4-26-2024

Building a Roadmap for Web Archiving: Organizational Sustainability in an American Research University Library

Ruth E. Bryan
University of Kentucky Libraries, ruth.bryan@uky.edu

Emily B. Collier
University of Kentucky Libraries, ebcollier@uky.edu

Building a Roadmap for Web Archiving:

Organizational Sustainability in an American Research University Library
International Internet Preservation Consortium, April 26, 2024

Ruth Bryan
University Archivist

Emily Collier
Assistant University Archivist

UK Libraries

Hi everyone! I'm Ruth Bryan and I'm Emily Collier, and we'll be presenting on building a roadmap for web archiving organizational sustainability in an American research university library.

Greetings from Lexington, KY, USA, and the University of Kentucky!

The Blue Grass Region of Kentucky: horses, bourbon, and oh, basketball!

Libraries

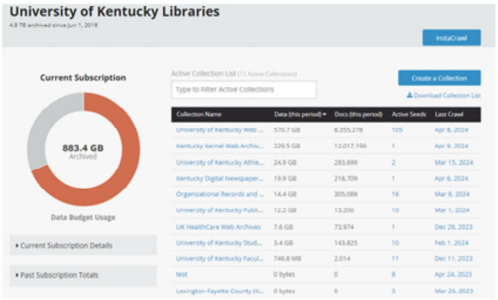
UKY Quick Facts:

- Established: 1865
- Enrollment: 33,885
- Colleges: 18
- Majors: 200+
- Budget: $6.8B

We bring you greetings from the University of Kentucky in Lexington, Kentucky, a state in the upper southeast of the United States. Kentucky is part of the Blue Grass region, which is known for thoroughbred horses, bourbon whiskey, and basketball. The University of Kentucky is a research-intensive school founded in 1865 with 18 colleges, more than 33,800 students and 200 majors, and a total budget of 6.8 billion dollars.

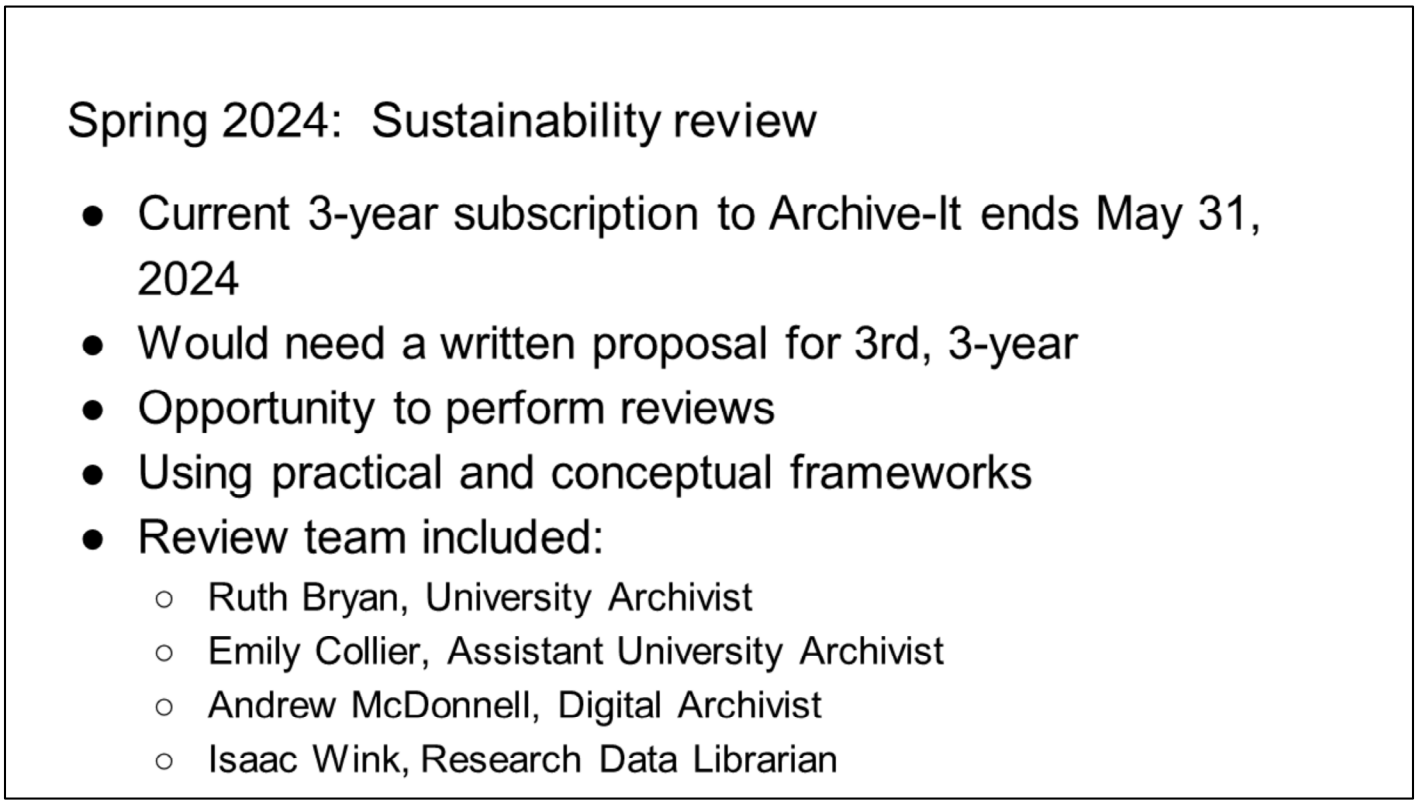
Web archiving at UKY overview

- Housed in the University of Kentucky Libraries
- Selects, preserves, and facilitates access to archived copies of websites and social media that are official and permanent records of the University of Kentucky; that are relevant to Special Collections Research Center (SCRC) collection priorities; and that support university faculty and student research and teaching.
- Began in 2018 with Archive-It
- Total subscription to date:
  - 4.7 TB
  - 82,395,451 documents
  - 10 web archives collections



Libraries

The web archiving program is housed in the Special Collections Research Center, part of the University of Kentucky Libraries. We preserve and provide access to websites and social media that are university permanent records, that support Special Collections' collecting priorities; and that support research and teaching. The program began in 2018 using Archive-It, which is our platform today. Our total holdings include 4.7 TB and more than 82 million documents organized into 10 web archives collections.

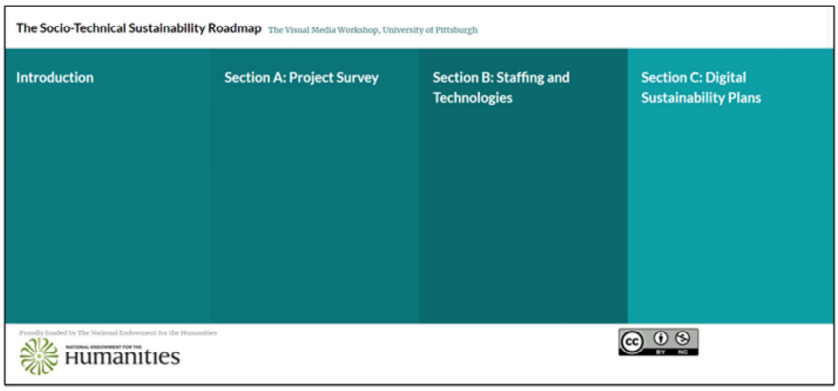


Because the current 3-year Archive-It subscription ends at the end of May, we decided to conduct a sustainability review of the program using both a practical framework and a conceptual framework. This review would contribute to writing a proposal for a third, 3-year subscription. The review team included Ruth and Emily, as well as Andrew McDonnell, the Digital Archivist (who's also presenting at IIPC), and Isaac Wink, the Research Data Librarian.

Practical: Socio-Technical Sustainability Roadmap (STSR)

- https://sites.haa.pitt.edu/sustainabilityroadmap/

- "How long should this last?" at project beginning
- Plan for lifespan
- "Intellectual goals"



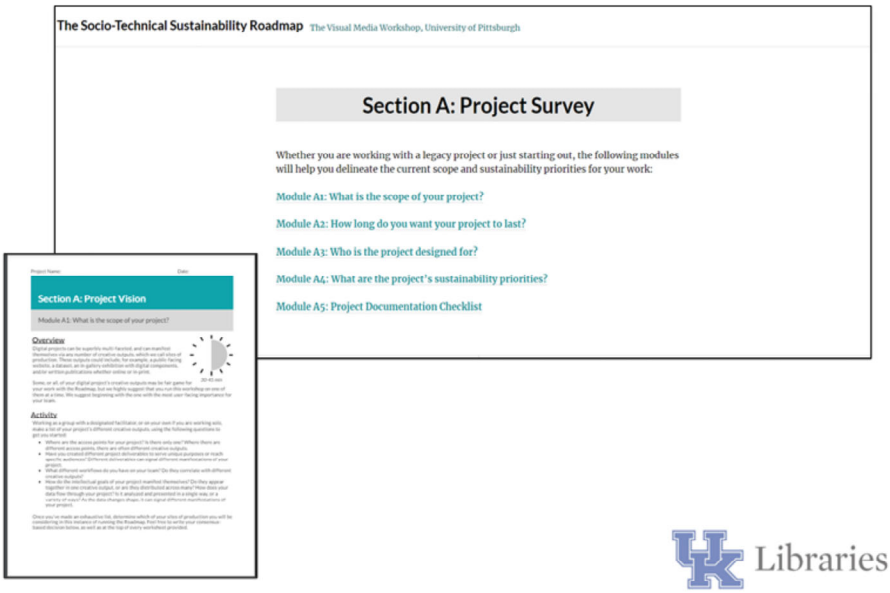
UK Libraries

The practical framework we chose is the Socio-Technical Sustainability Roadmap. The STSR is a practical series of activities designed to answer the question, "how long should this last?" at the beginning of a digital humanities project and then to plan for that lifespan taking into account the project's "intellectual goals."

"Digital humanities projects can, and should, have a variety of expectations of longevity, and these expectations should match individual intellectual and technological goals as well as a realistic assessment of funding opportunities."

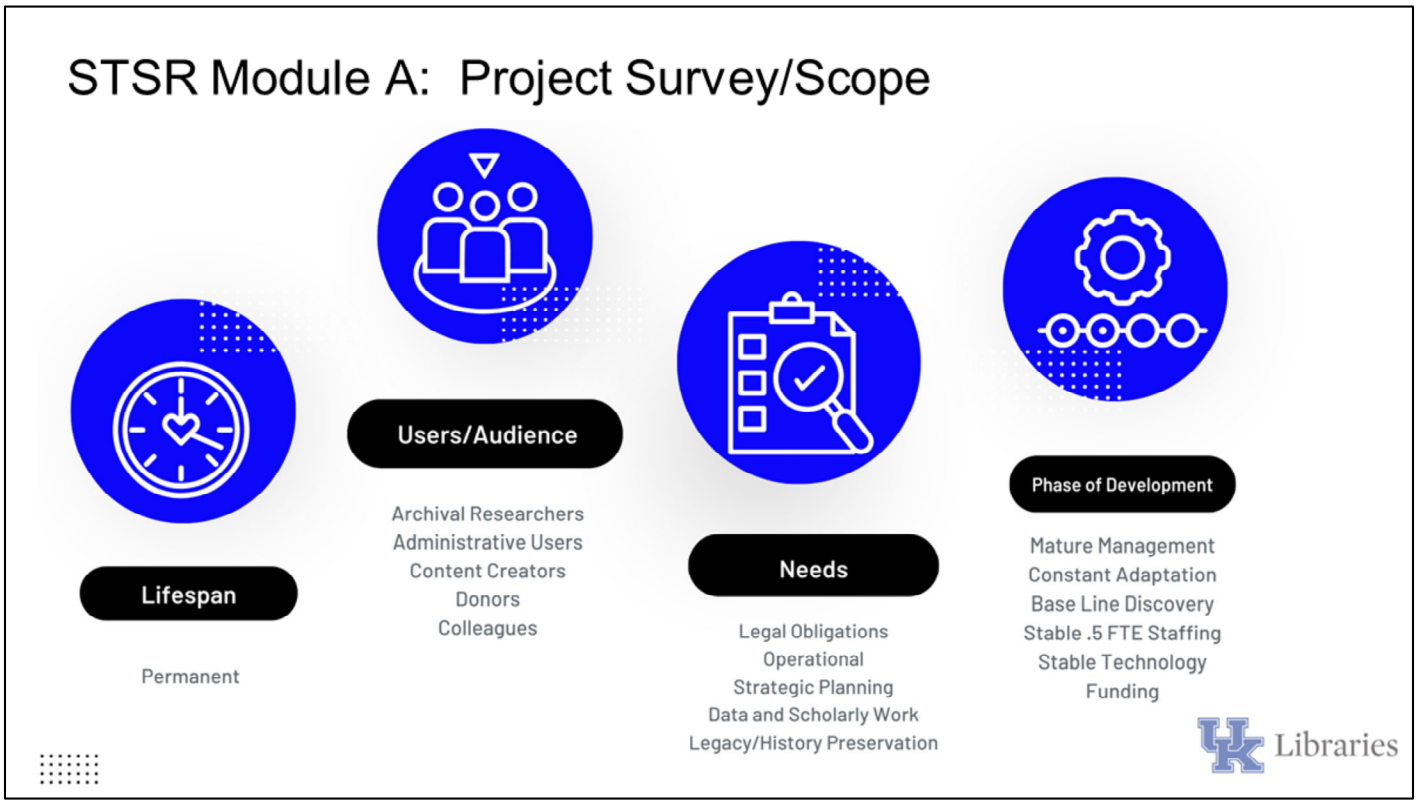
Practical: Socio-Technical Sustainability Roadmap (STSR)

- 3 modules:
  - Project Survey
  - Staffing and Technologies
  - Digital Sustainability Plan
- 3-year cycle



Libraries

The Roadmap comprises 3 modules. Each module has 4 to 5 sections. Each section has an activity sheet and/or a spreadsheet to fill out. The Roadmap should be reviewed and updated every 3 years of the project lifespan.



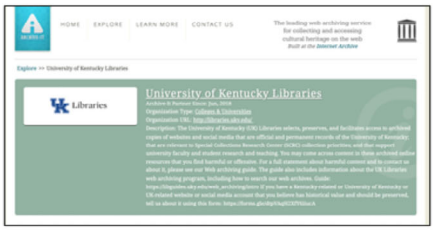
Module A prompts the team to survey or scope the project, including articulating the project's lifespan, users, phase of development, and its "significant properties," which include aspects such as access points, project deliverables, audiences, workflows and data flows, creative outputs, and intellectual goals.

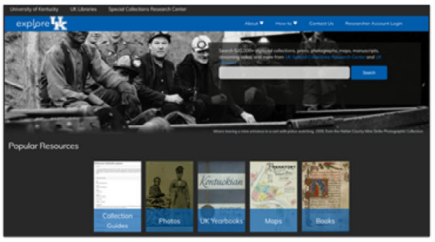
Our lifespan is permanent; our users include researchers, university administration, web content creators, donors, and colleagues. Our phase of development ranges from mature to stable to base line.

STSR Module A: Significant Properties

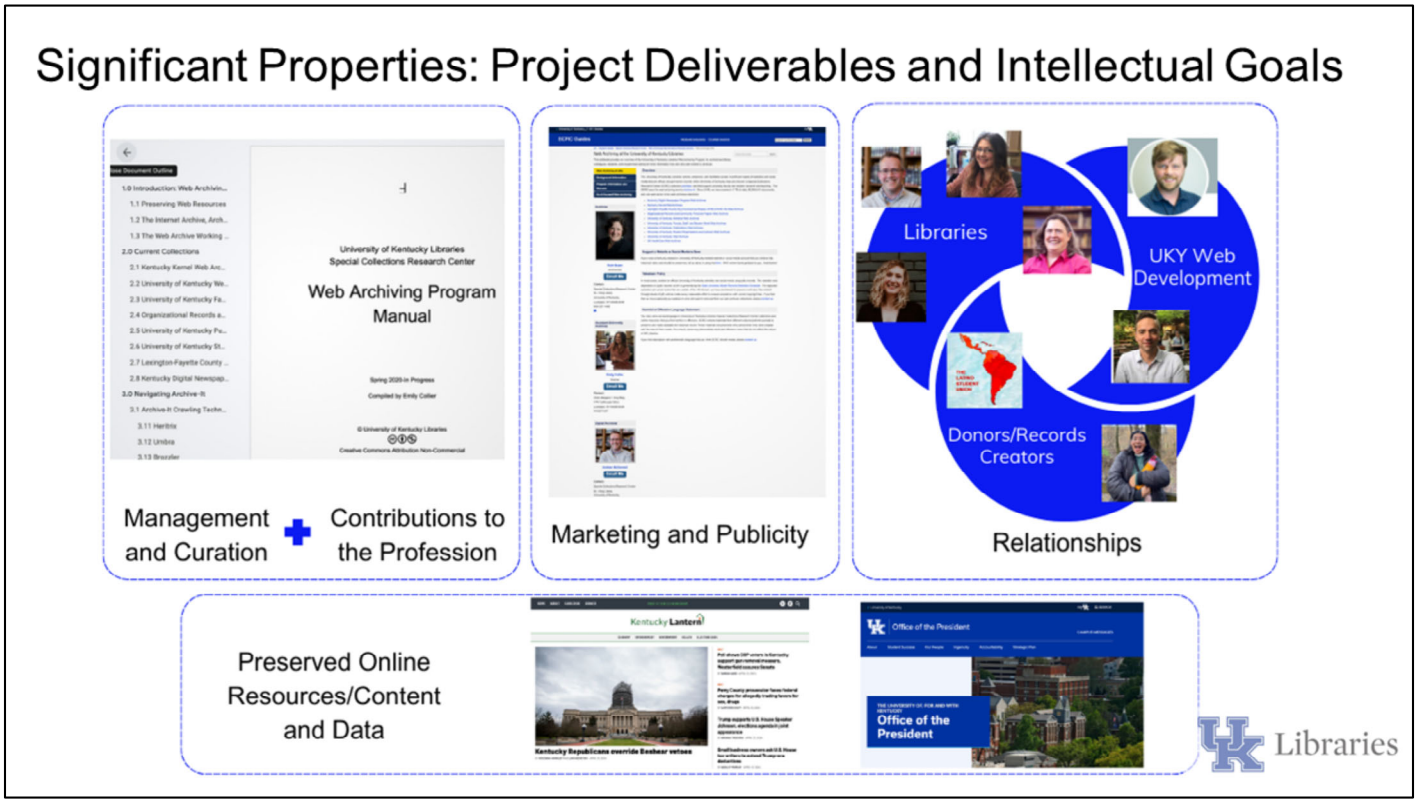
Access points:

- Archive-It
- Web archiving guide
- ExploreUK (digital library)
- Library catalog
- SCRC website





Our current and in development access points include (clockwise from top right): Archive-It, our library guide to web archiving, our digital library ExploreUK, our library catalog, and the Special Collections Research Center website.

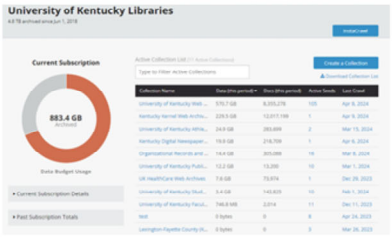


Our current and proposed future project deliverables and intellectual goals include macro and micro archiving of selected online resources for legal and cultural heritage reasons. We create and update policies, procedures, trainings, and workflows for web archival program management and curation. Our presentations and procedures contribute to the archival profession. We produce guides about the program for marketing, information, and publicity. And, we depend on and build relationships with university and community stakeholders, including university records creators, the centralized university web development team, and individual and organizational donors, who opt-in through deeds of gift to preserving their web content. We also offer an opt-out option through a take down policy, although we have yet to receive a request.

STSR Module A: Significant Properties

Libraries

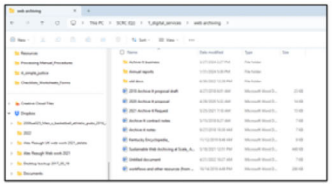
- Workflows and data flows



ARCHIVEWEB.PAGE
Interactive browser-based web archiving, a Webrecorder project.

Plausible

ArchivesSpace
a community served by Lyrasis

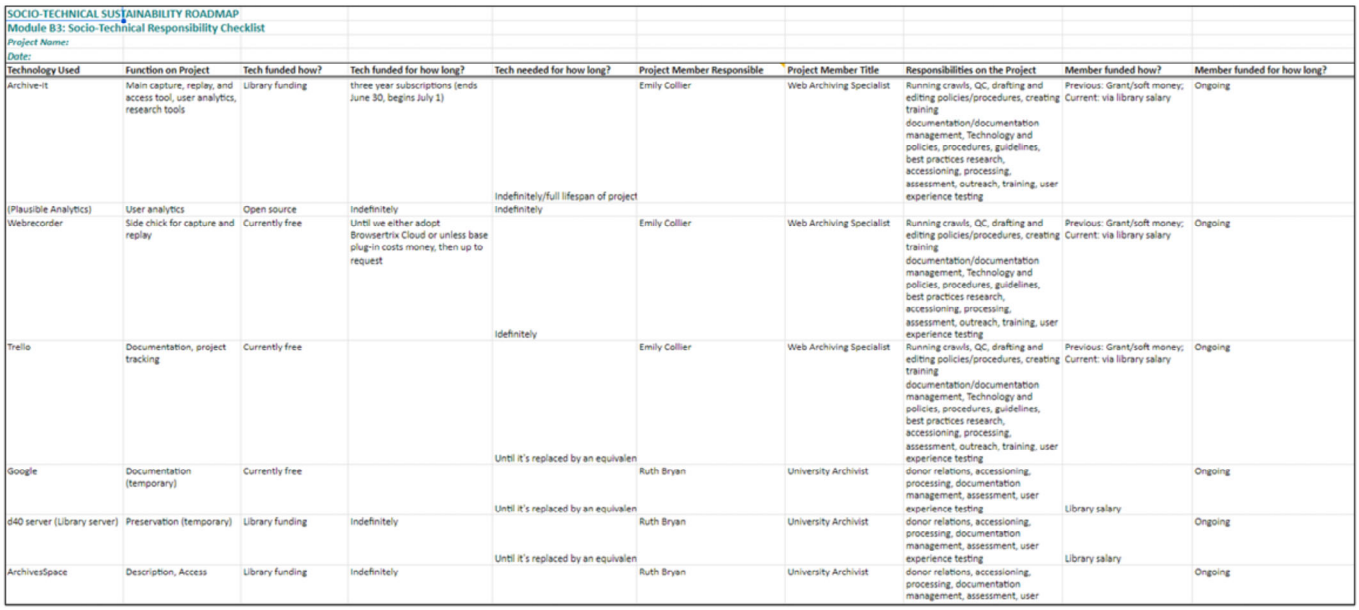


Drive

_zapier

Work and data flow through our web archiving program using a variety of platforms and tools, including (in clockwise order from the top) the Archive-It user interface for crawling, administration, and metadata, the Webrecorder archiveweb.page for crawling social media, our archival collection management system ArchivesSpace for collection metadata, Trello for project management, Google drive (which houses our working documents), and the Special Collections server partitions (where we store the final versions of our manuals and reports). In immediate future development is Zapier to automate data sharing between the university web development group and our web archiving Trello board. In the next 3 years, we would like to pilot the Archives Research Compute Hub for using web content as data and Plausible Analytics to better understand how our Archive-It content is used.

STSR Module B: Staffing and Technologies



SOCIO-TECHNICAL SUSTAINABILITY ROADMAP

Module B3: Socio-Technical Responsibility Checklist

Project Name:

Date:

Technology Used	Function on Project	Tech funded how?	Tech funded for how long?	Tech needed for how long?	Project Member Responsible	Project Member Title	Responsibilities on the Project	Member funded how?	Member funded for how long?
Archive-it	Main capture, replay, and access tool, user analytics, research tools	Library funding	three year subscriptions (ends June 30, begins July 1)	Indefinitely/full lifespan of project	Emily Collier	Web Archiving Specialist	Running crawls, QC, drafting and editing policies/procedures, creating training documentation/documentation management, Technology and policies, procedures, guidelines, best practices research, accessioning, processing, assessment, outreach, training, user experience testing	Previous: Grant/soft money; Current: via library salary	Ongoing
(Plausible Analytics)	User analytics	Open source	Indefinitely	Indefinitely					
Webrecorder	Side chick for capture and replay	Currently free	Until we either adopt Browsertrix Cloud or unless base plug-in costs money, then up to request	Idefinitely	Emily Collier	Web Archiving Specialist	Running crawls, QC, drafting and editing policies/procedures, creating training documentation/documentation management, Technology and policies, procedures, guidelines, best practices research, accessioning, processing, assessment, outreach, training, user experience testing	Previous: Grant/soft money; Current: via library salary	Ongoing
Trello	Documentation, project tracking	Currently free		Until it's replaced by an equivalen	Emily Collier	Web Archiving Specialist	Running crawls, QC, drafting and editing policies/procedures, creating training documentation/documentation management, Technology and policies, procedures, guidelines, best practices research, accessioning, processing, assessment, outreach, training, user experience testing	Previous: Grant/soft money; Current: via library salary	Ongoing
Google	Documentation (temporary)	Currently free		Until it's replaced by an equivalen	Ruth Bryan	University Archivist	donor relations, accessioning, processing, documentation management, assessment, user experience testing	Library salary	Ongoing
d40 server (Library server)	Preservation (temporary)	Library funding	Indefinitely	Until it's replaced by an equivalen	Ruth Bryan	University Archivist	donor relations, accessioning, processing, documentation management, assessment, user experience testing	Library salary	Ongoing
ArchivesSpace	Description, Access	Library funding	Indefinitely		Ruth Bryan	University Archivist	donor relations, accessioning, processing, documentation management, assessment, user		Ongoing

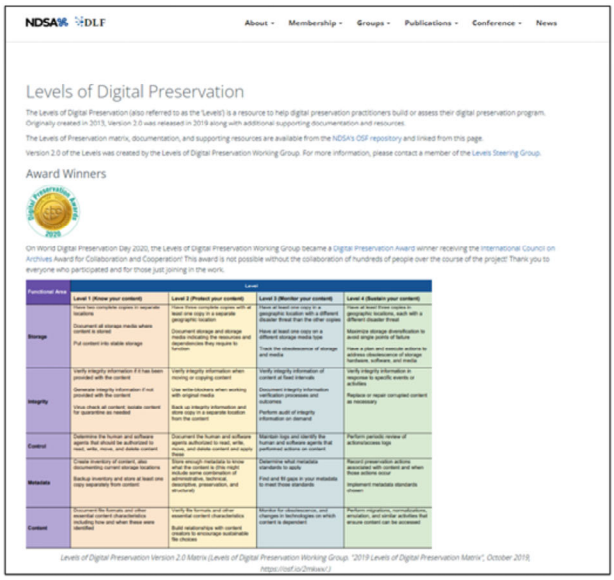
Section B prompts a consideration of who is on the project team and what their roles are and what the technological infrastructure of the project is. You then to develop a “Socio-Technical Responsibility Checklist” that maps staff onto their technology responsibilities, taking into account how and for how long staff and technologies are funded. For example, our main technology for web archiving is Archive-It, which is funded through the Libraries and which we anticipate needing indefinitely. Emily, as the web archiving specialist and assistant university archivist, has a variety of responsibilities for web archiving, and is currently in a permanent position funded by the Libraries.

Libraries

STSR Section C: Digital Preservation Plans/Checklist

- Access
- Backing up
- File Formats
- Metadata
- Permissions
- Data Integrity

	Level 1	Level 2	Level 3	Level 4
Access	Determine designated communities and significant properties Create and make available descriptive metadata, such as title, abstract, keywords, or other information that is useful for discovery	Ensure that designated communities can access significant properties of a project Have publicly available documentation, user guides, or other materials that make your work legible to users	Have a publicly available access and use policy	Provide access to the parts of the project that have become obsolete or difficult to access via a native environment and/or emulation



Section C prompts the project team to build digital sustainability plans using a modified version of the National Digital Stewardship Alliance's Levels of Digital Preservation. For each of the 6 functional areas, a table outlines a series of specific actions that can be taken to sustain digital projects with levels from 1, minimum, to 4, high.

(The NDSA is part of the Council on Library and Information Resources.)

STSR Section C: UKL results of Levels of Preservation

	1	2	3	4	5
Access Now					
Access Future					
Backing up Now					
Backing up Future					
Permissions Now					
Permissions Future					

	1	2	3	4	5
Metadata Now					
Metadata Future					
File Formats Now					
File Formats Future					
Data Integrity Now					
Data Integrity Future					

UK Libraries

This is a summary of our current and desired levels for access, backing up, permissions, metadata, file formats, and data integrity levels of preservation. In general, we were pleased to find that our levels are good for the small amount of staff we have doing web archiving. The file formats area didn't apply as the WARC/WACZ format is the only one currently. All these areas are high priority, and we would like to improve our sustainability across all. We added a level 5 for functional areas that we would like to sustain beyond or differently from the original levels.

STSR Section C5: Digital Sustainability Plan Checklist

Area	Action	Steps	Person	Time
Access, Metadata	Web archives part of search results in ExploreUK	Create and publish collection guides; allocate/find additional staff resources for description	Assistant Archivist + Archivist	2 years
Access	How to search web archives instructions	Write tips; publish on LibGuide; LibGuide connected to Archive-It account	Assistant Archivist + Archivist	2 months
Backing up, File Formats, Data integrity	Store WARC local copies in AWS Glacier	Request, approval, download, prep, ingest, accession	Archivist + Director of Information Technology + Digital Archivist	6 months
File Formats	Why loss of data uploading Webrecorder crawls to Archive-It?	Ask questions, read, attend conferences, explore other resources	Assistant Archivist + Digital Archivist	3 years

The team then created our Digital Sustainability Plan Checklist, the final outcome of the STSR. The checklist aggregates all the modules' work to create a comprehensive list of sustainability actions and their required staffing, technologies, and funding that can be accomplished within the three-year time frame. We identified 10 actions across all functional areas. Here, for example, are four of those. Some of them support sustainability in multiple functional areas.

Slide 14

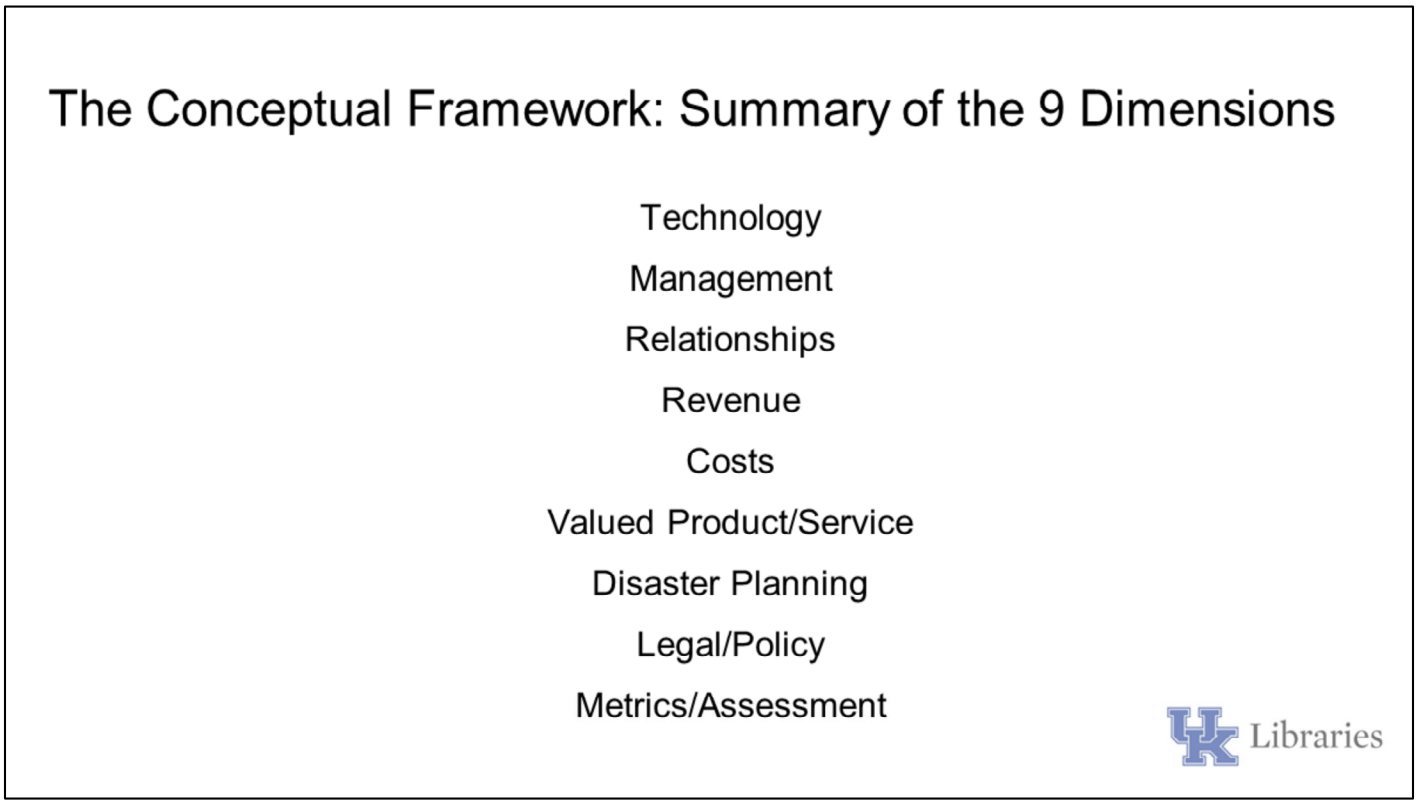
1 Now at 6:30 minutes from start to this slide.
Ruth Bryan, 4/14/2024

Conceptual: “A 9 dimensional framework for digital cultural heritage organizational sustainability”

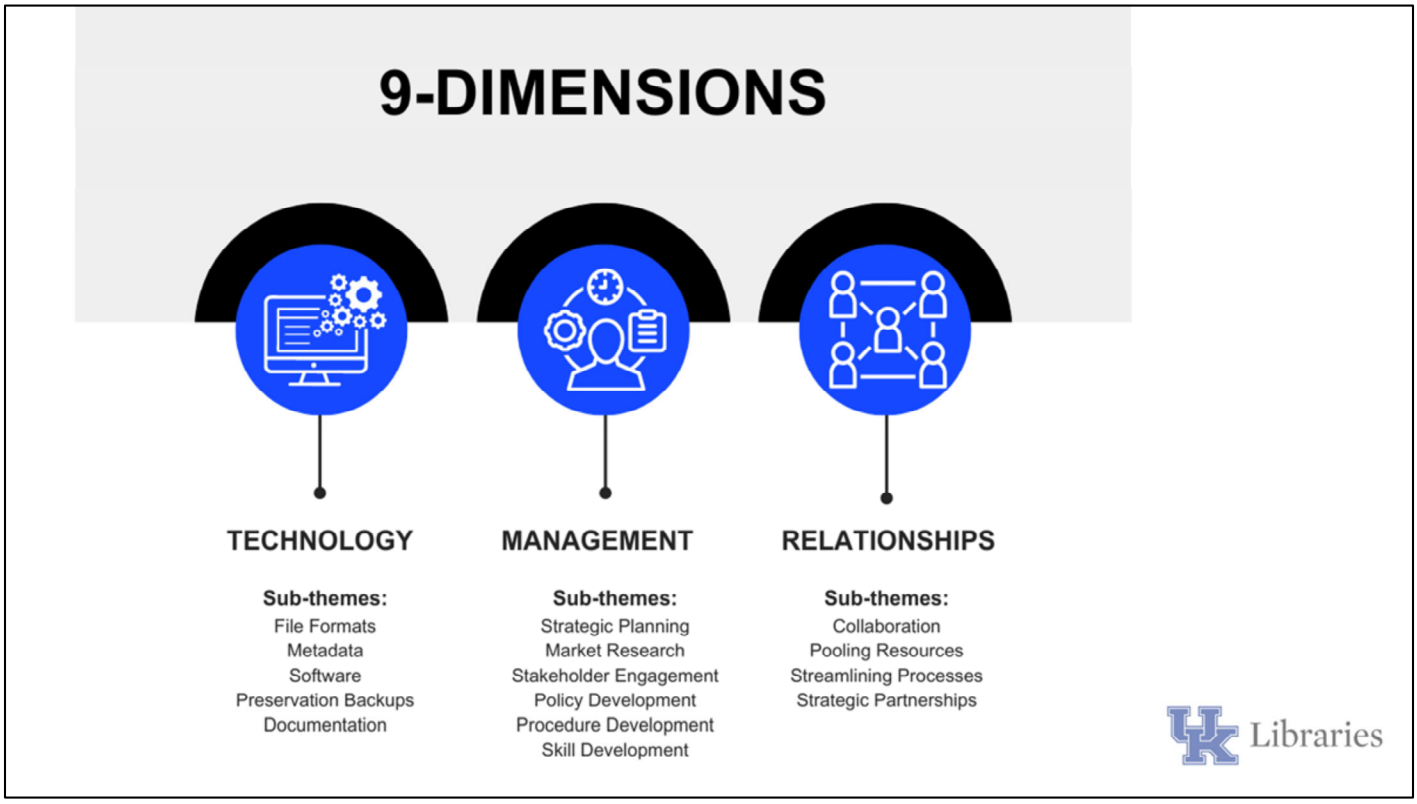
“...the arrangements of people and work practices that keep digital projects and services going over time, given ongoing challenges.”

Libraries

For our conceptual framework, we chose “A Nine Dimensional Framework for Digital Cultural Heritage Organizational Sustainability.” Eschenfelder and 5 other co-authors analyzed a sample of Library and Information Science literature from 2000-2015 and developed “definitions and coding rules for nine dimensions of organizational sustainability.” Their definition of sustainability is [quote].



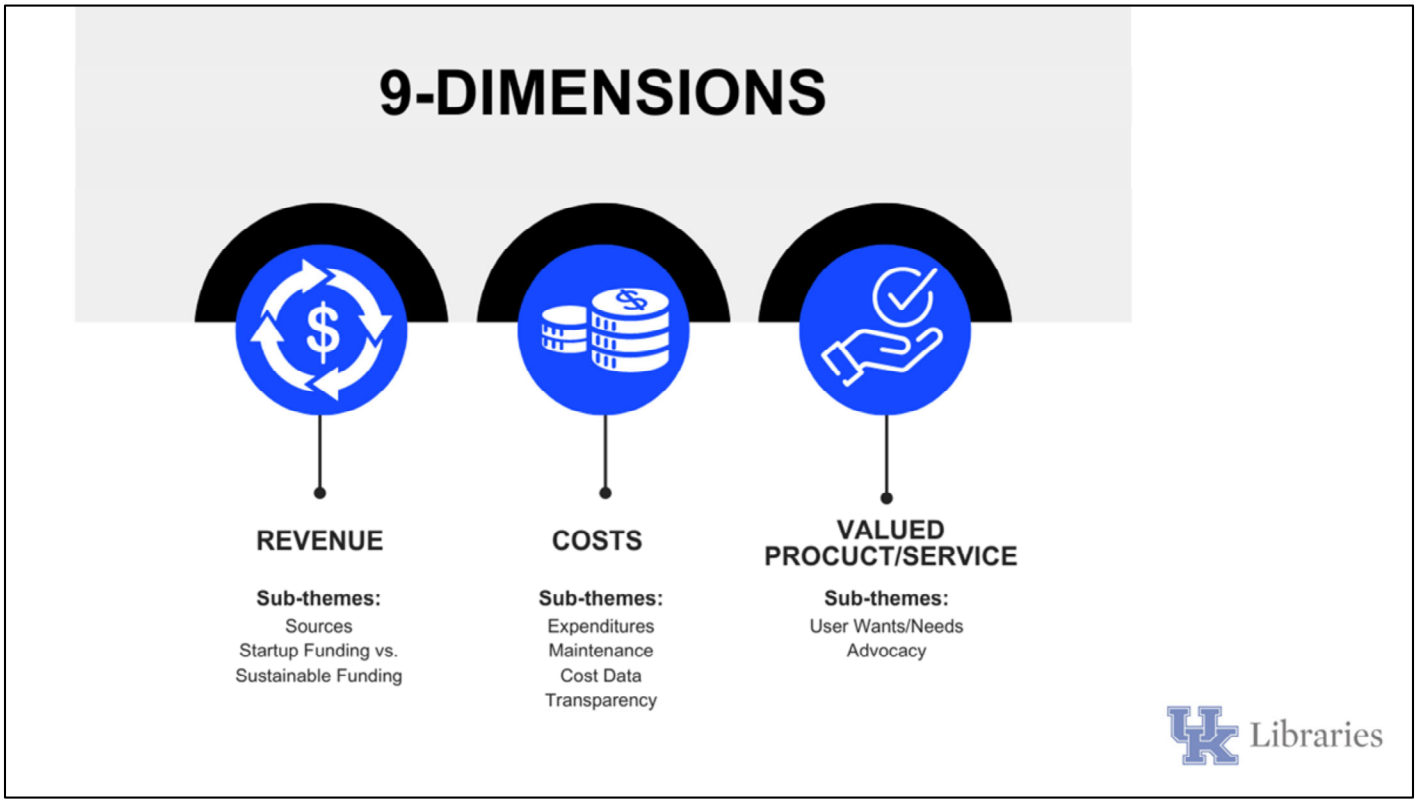
The 9 Dimensions that support this definition include technology, management, relationships, revenue, costs, valued product or service, disaster planning, legal and policy issues, and metrics or assessments. To summarize these briefly:



Our understanding of their discussion on technology included sub-themes like standards in areas like metadata and file formats; types and characteristics of software–whether commercial or open source, standard or common; backups/redundancies in preservation; and technology documentation

Management related to areas such as strategic planning, market research or marketing to identify audiences and services, stakeholder engagement, developing formal policies and procedures in contrast to ad hoc decision-making, and developing managerial skill sets.

Relationships dealt with areas like collaboration, pooling resources, streamlining processes, and strategic partnerships with resource-rich partners. These partnerships can allow organizations to reap the benefit of the “social capital” of that higher-resourced partner and their reputation.

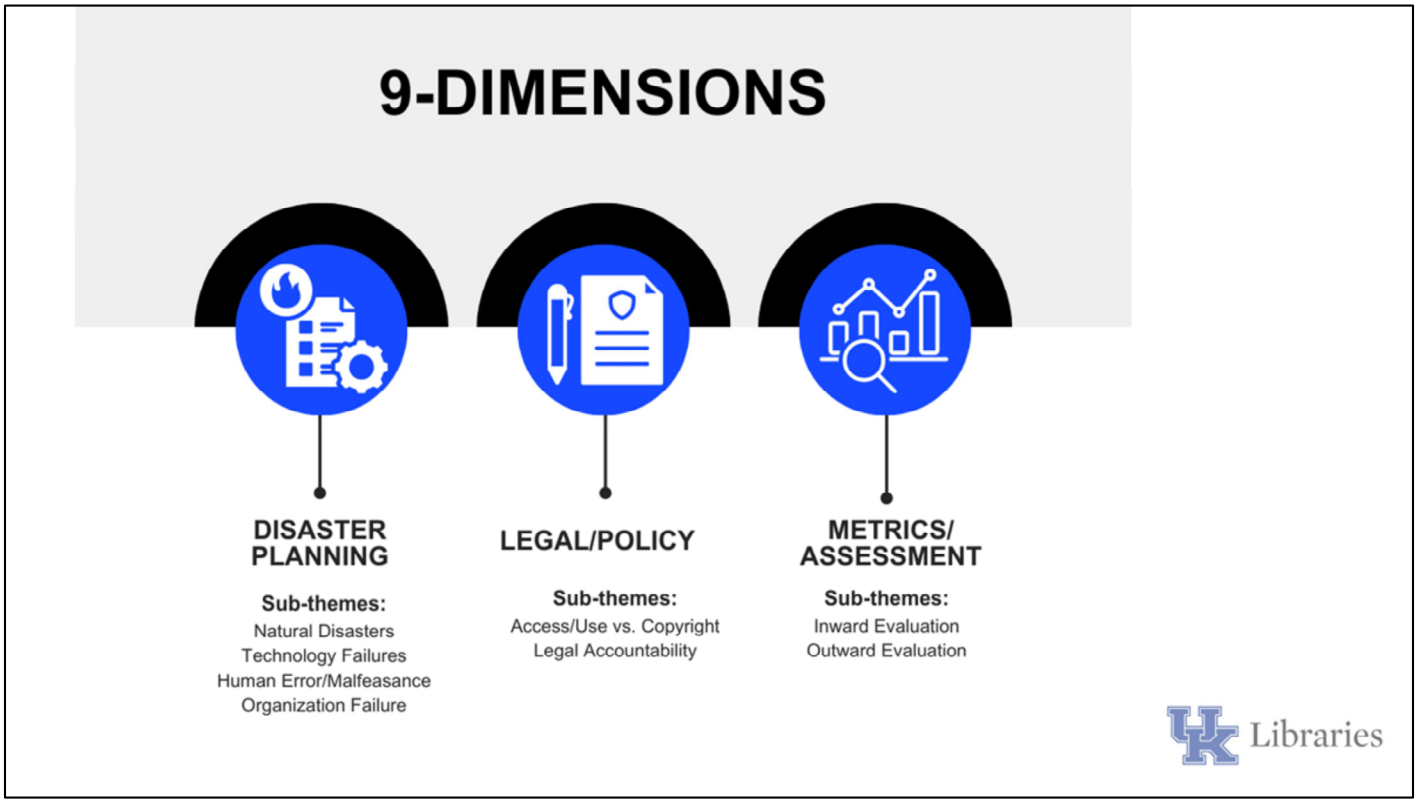


Revenue is required to sustain projects and can come from a variety of sources including endowments, donors, sponsorship, operating budgets, grants and more, and this is made easier with a “good reputation”. Important distinction that start-up funding is not considered the same as sustainable funding, so grant money for example, doesn’t weigh the same as operating budgets in terms of revenue.

Costs must be accounted for in relation to both expenditures and ongoing maintenance. Hard data is needed to plan for costs, but 9 dimensions also point out that while this can create transparency and accountability, this might actually hurt sustainability by making it easier for funders to make targeted cuts.

Valued product/service: User wants and needs must be kept in mind when offering tools and services, and getting continuous feedback is important because those needs are rarely static.

[An interesting sub theme discussed in relation to valued product/service was thinking about how value can lead people to new ways of using project data that were not conceptualized at the time that data was created.]

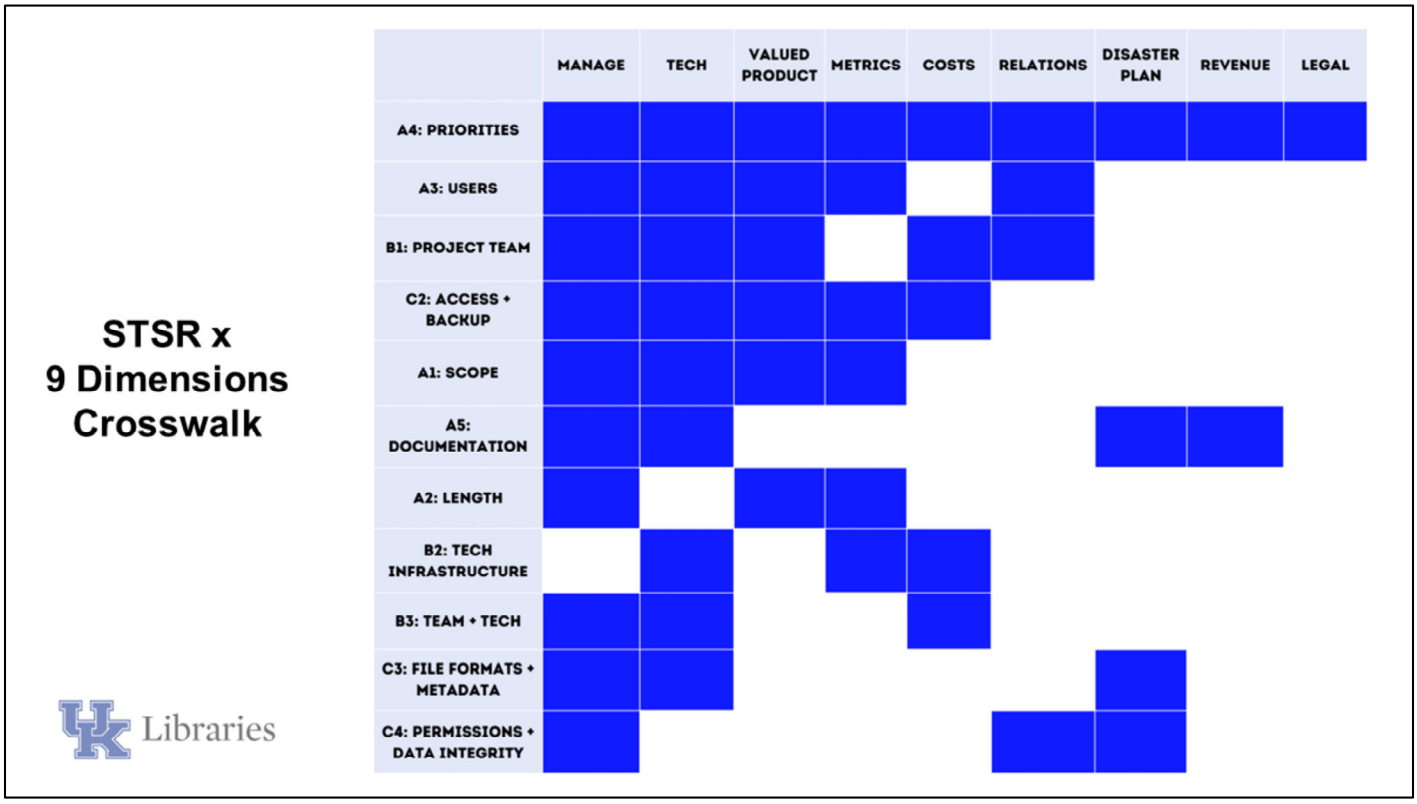


Disaster planning relates to sub themes like natural disaster, technology failures, human error or malfeasance, the cessation of the organization itself, and how to plan for the maintenance of your digital objects against these disasters.

Legal/Policy involves the balance of collecting, accessing and using works against copyright and other legal considerations.

And finally, metrics and assessment seeks project evaluation and can be inward facing, focusing on internal planning and risk management, or outward facing, demonstrating needs and impacts to stakeholders.

These 9 dimensions outline an idealized concept of what sustainability should look like. We took these ideas, and overlapped them with the STSR.

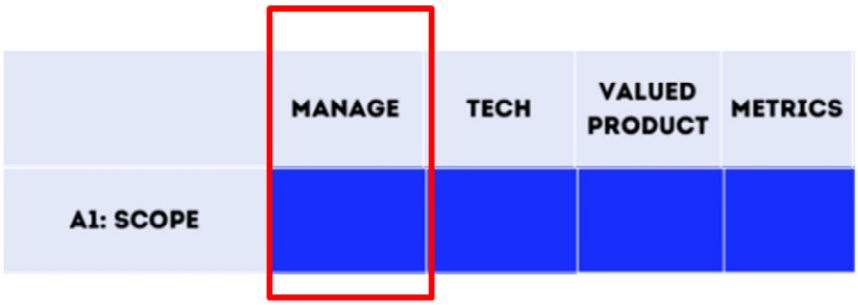


This Gantt chart represents the crosswalk, where the blue squares indicate the overlaps we found between the STSR and the 9 Dimensions. This led us to identify some pertinent issues in our web archiving sustainability, including gaps where we expected connections.

Slide 20

1 19-24 is 4:00 minutes.
Emily Collier, 4/15/2024

9 Dimensions x STSR Crosswalk: Pertinent Issues

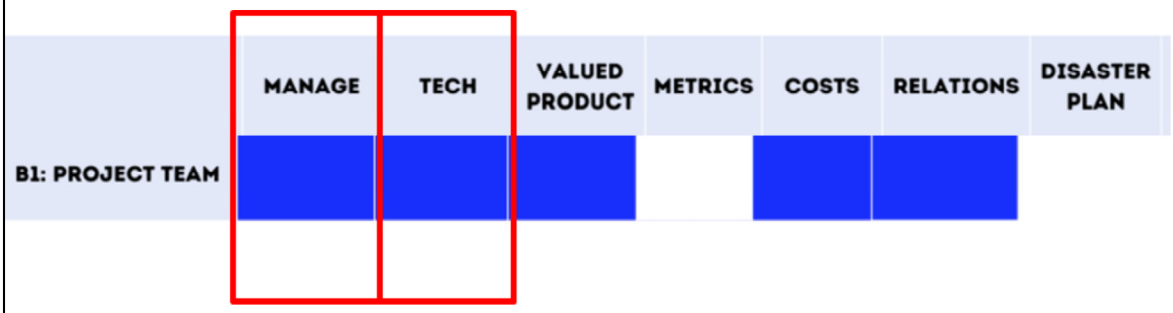


- Need to include planning for acquisition/selection and description changes.

Libraries

For instance, in section A1 of the STSR about project scope, we noticed that when considering project deliverables and team workflows, the management dimension doesn't include planning for changes in acquisition or description, which is a constant not just for web archiving, but digital archiving, generally.

9 Dimensions x STSR Crosswalk: Pertinent Issues

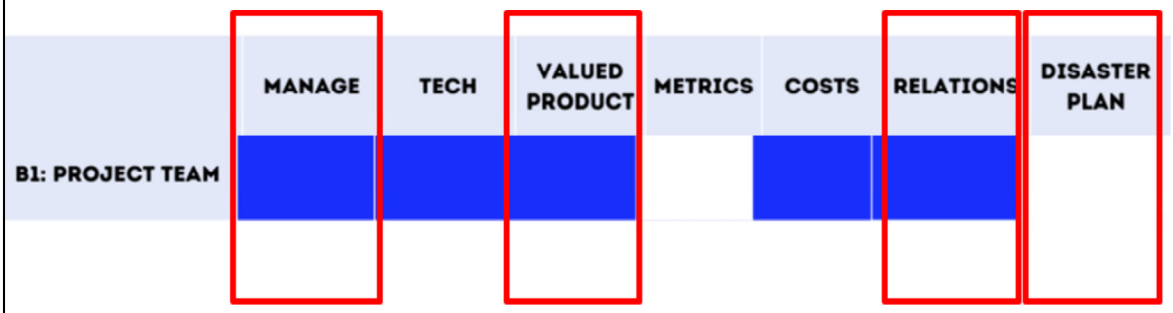


- Our management of web archiving is constrained by our single platform's development timeline and priorities.

Libraries

In section B1 about the project team, we noticed an interaction between the technology and management dimensions that isn't covered by the STSR. We benefit in many ways from the services Archive-It provides for web archiving, but as we rely on just the one platform, we are also constrained by Archive-It's capability and their development timeline and priorities, which might not align with our own.

9 Dimensions x STSR Crosswalk: Pertinent Issues

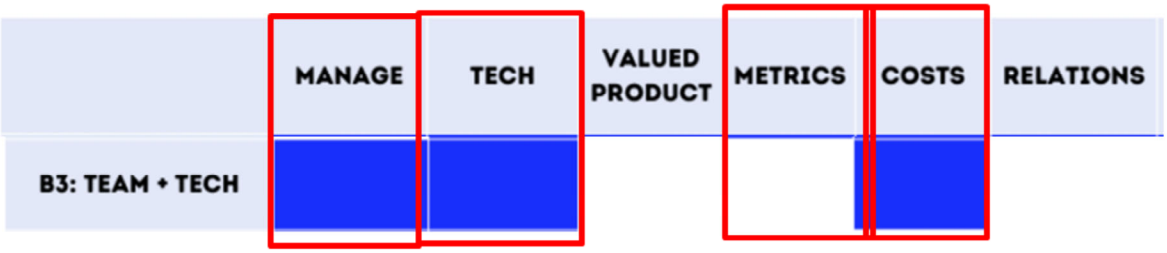


- Current Libraries staffing or administrative support could drop at any time.
- Partnership with University web development team relies on individuals' interest.
- Need to "sell" the value of web archiving

UK Libraries

However, the crosswalk brought to the forefront previously hidden interactions across the management, valued product, relationships, and disaster planning dimensions in relation to project team members. The web archiving program currently has Libraries administrative support, but it is in 3-year cycles. In addition, we have begun a collaboration with the university's central web development team. They will be providing us a master list of URLs and social media accounts in exchange for the value of the web preservation we provide which they can use for their own continuity of business needs. Both of these situations still rely on short-term support or on individual interest, which could stop or fall through at any time. We need to sell the value of web archiving to the extended team members and stakeholders we identified in module B to prevent the disaster of withdrawal of support.

9 Dimensions x STSR Crosswalk: Pertinent Issues

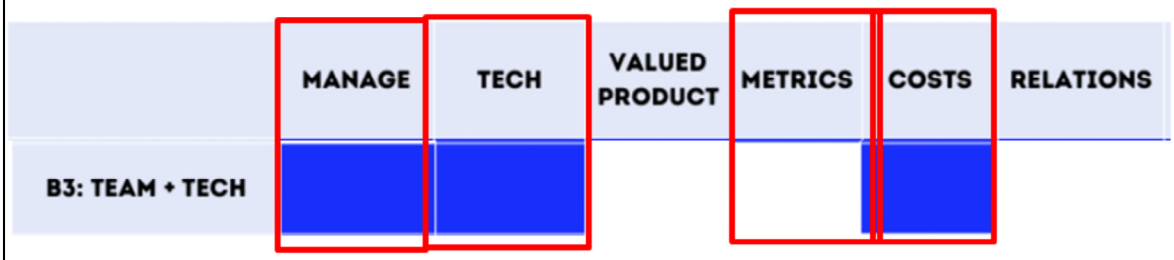


- Need to track and log not just costs for team and technologies, but also overall capacity supported by funding levels.
- Even with robust documentation, cost of web archiving includes ongoing research time and skill development.

Libraries

Moving on to section B3 of the STSR, in which we merged information about team members and their funding and our technologies and their funding, we noticed interactions among the management, technology, metrics and assessment, and costs dimensions that the roadmap didn't address. We need to track and log not just costs for the team and technologies, but also the overall capacity for web archiving that our current funding levels supports (i.e. there is a total cost of stewardship we need to calculate). For example, even with our robust curation documentation, team members need to continue learning and skill development, which adds to costs.

9 Dimensions x STSR Crosswalk: Pertinent Issues

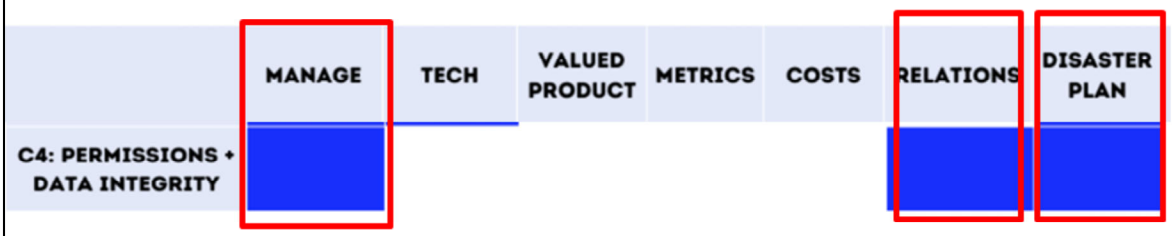


- Adding technology adds to total cost of stewardship–must update skills and develop and assess new workflows.
- Need to demonstrate the value of web archiving through use metrics.

UK Libraries

Adding new technology (for instance, our upcoming adoption of Zapier) adds to the cost of stewardship. Team members must not only update their skills, but also develop and assess new workflows. To sustain the web archiving program for the long term, we will need to demonstrate to our stakeholders, especially our funders, that the total cost of the program produces a valued product, and we'll need to expand our capacity to do this through gathering and analyzing usage and other metrics.

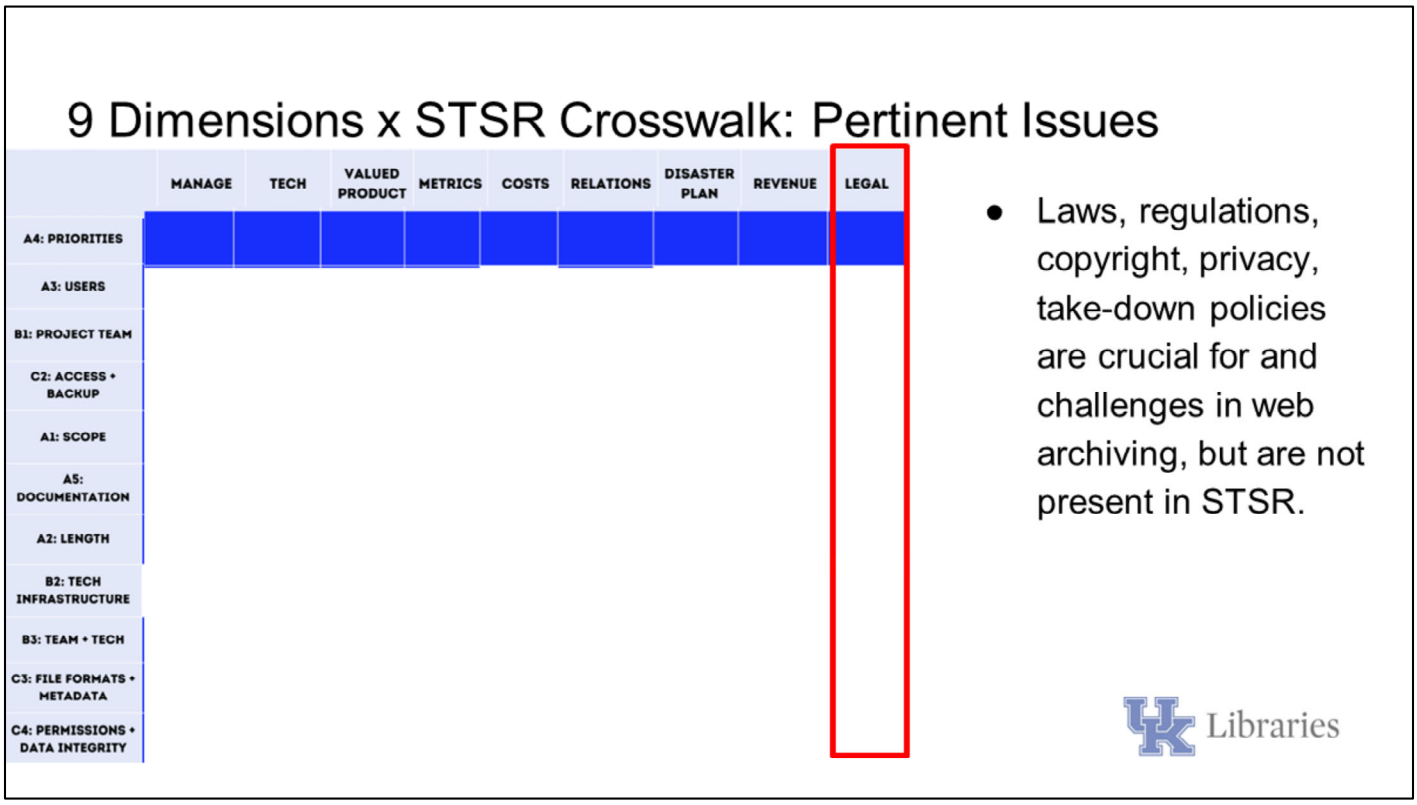
9 Dimensions x STSR Crosswalk: Pertinent Issues



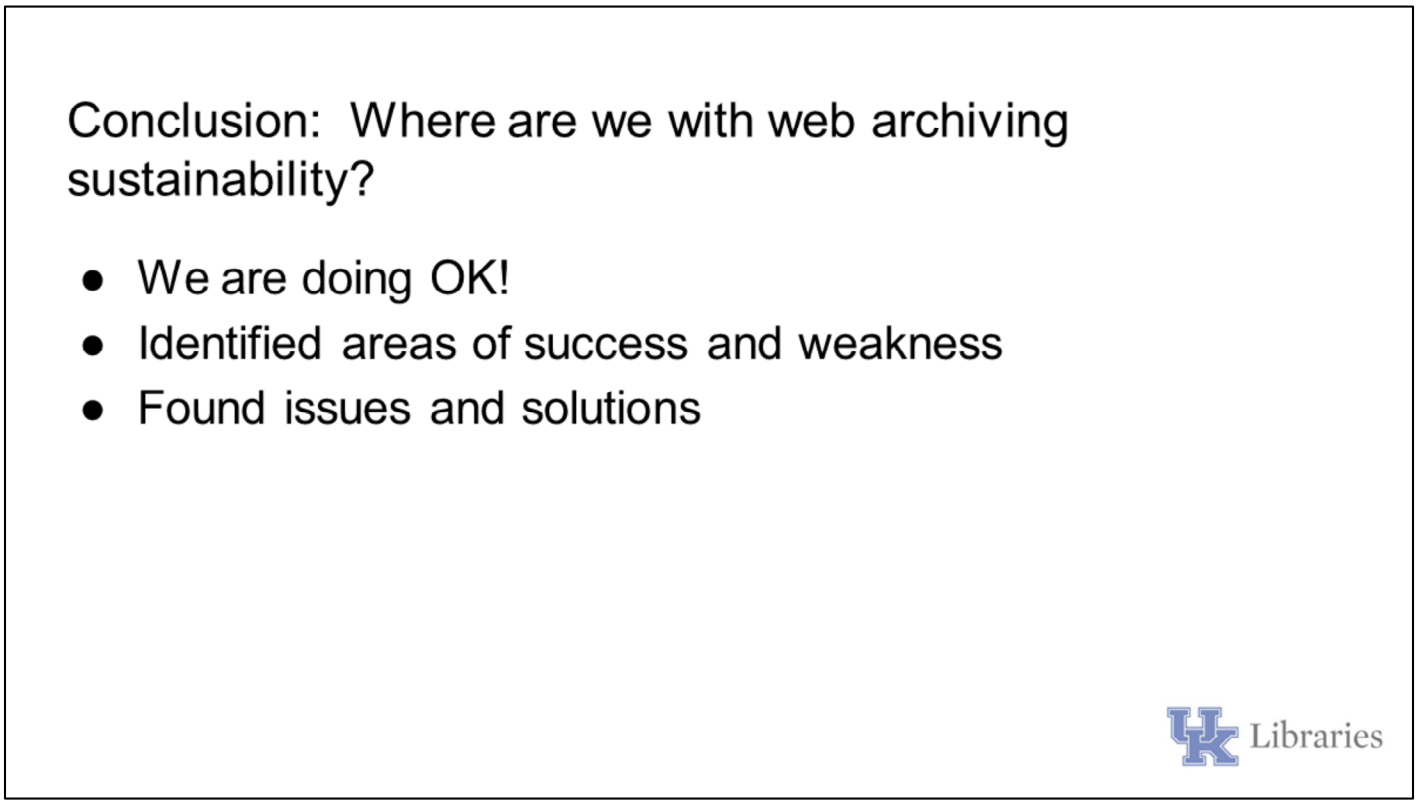
- Trust in existing and new team members is key.

UK Libraries

In the C4 section of the STSR, the review team evaluated our web archiving permissions functional area. We identified overlaps in this section among the management, relationships, and disaster planning dimensions. Specifically, trust in the integrity of existing and future team members and stakeholders is key to permissions and disaster planning. For example, we have given login permissions to the university web developers to our Archive-It account, trusting that they wouldn't compromise our work.



And, a final issue we noticed. The STSR doesn't include legal or policy issues as part of its sustainability considerations. Yet, laws, regulations, copyright, privacy, donor agreements, and take-down policies are crucial for and challenges in web archiving.



In conclusion, by going through this review, we found that our program is more sustainable than we had first thought.

By working as a team through both a practical and a conceptual sustainability framework and then crosswalking them, we identified areas of success and weakness, as well as issues we wouldn't have noticed otherwise. Finally, we discovered real, applicable solutions to some of our sustainability gaps.

Conclusion: Issues and Solutions

Issue	Solution	3-year steps
Complexity and constant change of uky.edu seed	• Build relationships to reduce staff effort/alleviate resources • Automation to reduce staff effort/alleviate resources	• Ongoing partnership with PR Marketing • Adoption of Zapier • Meetings with records creators
Reliance on a single preservation and access vendor/platform	• Back up WARCs locally • Download/create logs and reports • Documentation of rationale for user access levels • Research takedown/deletion options	• AWS Glacier storage • Workflow for creating/downloading and storing crawl logs (Webrecorder) and reports (Archive-It)

UK Libraries

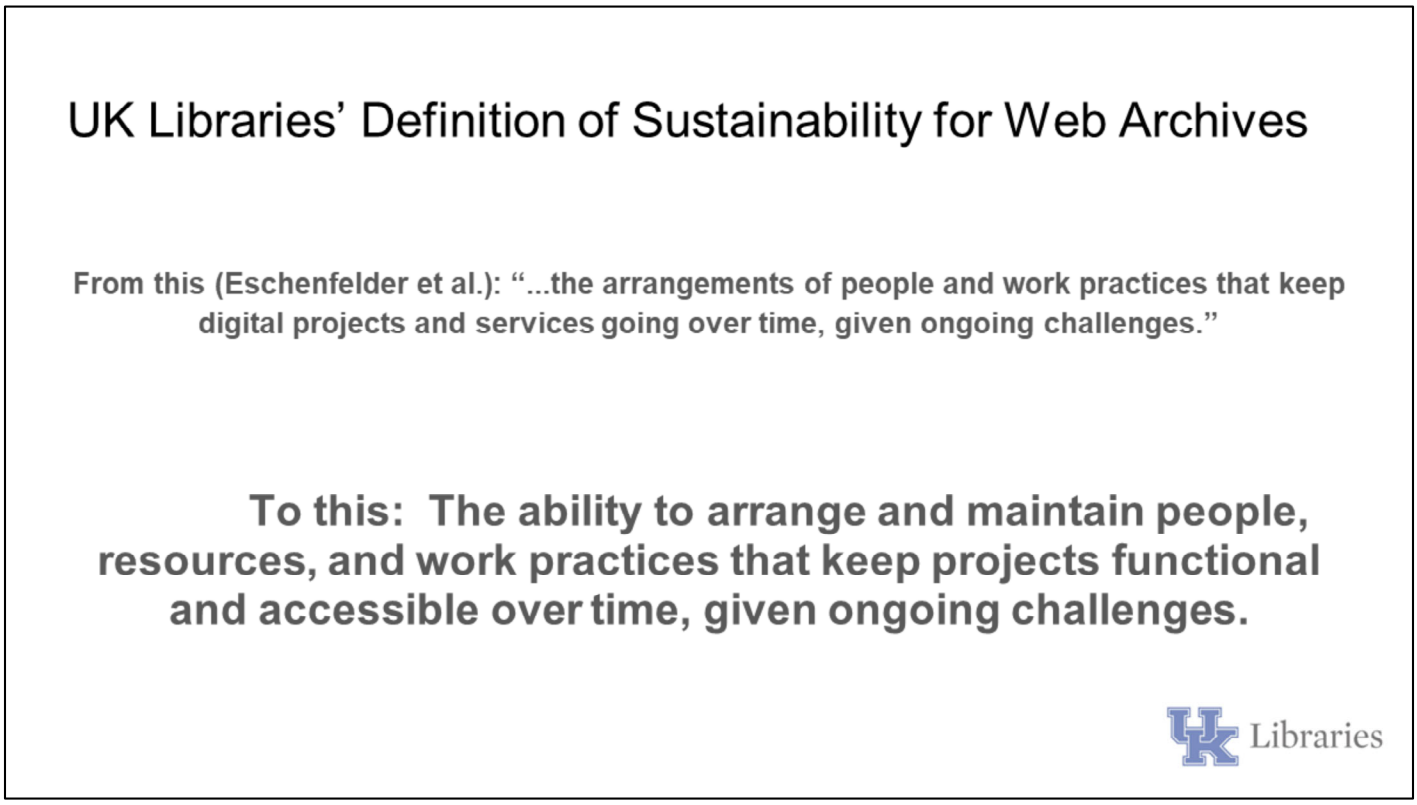
Issues include the complexity and constant change of the university website and reliance on a single vendor. Solutions include building relationships, automating processes, backing up WARCs locally, and expanding logs and documentation.

Conclusion: Issues and Solutions

Issue	Solution	3-year steps
Funding relies on individual support and goodwill	• Demonstrative value to stakeholders • Take advantage of short-term funding to build users	• Develop and use metrics tools (Plausible Analytics) • Seek ARCH temporary faculty position • Compete total cost of stewardship review
Access and discoverability is least developed area	• Add to current access points • Prioritize descriptive metadata • Advocate for and research better search functions	• Prioritize collection guides for ExploreUK by seeking more staff and shifting duties. • Add section on search to LibGuide • Pursue UK AI research option using local copies of WARCs.

UK Libraries

Other issues include relying on individual support and goodwill and underdeveloped access and discoverability. Solutions include demonstrating value to stakeholders by developing and using metrics tools and prioritizing creating descriptive metadata.



Finally, we have developed our own definition of sustainability for web archives, modified from the original Eschenfelder to: [quote]:

Resources that contributed to this work:

- Adam, U.A. and Kaur, K. (2023), "Empirical validation of IR sustainability model: leveraging on a PLS-SEM approach," *Digital Library Perspectives*, Vol. 39 No. 1, pp. 74-96. https://doi.org/10.1108/DLP-02-2022-0014
- DeRidder, Jody L. (2018). *Digital Curation Fundamentals*. Lanham, MD: Rowman & Littlefield.
- Eschenfelder, K. R., Shankar, K., Williams, R. D., Salo, D., Zhang, M., & Langham, A. (2019). "A nine dimensional framework for digital cultural heritage organizational sustainability." *Online Information Review*, 43(2), 182–196. https://doi.org/10.1108/OIR-11-2017-0318
- Visual Media Workshop at the University of Pittsburgh. "The Socio-Technical Sustainability Roadmap." Accessed March 2024. http://sustainingdh.net.
- Weber, Chela Scott, Martha O'Hara Conway, Nicholas Martin, Gioia Stevens, and Brigette Kamsler. (2021) *Total Cost of Stewardship: Responsible Collection Building in Archives and Special Collections*. Dublin, OH: OCLC Research. https://doi.org/10.25333/zbh0-a044.

Resources for web archiving produced by the University of Kentucky can be found at https://libguides.uky.edu/web_archiving

Contact us!

ruth.bryan@uky.edu ebcollier@uky.edu

UK Libraries

Thanks so much for listening, and we're happy to respond to your questions and comments.

Slide 32

---

2 Slides 23-35: 5 mins.

Ruth Bryan, 4/14/2024