titillating as these images are, it’s easy to forget that they’re the work of an army of invisible laborers—the Google hands. This is the subject of an art work by the Brooklyn-based artist Andrew Norman Wilson called “[ScanOps](http://www.andrewnormanwilson.com/ScanOps.html).” The project began in 2007, when Wilson was contracted by a video-production company to work on the Google campus. He noted sharp divisions between the workers; one group, known as ScanOps, were sequestered in their own building. These were data-entry workers, the people to whom those mysterious hands belonged. Wilson became intrigued by them, and began filming them walking to and from their ten-hour shifts in silence. He was able to capture a few minutes of footage before Google security busted him. In a letter to his boss explaining his motives, Wilson remarked that most of the ScanOps workers were people of color. He wrote, “I’m interested in issues of class, race and labor, and so out of general curiosity, I wanted to ask these workers about their jobs.” In short order, he was fired.

<https://www.newyorker.com/books/page-turner/the-artful-accidents-of-google-books>

**Why Digital Archives**

The stakes are high indeed, as UNESCO reminds us:

“Documentary heritage reflects the diversity of languages, peoples and cultures. It is the mirror of the world and its memory. But this memory is fragile. Every day, irreplaceable parts of this memory disappear for ever.”

UNESCO Memory of the World Programme, quoted by BL`s Endangered Archives Programme

**Data Loss**

Like physical books, digital archives have their own vulnerabilities: to server shutdown or server failure or data corruption; or to the obsolescence of software and hardware in which an archive is housed.

Let me tell you here the story of the two Doomsday Books.

The first is a Latin manuscript of the eleventh century: for taxation purposes, William the Conqueror ordered a record made of landholders and their property in much of England and parts of Wales. The book was finished in 1086 and it resides at The National Archives at Kew, London.

The second Doomsday Book was a BBC initiative, 900 years later: an attempt to build a “digital snapshot of the country,” a documentation of everyday life in communities around the United Kingdom. Over a million people contributed.

The BBC Domesday Project was stored on adapted laser discs. Reading these discs required a state-of-the-art microcomputer, the Acorn BBC Master, customized with a specially produced laserdisc player.

Unfortunately, the machines were very expensive. They became obsolete. And the laser discs and the data on them became unreadable. Meanwhile, the medieval manuscript endured, almost as legible as it was a thousand years ago.

Fortunately, there’s a happy ending. Since the late 90s, academics and technologists around the world tried to recover the data. In 2011, the BBC’s own technologists, George Auckland and his Innovations Team, completed the extraction of the laser disc.

**Keep it Secret; Keep it Safe**

What can born-digital projects do to prevent losing data to obsolescence? Let’s take a look at another project: U of T’s very own Dictionary of Old English. It lives on the fourteenth floor of Robarts. The DOE studies the earliest form of the English language, from 600 to 1150. Founded in 1970 at the University of Toronto, the Dictionary and its Old English corpus predate the Internet by more than a decade.

In the early seventies, the Dictionary of Old English Project created an electronic corpus (a body of texts) that included at least one version of every known Old English text—typed on a typewriter onto Scantron sheets, because time on the mainframe computer was very expensive (Ashley Crandell Amos, "Computers and Lexicography: The Dictionary of Old English," Editing, Publishing, and Computer Technology, AMS Press, Inc., 1988). These Scantron sheets were scanned, and so the text was digitized. The piece you see above is from the Old English poem Beowulf:

HWÆT, WE GAR-DEna in geardagum,

þeodcyninga þrym gefrunon,

hu ða æþelingas ellen fremedon!

(Listen, we have heard of the Spear-danes in days of old, the kings of the people and their glory, how the princes performed deeds of courage.)

And here is the DOE Corpus in 2016, accessible online.

What did the DOE do right?

First, lightweight data: all its data, which is text, comes up to about 210 MB of data.

Second, data format: the corpus adopted a community-supported, international encoding standard (TEI) as soon as that standard became available.

Third, software platform: the DOE largely relies on open-source, community-supported software (Apache Solr). Open-source means that source code is made available with license that permits others to study the code, change the code, and develop the code collaboratively.

And fourth, the DOE is surrounded by a vigorous community of practice: Anglo-Saxonists and language historians around the world use it in their research..

How long do hardware and software platforms live? What’s their life expectancy?

Like physical books, digital archives have their own vulnerabilities: to deliberate erasure and censorship; to server failure or data corruption; or to the obsolescence of software and hardware in which an archive is housed.

**Data Access & Storage**

Data access and storage are the first point of vulnerability. Best practice is several variations on not putting all of your eggs in one basket.

Storing the data on your own machine is very risky. If you put it in Dropbox or Google Drive, that’s better, but not perfect: what if you or a team member deletes the data by accident? Even better if you have multiple off-site backups: so if e.g. Dropbox has a security failure and your own computer is compromised, you can still get your data from the hard-drive at your parents’. Even better: GitHub (which is a platform that tracks versions of your data, where you have the history of your data and can go back to earlier versions).

Libraries and cultural heritage institutions manage these risks in similar ways. Well-kept archives possess multiple off-site backups and ; they live in repositories with technical safeguards against data degradation.

If one institutional repository is nice, more are even better. LOCKSS is an open-source preservation system that sustains a distributed network of institutional repositories. It’s like storing your backup hard-drive at your parents’ and having them store their hard-drive at your house. In other words, everyone protects everyone else’s data. LOCKSS stands for “lots of copies keep stuff safe”:by storing data across a distributed network of institutional repositories.

This can work if e.g. a natural disaster, like a flood, brings down servers in one physical location. It can also work against deliberate attempts to destroy data or make it inaccessible. For example, in 2013, U.S. Congress temporarily shut down the U.S. federal government in order to block Obamacare. A slew of important government websites, from the Library of Congress and the National Oceanic and Atmospheric Administration to NASA, went dark Users could no longer access their services. The Internet Archive, however, had archived captures of the sites – and made them available through its Wayback Machine.

And for additional protection, there are dark archives, for disaster recovery. These keep data secret and inaccessible until a trigger event such as a natural disaster or another kind of catastrophic data loss.

**Standards & Software**

After data access and storage, standards and software choices are another point of vulnerability. Wise data curators avoid dedicated, one-off, artisanal software platforms made by commercial providers. If a commercial provider stops operating, their software dies with them. Instead, a safer choice are Open-source, community-supported standards and software. If standards are community-supported, it means your archive will be able to interoperate with other archives. IF your software is open-source, that means the code is available online, and you can look at it, modify it, fix it, share it—and a whole community of developers can do the same.

Then there is obsolescence. Software platforms go out of date, become obsolete, and eventually die. There are two ways of dealing with this. First, migration: this involves moving your data from an obsolete, less stable platform and format into a newer, more stable platform and format. The other way is emulation. This involves recreating the—now obsolete—original environment of a digital object in a new, stable software platform. This is what you do if you want to play old arcade video games on your PC.

**Political Suppression**

Digital archives are vulnerable to politically motivated suppression, including data deleted outright, data no longer accessible to the public.

Having an archived copy of a site elsewhere enabled users to keep accessing the data. In January 2017, the University of Toronto, UPenn, and the Internet Archive hosted a hackathon to archive data that would be endangered during Donald Trump’s presidency. The aim of these librarians, scholars, and community members was

“To archive the federal online pages and data that are in danger of disappearing during the Trump administration. [The event] focused on preserving information and data from the Environmental Protection Agency, which has programs and data at high risk of being removed from online public access or even deleted. This includes climate change, water, air, toxics programs.” (Michelle Murphy, <https://technoscienceunit.org/2016/12/04/guerrilla-archiving-event-saving-environmental-data-from-trump/>).

Even if this data is deliberately deleted from US government sites, parking copies of it in other jurisdictions means it will be available in the future.

**Disasters**

Digital archives are vulnerable to war & natural disasters.

Across history, we have terrible stories of libraries destroyed in natural disasters. An example from the ancient world is Library of Celsus, Ephesus, Anatolia, now part of Selçuk, Turkey – destroyed by earthquake and/or fire and/or Goth invasion in 3rd century AD. More recently, the Canadian Ice Core Archive:

Freezer fails,  temperature rises to about 100 degrees F, ice core samples from Canadian Arctic “which contained 22,000 years’ worth of atmospheric information, were entirely or partially destroyed” (Martin Sharp, cited by Tatiana Schlossberg, “An Ice Scientist’s Worst Nightmare,” New York Times, 11 April 2017)

Digital archives are also physical – so they partake of these risks.

With climate change:

“collection evacuation risks, prompted by wildfires, floods, and hurricanes

“long-term relocation decisions due to sea-level rise and coastal erosion, or if a weather event is so devastating, rebuilding is inadvisable or impossible

“increasing infrastructure and preservation costs when current HVAC systems can’t keep up with future increases in temperature and humidity” (Eira Tansey, “When the Unbearable Becomes Inevitable: Archives and Climate Change,” 2017, http://eiratansey.com/2017/05/16/fierce-urgencies-2017/)

LOCKSS, dark archives, community building

**Misrepresentation**

Cultural memory and digital archives are vulnerable to another factor: not obvious political suppression, but a sneakier kind. This is what Nowviskie (2017) calls the ‘library problems of misrepresentation, thwarted agency, and structural neglect.” What happens to the cultural memory of marginalized people? Of marginalized cultures? What happens to cultural memory when it’s stored in someone else’s systems, described through categories that are pejorative or marginalizing? What happens to the cultural memories that are collected via appropriation or not collected at all?These archives represent knowledge that, though it may not be explicitly censored, is nevertheless endangered. Nowviskie, Marisa Elena Duarte, and Miranda Belarde-Lewis, among others, discuss anti-racist or anti-colonial libraries and archives in a US context (Nowviskie 2017; Duarte and Belarde-Lewis 2015). And they ask important questions: as we build digital archives, who decides what cultural memory is worthwhile? Who decides how this knowledge is archived? Who decides how this knowledge is described? Who decides who can access this knowledge?

In a Canadian context, when it comes to the erasure of cultural memory, we have difficult and terrible history to contend with. Truth and Reconciliation Commission (TRC) is an initiative dedicated to telling and documenting the residential school system in Canada. The residential school system was a colonial school system that separated Indigenous children from their families and communities, forced Indigenous children (often through horrific physical and psychological abuse) to adopt settler cultural and religious norms, and deliberately aimed at the erasure of Indigenous cultures and spiritual practices (Truth and Reconciliation Commission 2015). The digital archives of the National Centre for Truth and Reconciliation document the history of residential schools in Canada. They do not restore the searing losses and injustices of the past and the present. They can, however, function as ‘a place of learning and dialogue where the truths of their [the survivors’] experiences are honoured and kept safe for future generations,’ with the hope that ‘their families, communities and all of Canada would learn from these hard lessons so they would not be repeated,’ and where they can share ‘the wisdom of the Elders and Traditional Knowledge Keepers on how to create just and peaceful relationships amongst diverse peoples’ (National Centre for Truth and Reconciliation). An archive that documents profound harm to Indigenous people—including the banning and endangerment of cultural knowledge—becomes cultural memory in its own right, ‘a legacy gift to all Canada’ (National Centre for Truth and Reconciliation).

Where do digital frameworks come into this?

Digital archives provide ways into this cultural memory. Are these ways respectful? Are these ways ethical? There are digital platforms that come out of precisely this endeavour of representing cultural memory of marginalized cultures in respectful ways. One example of a magnificent DH project is MOOK-oo-too:

“Mukurtu (MOOK-oo-too) is a grassroots project aiming to empower communities to manage, share, narrate, and exchange their digital heritage in culturally relevant and ethically-minded ways.”

“In 2007, Warumungu community members collaborated with Kim Christen and Craig Dietrich to produce the Mukurtu Wumpurrarni-kari Archive.

Mukurtu is a Warumungu word meaning ‘dilly bag’ or a safe keeping place for sacred materials. Warumungu elder, Michael Jampin Jones chose Mukurtu as the name for the community archive to remind users that the archive, too, is a safe keeping place where Warumungu people can share stories, knowledge, and cultural materials properly using their own protocols.

Growing from this community need, Mukurtu CMS is now an open source platform flexible enough to meet the needs of diverse communities who want to manage and share their digital cultural heritage in their own way, on their own terms.”

A safe keeping place. A a glory-fast speech and its strong foundation. With that, we return to the riddle of our bookworm; to a meditation, more than 1000 years old, about collective memory and its safekeeping. What does Nowviskie tell us about this collective memory? What are ways to hold on to knowledge and history and our creations? Communities of practice, ethics of care, whispered translations. Ultimately, what will survive of us is what we build communities around. The worm riddle carries the echoes of Old English poetic tradition. We hear in this poem the sound of relentless time the destroyer – because time has not yet destroyed the memory of these poems. Paradoxically, we understand it as a meditation about time and loss because this poetic language is not yet lost. And in this poetic language, in this poetic tradition, loss doesn’t mean just the works of humankind falling apart. Human works crumble, in Old English poetry, when? When they’re uninhabited. When they’re no longer used. When they’re buried and forgotten. Cultural memory: the worm may destroy it. But not as long as there are people who hold it. Not as long as the song is still sung.