**What is the GitHub Archive Program?**

* Its mission is to **archive and protect public open source code** so it isn’t lost over time — for reasons like media degradation, organizational shutdowns, or catastrophic events. [GitHub Archive Program](https://archiveprogram.github.com/)
* It recognizes that “as today’s vital code becomes yesterday’s historical curiosity, it may be abandoned, forgotten, or lost.” [GitHub Archive Program](https://archiveprogram.github.com/)
* The program stores multiple copies of code across different organizations, media formats, and geographies to increase durability and resilience. [GitHub Archive Program](https://archiveprogram.github.com/)

**Key features and components**

Here are some of the program’s flagship initiatives and strategies:

| **Initiative** | **What it is / how it works** |
| --- | --- |
| **Arctic Code Vault** | On **February 2, 2020**, GitHub took a snapshot of every active public repository and archived it onto a durable film designed to last up to 1,000 years. That archive is stored in a hardened archive deep within a decommissioned coal mine in **Svalbard, Norway** (Arctic region). [GitHub Archive Program](https://archiveprogram.github.com/) |
| **Greatest Hits archives** | GitHub works with libraries (e.g. Bodleian Library in Oxford, Bibliotheca Alexandrina in Egypt, Stanford Libraries) to store “Greatest Hits” — the most relied upon open source projects — in museum-quality archives around the world. [GitHub Archive Program](https://archiveprogram.github.com/) |
| **Partners & collaborators** | GitHub partners with preservation and archival organizations like **Software Heritage**, **Internet Archive**, **Long Now Foundation**, **Project Silica**, and others to build redundancy and resilience. [GitHub Archive Program](https://archiveprogram.github.com/) |

**Why it matters / significance**

* **Longevity of digital heritage:** Many critical software projects are open source (Linux, compilers, libraries). Losing them would mean losing major parts of modern infrastructure.
* **Protection from failure or disaster:** Code stored only on servers or conventional media is vulnerable to data center failures, bit rot, natural disasters, or organizational collapse.
* **Historical and academic value:** Future generations or researchers might want to look back at the evolution of software, open source practices, or code design trends.
* **Encouraging open source culture:** It underscores that open source is part of the shared cultural and technical heritage, not just a utilitarian license model.

If you like, I can dig up technical documentation or whitepapers about how GitHub encodes, preserves, and verifies the archived data (error checking, media durability, etc.). Do you want me to find that for you?

**You said:**

do a well tailored research about the topic in the link with credible sources and citations

**ChatGPT said:**

Thought for 29s

Great — here’s a focused, well-sourced research briefing on **GitHub’s Archive Program** (the site you linked). I explain what the program is, how it works (media, selection rules, partners), why it matters, strengths & limits, and where to read the original technical/primary sources. I cite the most important claims so you can follow up.

**Executive summary**

The **GitHub Archive Program** is a long-term preservation initiative that aims to protect public open-source software for future generations using a multi-layer archival strategy: (1) “hot/warm” live copies and web crawls (Internet Archive, Software Heritage), (2) periodic durable physical deposits (the **Arctic Code Vault** stored on Piql’s archival film in Svalbard), and (3) experimental ultra-long-term media (Microsoft Research’s **Project Silica**). GitHub took a comprehensive snapshot of active public repositories on **02 Feb 2020** (≈21 TB of data) and has distributed “greatest hits” copies to historic libraries for redundancy. The program follows LOCKSS/“pace layers” thinking (lots of copies, different media and update cadences) to maximize survivability and future recoverability. [GitHub Archive Program+1](https://archiveprogram.github.com/?utm_source=chatgpt.com)

**Mission & rationale**

* **Why archive GitHub?** Public repositories contain a major portion of the world’s software heritage — infrastructure, libraries, tools and documentation. GitHub’s program treats open-source code as cultural and technical heritage that could be lost to bit-rot, company failure, or catastrophic events, and so aims to preserve it using multiple redundant approaches. [GitHub Archive Program+1](https://archiveprogram.github.com/?utm_source=chatgpt.com)

**Core components & methods**

**1) The Arctic Code Vault (very-long-term physical cold storage)**

* **What happened:** GitHub captured a snapshot of active public repositories on **02/02/2020** and archived that snapshot to hardened archival film (stored in the Arctic World Archive in Svalbard, Norway). The deposited snapshot is reported as about **21 TB** of content, written on **~186 reels** of Piql archival film; data on the reels were encoded (QR-style) and include a human-readable index and recovery guide on each reel to help future readers decode the data. The archive is intended as a millennial-scale cold store (design life ≈1,000 years). [GitHub Archive Program+1](https://archiveprogram.github.com/arctic-vault/?utm_source=chatgpt.com)

**2) Greatest-Hits physical donations (museum-quality duplicates)**

* GitHub assembled a curated “Greatest Hits” set (the most-starred and most-depended-upon projects) and donated microfilm reels/boxes to major libraries (Bodleian Library at Oxford, Bibliotheca Alexandrina, Stanford Libraries, plus a GitHub display copy) to provide geographically distributed museum-quality redundancy and public stewardship. [GitHub Archive Program+1](https://archiveprogram.github.com/greatest-hits/?utm_source=chatgpt.com)

**3) Ongoing digital archiving (Software Heritage & Internet Archive)**

* **Software Heritage** ingests code archives and preserves full development histories; GitHub partnered with them to push GitHub content into a persistent, open archive.
* **Internet Archive** began archiving GitHub repositories (WARCs, Wayback etc.) as a warm-layer, publicly accessible backup that is updated more frequently. These accessible partners provide the “hot/warm” layers in GitHub’s multi-pronged strategy. [Software Heritage+1](https://www.softwareheritage.org/2019/11/14/github-helps-preserve-source-code/?utm_source=chatgpt.com)

**4) Experimental / ultra-long media: Microsoft Research’s Project Silica**

* GitHub partnered with Microsoft Research on **Project Silica** (data written into silica/quartz glass using femtosecond lasers) as an experimental path for storage potentially lasting many thousands of years; this adds media diversity beyond film and web crawls. [Microsoft+1](https://www.microsoft.com/en-us/research/project/project-silica/?utm_source=chatgpt.com)

**Technical details (encoding, readability, and access)**

* **Media & encoding:** Piql’s film (piqlFilm) encodes compressed data in machine-readable patterns (QR-like) plus a human-readable recovery guide. The design choice intentionally trades densest storage for extreme durability, readability without proprietary hardware, and resilience to electromagnetic hazards. GitHub’s public pages and Piql describe the film reels and the inclusion of recovery guides intended to help future generations decode the data. [GitHub Archive Program+1](https://archiveprogram.github.com/arctic-vault/?utm_source=chatgpt.com)
* **Snapshot inclusion rules:** The 2020 snapshot included every active public repo meeting GitHub’s snapshot criteria at the time (commits in specified windows, stars thresholds and so forth); by default public repos are included in the Archive Program unless the user opts out via repository settings. GitHub documents that partners can archive public repos via the public API. [Reddit+1](https://www.reddit.com/r/programming/comments/hstcpi/github_achives_all_of_the_repositories_present_on/?utm_source=chatgpt.com)

**Preservation strategy & governance**

* **LOCKSS + pace layers:** GitHub explicitly adopted a multi-layer “pace layers” / LOCKSS approach (Lots Of Copies Keep Stuff Safe): hot (live GitHub, torrents), warm (Internet Archive, Software Heritage), and cold (Arctic Code Vault, Project Silica). Different layers are updated on different cadences (near realtime → monthly/yearly → multi-year). This diversity reduces single-point-failure risk and improves the odds that code remains recoverable with future technologies. GitHub also works with library partners and preservation specialists for governance and access stewardship. [WIRED+1](https://www.wired.com/beyond-the-beyond/2019/11/dead-media-beat-long-nowgithub-software-archive?utm_source=chatgpt.com)

**Strengths & benefits**

* **Redundancy across media and geography** (Svalbard + major libraries + online archives) reduces risk of total loss. [GitHub Archive Program](https://archiveprogram.github.com/approach/?utm_source=chatgpt.com)
* **Human-readable recovery guides** and non-proprietary encodings on film increase the chance future readers can decode the data even if specific technologies vanish. [GitHub Archive Program](https://archiveprogram.github.com/arctic-vault/?utm_source=chatgpt.com)
* **Partnership with established archives** (Software Heritage, Internet Archive, Bodleian, etc.) binds technical expertise and public trust into the program. [GitHub Archive Program+1](https://archiveprogram.github.com/partners/?utm_source=chatgpt.com)

**Limits, risks & criticisms**

* **Single-time snapshot vs continuous preservation:** the Arctic deposit is a snapshot (02/02/2020); it is a powerful time capsule but not a substitute for continuous archival ingestion. GitHub’s plan is to update deposits periodically, but the snapshot itself will reflect the software landscape at one point in time. [GitHub Archive Program](https://archiveprogram.github.com/arctic-vault/?utm_source=chatgpt.com)
* **Readback & future decoding complexity:** while the film and index are designed for human readability, recoverability still depends on future parties preserving the necessary interpretative documentation and capability. Experimental media like Project Silica require future decoding pipelines and preservation of reading technologies/algorithms. [Microsoft+1](https://www.microsoft.com/en-us/research/project/project-silica/?utm_source=chatgpt.com)
* **Legal / copyright & access considerations:** distributing copies of repositories into multiple archives raises potential copyright and license management questions (who can access, under what terms). Partner archives like the Internet Archive have faced legal challenges on other collections — this is a non-trivial governance area for long-term access. [Reuters](https://www.reuters.com/legal/litigation/internet-archive-must-face-record-label-copyright-claims-judge-rules-2024-05-16/?utm_source=chatgpt.com)

**Practical implications (for developers / project owners)**

* **Public repositories are included by default.** If you don’t want your repository included in a physical snapshot, GitHub docs explain opt-out/control settings (check repo visibility and Archive Program policy). For canonical policies see GitHub docs and the Archive Program pages. [GitHub Docs+1](https://docs.github.com/en/repositories/archiving-a-github-repository/about-archiving-content-and-data-on-github?utm_source=chatgpt.com)
* **If you depend on reproducibility or provenance,** keep separate project metadata, README, build instructions and non-binary source artifacts in the repo — these increase the value of archived code for future readers. The Archive Program’s human-readable index is helpful, but rich repository metadata improves recoverability.

**Primary/credible sources (read next)**

* GitHub Archive Program — official site and approach pages. [GitHub Archive Program+1](https://archiveprogram.github.com/?utm_source=chatgpt.com)
* Arctic Code Vault page (technical summary of the Piql reels deposit). [GitHub Archive Program](https://archiveprogram.github.com/arctic-vault/?utm_source=chatgpt.com)
* Piql announcement on storing GitHub data to piqlFilm. [Piql](https://www.piql.com/news/githubs-code-now-stored-forever-on-piqlfilm/?utm_source=chatgpt.com)
* Microsoft Research — Project Silica project page (experimental quartz/glass archival technology). [Microsoft](https://www.microsoft.com/en-us/research/project/project-silica/?utm_source=chatgpt.com)
* Software Heritage — announcement and role archiving GitHub content. [Software Heritage](https://www.softwareheritage.org/2019/11/14/github-helps-preserve-source-code/?utm_source=chatgpt.com)
* Coverage & analysis (Wired / Long Now / news outlets) on pace layers / LOCKSS context. [WIRED+1](https://www.wired.com/beyond-the-beyond/2019/11/dead-media-beat-long-nowgithub-software-archive?utm_source=chatgpt.com)

**Short conclusion**

GitHub’s Archive Program is a credible, multi-partner, multi-media preservation effort that balances immediacy (Internet-accessible archives) with extreme longevity (Arctic film + experimental glass). Its design (multiple copies, different media, human-readable indexes, trusted library partners) follows established archival principles and significantly raises the prospects that public open-source software will survive for centuries. However, the Arctic deposit is a **snapshot** (2020) and does not replace the need for continuous, curated, legally aware digital preservation strategies; readback and governance remain active challenges.