

NCDS Final Report: Migrating digital research data across University of Michigan's Deep Blue repositories

Cate Gvon

August 14, 2026

Mentored by Joanna Thielen, Team Lead and Data Curation Specialist for Science & Engineering, University of Michigan Library

Human-data sharing and eligibility

A central question in evaluating the HeLa cell image collection was whether it could appropriately be migrated to Deep Blue Data as openly accessible research data. Because of the ethical and historical context of the HeLa cell line, the collection was reviewed against three policy frameworks: the NIH rules governing HeLa genomic data, the OSTP desirable characteristics for repositories storing human data as implemented by Deep Blue Data, and the U.S. Department of Justice Bulk Sensitive Data Regulations.

The first framework considered was the agreement between the National Institutes of Health and the Lacks family governing access to HeLa genomic sequence data. Under this framework, HeLa genome sequence data submitted to NIH are made available through the controlled-access database of Genotypes and Phenotypes (dbGaP), and requests for access are reviewed through the HeLa Genome Data Access Working Group. These protections were established in recognition of the circumstances under which Henrietta Lacks's cells were obtained and the potential privacy implications that genomic information may have for her descendants. However, the legacy collection examined for this project does not contain whole-genome sequence data, other HeLa genomic datasets, clinical information, or identifying information about Henrietta

Lacks or members of her family, so these access requirements therefore do not apply directly to this collection.

The collection was also reviewed against the Department of Justice's (DOJ) Bulk Sensitive Data Regulations. These rules restrict the sharing of certain categories and quantities of sensitive U.S. personal and government-related data. Deep Blue Data incorporates these requirements into its human-data review and does not accept deposits subject to the rule for unrestricted worldwide dissemination. Because the collection consists of microscopy images and experimental metadata rather than genomic sequences, biometric identifiers, health records, geolocation data, or personal identifiers, the regulations do not appear to apply in this case.

The final framework this collection was assessed against is Deep Blue Data's implementation of the Office of Science and Technology Policy (OSTP) guidance for storing human data. Deep Blue Data is an open-access repository and therefore does not rely on mechanisms such as tiered access or review of individual download requests. The repository only accepts human data for public sharing only when they are appropriately de-identified, present minimal risk if identity were inadvertently discovered, and are consistent with applicable consent or participant agreements.

The HeLa images themselves do not contain identifiable participant information and therefore meet the repository's expectations for de-identification and open access. The less clear issue is fidelity to consent because Henrietta Lacks's cells were originally taken without her consent, making it difficult to apply a contemporary consent-based standard straightforwardly to this historical material. However, the absence of original consent does not place these microscopy images within any human data sharing restrictions, nor does it make the images identifiable

human-participant data. Ultimately, Deep Blue Data may reasonably treat the collection as non-identifiable experimental imaging data suitable for open access, provided that curator review confirms the absence of genomic sequences, identifiable information, clinical data, and other regulated content. Rather than treating the lack of identifiable information as resolving issues around consent, the migration review could document the historical circumstances of the HeLa cell line and specify that the collection does not contain human genomic data or identifiable information.

Taken together, these three reviews support migration of the collection which is non-identifiable and does not contain information that would require restricted access. However, the expectation that human data be shared consistently with participant consent is difficult to reconcile fully with the documented history of the HeLa cell line.

Preserving scientific context

Although 1,344 separate records existed in Deep Blue Documents, those works collectively represent a coherent imaging experiment. Reproducing the same record structure in Deep Blue Data would reduce usability of the research dataset. This migration therefore provides an opportunity to consolidate the works into a single deposit while preserving the distinctions among the underlying image files. The individual Deep Blue Documents records contain substantial microscopy metadata, including 512×512 pixel dimensions, 12-bit CCD imaging, Cellomics KineticScan instrumentation, a $20\times$ objective and $200\times$ magnification, exposure times, fluorescence channels, probe concentration, and incubation and washout procedures. These details are necessary for interpreting the images and understanding how they were produced. Preserving this metadata also aligns with broader practices in the bioimaging community for

microscopy images. The migration should therefore preserve both the image files and the metadata necessary to interpret them. Where the original files use specialized or compressed formats, the original bitstreams should be retained as preservation masters.

Ethical description and provenance

Finally, the fact that the images can be shared openly does not remove the ethical significance of their provenance. The history of the HeLa cell line is inseparable from questions of consent, recognition, and the governance of human biological materials. For this reason, migration into a repository designed for long-term research-data stewardship creates an opportunity to describe the collection more explicitly and accurately.

A concise public note can acknowledge the origin of the cell line while also explaining what the collection actually contains. For example:

This collection contains microscopy images of cultured HeLa cells and associated experimental metadata. The HeLa cell line originated from cervical cancer cells taken from Henrietta Lacks in 1951 without her knowledge or consent. The collection does not contain human genomic sequence data, identifiable participant information, or clinical records.

Such a note gives users enough context to understand the provenance of the materials without conflating the ethical history of the cell line with the particular features of this dataset. It also makes explicit why the collection can remain openly accessible while still requiring thoughtful description.

Recommendations for migration of HeLa cells collection

Based on the policy review, metadata standards for microscopy images, and ethical considerations described above, the HeLa collection can be migrated into Deep Blue Data as an openly accessible research dataset, as the contents have been reviewed to confirm the absence of regulated or identifiable human data.

Recommended actions:

1. Consolidate the 1,344 Deep Blue Documents records into a single coherent Deep Blue Data deposit while preserving essential microscopy and experimental metadata.
2. Use CC BY 4.0 based on depositor responses to the migration and licensing request.
3. Preserve provenance by retaining original compressed files as preservation masters and documenting any derivatives.
4. Tombstone the original records and link them to the new Deep Blue Data DOI to maintain repository continuity.
5. Add a concise historical and ethical note acknowledging the origin of the HeLa cell line.
6. Document the open-access eligibility decision, including confirmation that the collection contains no regulated or identifiable human data.

Conclusion

This project demonstrates that migrating legacy research data is not simply a matter of transferring files between repository systems. Effective migration requires careful review of eligibility, metadata, licensing, depositor communication, and provenance so that datasets remain understandable, discoverable, and reusable in their new repository context. The workflow developed for moving research datasets from Deep Blue Documents to Deep Blue Data provides

a systematic approach for addressing these needs while preserving continuity with the original records.

The HeLa cell image collection also illustrates how migration can surface questions that extend beyond technical repository management. Consolidating the collection improves its usability, but responsible stewardship also requires preserving its scientific context and acknowledging the historical circumstances of the HeLa cell line. Together, these considerations show how repository migration can serve not only as infrastructure maintenance, but also as an opportunity to improve metadata, clarify provenance, and bring legacy research materials into closer alignment with current standards for access and stewardship.

Finally, this project benefitted enormously from the mentorship of Joanna Thielen, collaboration with colleagues Jenna Kincaid and Brittany Johns, and the support of Jeremy Morse, Peter Cerda, and Molly Hirst.

Funding

This project has been funded with Federal funds from the National Library of Medicine (NLM), National Institutes of Health (NIH), under cooperative agreement number UG4LM01234 with the University of Massachusetts Chan Medical School, Lamar Soutter Library. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.