JIM COWIE, INTERNET HISTORY INITIATIVE

Envisioning an Internet Data Trust

Strategies for Century-Scale Preservation of Infrastructure History

January 2026

Executive Summary

The Internet's infrastructure history is the story of the evolution of the 21st century. Over five decades, that story has been written in operational and measurement datasets collected by a diverse range of stakeholders: researchers, network operators, commercial interests, governments, and civil society organizations. The insights from these multi-decade holdings will underpin the next century of critical research in public policy, security, economics, development, and the history of technology. Yet despite their importance, much of this Internet measurement legacy – our century's digital heritage – remains fragile, fragmented, and at continual risk of loss.

This report argues that existing approaches to preserving Internet measurement and infrastructure data are no longer adequate to the scale, complexity, and longevity of the challenge. The institutions around the world that hold this data are coming under increasing pressure: vulnerable to funding cycles, legal uncertainty, and generational loss of expertise. As a result, irreplaceable records of the Internet's evolution are at continuous risk of being lost in a disaster or quietly deleted to save money.

To address this gap, this report proposes the establishment of an **Internet Data Trust (IDT):** a dedicated, independent fiduciary institution charged with the long-term preservation and stewardship of Internet measurement and infrastructure data in the public interest. The Trust would operate under a legally binding mandate to preserve designated datasets over century-scale time horizons, while respecting the sovereignty, rights, and obligations of contributing institutions.

The Trust is intentionally narrow in purpose. Its core responsibility is preservation: ensuring that critical historical datasets remain intact, verifiable, and accessible to future generations of researchers and policymakers. The Trust would function as a cooperative safety net, providing durable legal, financial, and technical infrastructure that individual institutions cannot sustain on their own.

Alongside the preservation function, the report envisions a complementary operating entity — the **Internet Data Initiative (IDI)** — to support research access, capacity building, and methodological continuity. The Initiative would curate a living research collection derived from preserved holdings, helping to transmit community expertise across generations of researchers.

The report argues that a trust-based institutional model offers key advantages over potential alternatives. By focusing on fiduciary stewardship, the Trust can align preservation obligations with appropriate legal and financial structures. By carefully mapping capabilities and needs across a federated community of contributing institutions, it can serve as a coordinating point for mutual assistance, bringing preservation resources to bear where they are most urgently needed. And by embedding public-interest duties at the core of its governance, it can maintain legitimacy across jurisdictions and over time.

The remainder of the report proceeds as follows. It first surveys the current preservation landscape, and the failure modes that put historical datasets at risk. It describes feedback from an initial December 2025 convening in Barcelona, at which more than 60 members of the global Internet measurement community gathered with professional librarians and archivists, shared their institutional experiences, and provided feedback on an initial set of community principles to guide the formation of the Trust.

Subsequent sections translate these principles into proposed governance structures, examine potential alternative forms of legal organization, and outline the Trust's financial and organizational path to launch. This path combines catalytic startup funding, a long-term preservation endowment, and recurring operational income — together forming a structure designed to support preservation operations despite a century of economic and political volatility.

Taken together, the report advances a single core claim: that preserving the Internet's measurement and infrastructural data history is a public-interest obligation that can be addressed, in best Internet tradition, through collaboration among a broad community of stakeholders, within institutional forms explicitly designed for longevity.

Contents

An Internet Data Trust constitutes a collective insurance policy for the Internet's historical measurement data legacy.

Introduction

As the Internet enters its second half-century, societies are struggling to respond to rapid changes in the world's interlocking complex systems. The Internet was not the sole trigger, but its global spread has clearly catalyzed a broad range of interrelated technological and societal changes, from the emergence of AI to a crisis of confidence in traditional institutions, all mediated by historical processes that may take generations to fully appreciate.

Meanwhile, as part of the explosion of change ushered in by the Internet, the tools of digital preservation and data science are evolving at a furious rate. Fueled by scalable storage and computational resources, open data, and open tools, the next decade has the potential to connect a vastly expanded community of researchers to the Internet's history, allowing scientists to seek firmer conclusions about the causal relationships that have connected technological change, global networking, and sustainable development goals. A new generation of scalable data science and artificial intelligence tools have the potential to make data-driven history legible to policymakers and social scientists who would previously have struggled to access the relevant information buried in exabytes of technical data.

What We Stand to Lose

We live in an age of precarity, in which the Internet's custodial institutions may face fundamental, or even existential, challenges. Public funding for measurement research is under pressure, even as our ability to instrument the Internet, collect operational telemetry, and mine it for insight has grown hugely. And as measurement data grows relentlessly over time, custodial institutions must weigh the costs of preservation of decades-old data against their responsibilities to present-day stakeholders.

Those decades of accumulated data —the Internet's measurement and infrastructural history — constitute the primary sources that underlie our institutional memory of how we arrived at this point in the Internet's evolution, from the early days of the dot-com expansion to today's landscape of hyperscale cloud services and a largely mobile edge. These datasets record in detail how the Internet grew with each passing year: how cooperation and competition emerged among early Internet service providers, how traffic flowed within and between world regions as consumer demand grew, where the network was slow and where it was fast, where services and content were hosted and how they were consumed. These original measurements of a constantly changing global Internet are irreplaceable resources: once lost, they can never be performed again.

Collective Opportunity

In the pages that follow we will survey this landscape, sketching the potential to create a new kind of global institution – an **Internet Data Trust** – to

preserve the Internet's measurement history and make it available to the researchers who will one day study the Internet's impact on human society in our century.

An Internet Data Trust could serve as the locus for creative and alternative preservation and interpretation strategies, provide assistance and support to custodial institutions, and preserve critical historical datasets in the public interest. Assigned the mission of century-scale preservation, and entrusted with sufficient resources to curate those datasets and make them legible to future researchers and historians, such a Trust would constitute a collective insurance policy for the Internet's historical measurement data legacy.

This report was compiled in late 2025 with the support of the Internet Society and with input from a diverse range of institutions that maintain the Internet's historical data. It surveys the challenges and suggests one way forward, incorporating extensive feedback from a December 2025 workshop held during ISOC's Pulse Measurement Week in Barcelona, in which more than 60 members of the Internet Measurement community considered the preservation mission and refined the Trust's core principles. This report examines in turn:

- The urgency of the need for an Internet Data Trust, and the research opportunities that such a Trust would unlock;
- The stakeholder landscape: the primary kinds of organizations that would form the Trust's membership community;
- The articulation of the Trust's mission, and the key operating principles that would guide the Trust in carrying out that mission;
- The technical and operational challenges facing such a Trust as it carries out its century-scale preservation mission;
- The legal and governance aspects of the Trust, anticipating the ways that the membership would deal with potential challenges to Trust operations, and finally
- The sustainability strategy for the Trust, enumerating the kinds of catalyst funding and long-term capital resources that the Trust would need to access to sustain century-scale operations.

Crisis and Urgency

For 50 years, network operators and researchers have probed and measured the Internet's connectivity and performance. From the very beginning, they recorded how organizations connected to each other, how they named their servers and their services, how packets traversed submarine and terrestrial cables to knit the world together, and how network operators gathered at Internet exchange points to build an Internet that was more resilient, more intricately interconnected, and more secure, reaching more people with every passing year.

But the accumulated records of their observations also grew and grew, each observation an irreplaceable point-in-time measurement of a constantly changing global network. And in a world obsessed with real-time monitoring and millisecond-scale optimization, the value of the historical perspective locked away within exabytes of this decades-old technical data has sometimes been unclear. This has created a crisis of legibility: absent a clear, motivating story connected to societal themes, these scattered historical repositories are not serving as effective advocates for their own preservation.

Meanwhile, with each passing year, the experts who collected and worked with these datasets for their entire careers are coming closer to retirement. Freshly minted PhDs who entered Internet studies in the 1990s are now nearing retirement, and their advisors have long since left the stage. There's a real risk that as these experts disappear from the research conference circuit, their tacit knowledge and their tools will vanish, along with their

websites[1], and ultimately the datasets themselves, through accident or neglect or someone's desperate attempt at cost savings.

First opportunity: coordinate for mutual aid, avoid extinction events

At this particular moment in history, fragmentation of preservation efforts, resource exhaustion, and the loss of knowhow and context are critical problems for the Internet community to solve, if the lessons from the Internet's engineering history are to be preserved for future generations. The Trust's motivating vision therefore begins with the creation of a coordinating layer among the Internet's data-custodial institutions, creating a durable legal and technical trust framework that can reduce friction, lower the risks of loss for the Internet's digital heritage, and make century-scale cooperative preservation economically and politically feasible, while throwing a lifeline to the data holding organizations that are struggling to preserve the Internet's endangered history.

Second opportunity: organize a living repository for a new century of research

By itself, however, long-term preservation is an abstract goal. The Trust also provides a unique opportunity to use century-scale preservation as a concrete motivator for present-day change. Finding a new home for the Internet's endangered datasets also creates the opportunity to reorganize and optimize that data for public access, coordinate the efforts of its present-day custodians, and make investments to make historical datasets legible to present-day audiences and more broadly accessible to present-day researchers, even as it's being set aside for future historians. A functional Research Collection should complement the Preservation Collection, actively probing and testing our assumptions about the utility and long-term legibility and accessibility of the datasets we are preserving.

In this vision, the Trust becomes a fresh convening vehicle, tapping into the energies and skills of historians, economists, archaeologists of the contemporary, librarians and digital preservationists: exactly those professional communities whose deep expertise lies in connecting the dots between historical observation and human needs.

Under one institutional roof, once safely preserved and open to the public, the Trust's research collections will connect lessons from the Internet's history to a range of broader technology stories that are critically important to understand in our times, including:

- continuing technology-driven transformations in employment and education
- the necessity of resilience and the social costs of infrastructural outages
- the political economy of international trade and information services
- the social impacts of regionalized technology diffusion
- the cognitive infrastructure that delivers artificial intelligence to global markets

In other words, the Internet Data Trust represents our decade's best chance to make the Internet's history freshly relevant, in response to a historical moment that seems otherwise precarious. It's a roadmap for the Internet's institutions to coordinate their visions for data preservation, coupled to the promise of creating entirely new kinds of desperately needed community sensemaking atop that digital heritage.

[1] For example: Pinger (https://www.slac.stanford.edu/pinger/) was nearly lost forever due to the end of a funded research project, with no provision for terminal preservation.

Stakeholder Landscape

There are, broadly speaking, six major categories of stakeholders whose interests must be considered, aligned, and counterbalanced in the formulation and administration of the Trust.

First, and most immediately, we must consider the **producers and users of Internet measurement data,** who are also generally the custodians of accumulated measurement datasets in need of preservation. These may comprise Internet registries (e.g., RIPE NCC), academic research institutions (CAIDA), nonprofit corporations (NSRC, M-Lab), public benefit corporations (DNS OARC), and international treaty organizations (PCH), to name just a few of the diverse forms of historical organization.

Second, we consider the community of **academic and industrial researchers,** now or in the future, who rely on accumulated and ongoing Internet measurement to advance our understanding of the Internet, its operation, and its influences on human society at large, and who make their research freely available to the public. These include researchers whose job is to understand the Internet and similar technical infrastructures, but also (critically) those social scientists and historians who will effectively use technical data to tell the stories of our century.

Third, we consider the community of **commercial and noncommercial service providers** (network operators, infrastructure services, OTT providers) who rely to a greater or lesser degree on publicly available Internet measurement research to provide their services, optimize Internet operations, or make their Internet-delivered offerings more resilient. These users rely on insights derived from public Internet measurement, though they have not always been forthcoming in their support for preserving that history.

Fourth, we consider the community of **librarians, archivists, and digital preservation specialists,** who advocate for the preservation and organization of human knowledge, bringing centuries of expertise to bear on the challenges of prioritizing the rescue of heritage data. These are the voices that can guide the Trust's roadmap, showing how to balance the competing concerns of preservation and open access, and how to make hard decisions about what history we can afford to preserve.

Fifth, we consider the space of potential **sponsors and guarantors** of the Trust's public interest mission, including those who advocate for the Internet itself (eg., the Internet Society and its chapters), and those who rely on accurate historical data for development of local, national, regional, and global policy (governments, IGOs). These institutions have the most immediate perspective on the Trust's public interest mission, and the resources to help ensure its sustainable development.

Sixth and finally, we must consider the needs of a community easily overlooked, because most of them have not been born yet: **those whose lives will depend on the world's technical infrastructures in 2125.** This is a community to whom we owe a primary duty of care, too often left out of our present-day plans because of their invisibility and silence.

Barcelona Convening: Custodial Institutions and Researchers

In December 2025, with the Internet Society's support, we convened a workshop in Barcelona to explore the Internet Data Trust concepts and hear from data holder institutions about their hopes and concerns for long-term data preservation. More than 60 researchers and representatives of data-holding institutions attended the afternoon's sessions, sharing their experiences and concerns, and collaborating on a set of guiding principles for Trust operation.

The Internet's custodial institutions have a long, rich history of maintaining their own research groups, working together to produce integrative research based on shared data, and ensuring that

data remains open and interpretable for the creation of dashboards and other interactive tools that can reach a broader policymaking audience. But successful collaborations between custodians and researchers have created a largely uncompensated mesh of dependence and obligation, making it even more important (and more expensive) for custodians to guarantee continued operational access to their historical datasets.

What we heard in Barcelona were signs of a community under significant stress, but also one prepared to establish patterns of mutual aid and pursue shared concerns. The presenters shared stories of preservation burdens that tend to increase without limit, in a world where it has become increasingly difficult to find resources (money, but also time and attention) to dedicate to preservation.

Key Takeaways

One workshop takeaway was that the costs and benefits of maintaining historical datasets tend to become unaligned over time, for organizations that are focused on the needs of a broad user community. In conversation with these data custodians, we heard that users are primarily focused on operational impacts in the moment ("What does the Internet look like today? What outages or attacks are my own users subjected to today?"). To these user communities, historical data may seem less immediately useful in the construction of alarms and real time dashboards.

As a result, if nobody speaks for the datasets that enable long-horizon longitudinal studies, there was concern that those datasets might be at risk of deprioritization or deletion. Faced with storage costs for maintaining a warm collection that few researchers will use, the temptation is strong to prune old data holdings, or send older data to tiers of storage (like AWS Glacier, or Google Cloud Archive) that make the data less available, shifting costs to future users who may never arrive as a result.

We also heard clear frustration around the institutionally siloed nature of long term data storage: the sense that if data offerings from various repositories could be "brought into the light" together, with appropriate curation and metadata and searchable interfaces to make them mutually intelligible to researchers, it might be possible to answer entirely new kinds of questions (potentially at the expense of raising new privacy concerns).

Data privacy regulations were also frequently cited as a potential motivator for failing to retain datasets, for fear of liability for the maintenance of whatever personally identifiable information (PII) may be lurking in the historical record.

We suggest that the Trust can serve to restart conversations around the preservation of digital heritage, recognized by UNESCO since 2003 as a valid goal in tension with, rather than in conflict with, citizen rights to data privacy, particularly when viewed through the lens of century-scale preservation and the needs of future generations.

The Internet Data Trust concept should respond to these recurring institutional pain points, by providing custodial institutions with a clear way out of their collective preservation dilemma: building a community of practice and mutual aid around common challenges of long-term digital preservation.

"Establish a legally constituted Internet Data Trust to protect and preserve the world's Internet measurement and infrastructure data, as part of humanity's collective heritage, to guide future development in the public interest."

Mission Statement and Key Principles

The Trust's mission statement is the core statement of purpose from which its core fiduciary responsibility flows. It should be general enough to last a century, but specific enough to guide the Trustees in everything that follows.

The principal motivating vision is to create a durable legal and technical trust framework that will lower the existential risks for the Internet's historical datasets, by making cooperative preservation among stakeholders economically and politically feasible. The trust achieves this risk reduction by placing the archives in the service of the public good — promoting access and research as well as long-term preservation.

Feasibility is achieved by creating a community of support among the participating institutions, so that the Trust's mission-directed needs are met, and all members receive proportional benefit in exchange for their support. Benefit may take many forms (financial, ease of access, sharing of burdens, recognition), but in the end, no single member should realize disproportionate benefit at the expense of others.

To find this balance, we began by articulating a set of core values, or principles, that could support the mission and give it purpose – the "why" to complement the "what." These principles guide the operation of the Trust and the behavior of both contributors and users of the entrusted data.

We presented an initial set of draft principles to the December 2025 workshop in Barcelona, where more than 60 stakeholders worked to debate, revise, and condense them, in preparation for the generation of a first nonbinding Memorandum of Understanding.

Together, such a set of governing principles attempts to create a self-consistent mesh of obligation and benefit among the participating members of the Internet data preservation community. With careful attention to the mechanisms of governance, the principles should create a mesh of trust that can provide stable support for preservation activities over a timespan longer than the lifetime of the participating institutions.

Community Feedback and Guidance

Workshop participants' key recommendations in this first round of revision included:

- Organize principles for clarity into Mission, Core Principles, and Operating Principles, providing short labels for better interpretability.

- Clearly identify preservation as the Trust's top-line responsibility; add access and the advancement of knowledge as mission-level public benefits.
- Make realistic commitments around long-term intelligibility for future researchers ("you cannot guarantee 100 years of intelligibility"). Participants with library and archival experience added emphasis on transparency, provenance, and integrity.
- Consolidate and simplify privacy / no-harm language into a single, clear principle.
- Emphasize preservation of open-source software, rather than maintenance of open-source software ("thou shalt not!")
- Add provisions for the governance of content ingestion that protect the Trust from DoS-like overload, misalignment, or future conflict with preservation commitments. Make it clear that accepting assets into the Trust is an explicit governance function ("if it goes in, it doesn't come out").
- Clearly separate values from mechanisms, reducing confusion between aspirations and obligations.
- Incorporate more explicit definitions of "public interest" and "public good," given the potential for variation across global jurisdictions.
- Address contributors' concerns about fairness, revenue impacts, and disincentivization.

Draft Trust Principles

After incorporating the workshop's recommendations, the following definitions and principles are offered as a starting point for a potential Memorandum of Understanding.

Public Interest — The long-term societal benefit derived from preserving, understanding, and accessing Internet measurement and infrastructure data, independent of the interests of any single state, corporation, or stakeholder.

Public Good — A resource whose benefits are non-rivalrous, widely shared, and preserved for future generations.

Harm — Risks to privacy, political repression, discrimination, commercial exploitation, cybersecurity vulnerabilities, or substantial disadvantage to contributors or communities.

Mission Principles

- **Preservation as a Global Public Good.** The Trust exists to safeguard the world's Internet measurement and infrastructure record for the benefit of future generations. Preservation of these materials constitutes a public good of global significance, underpinning research, accountability, transparency, innovation, and long-term understanding of the Internet's evolution. All other activities of the Trust must be consistent with this mission.
- **Stewardship in the Public Interest.** Materials contributed to the Trust are held in fiduciary care for the world's public, present and future. Contributors retain sovereignty and intellectual property, while granting the Trust irrevocable, perpetual rights to preserve and make available the contributed materials in service of this public interest. No successor entity may privatize or reclaim entrusted materials.
- **Advancement of Knowledge through Responsible Access.** The Trust promotes research, education, and societal understanding by enabling responsible access to preserved materials. Public benefit – not private exclusivity – governs the Trust's access framework.

Core Principles

- **Integrity, Provenance, and Verifiability.** Entrusted materials must be accompanied by transparent documentation of origin, methods, processing, and transformation. The Trust shall maintain verifiable provenance and ensure integrity through open auditing, versioning, cryptographic verification, and clear disclosure of known limitations or uncertainties. Trustworthiness of the holdings is a foundational asset of the institution.
- **Privacy, No-Harm, and Ethical Obligation.** The Trust must minimize risk of harm to individuals, communities, and institutions, especially where data contain human-related information or where disclosure could cause political, economic, or security risks. Privacy, confidentiality, and ethical obligations apply across jurisdictions and generations. Restrictions on access must be justified, documented, and reviewed regularly.
- **Legal Compliance and Respect for Sovereignty.** Preservation must honor applicable law, respect data sovereignty, and maintain compliance with contributor agreements. Contributing institutions retain rights consistent with their legal frameworks while enabling the Trust to preserve and provide access in ways aligned with its mission.

Operational Principles

- **Openness and Interoperability.** The Trust employs open standards, open formats, and open interfaces to maximize longevity, transparency, and cross-community research use. Where openness conflicts with legitimate obligations (privacy, security, licensing, or sovereignty) the Trust shall select the least restrictive feasible approach.
- **Governed Acceptance of Materials.** Acceptance of materials into the Trust is a deliberate and governed process. The Trust must evaluate submissions for mission relevance, legal and ethical compliance, sustainability, long-term feasibility, and potential to overload or destabilize the archive. There is no obligation to accept all offered materials.
- **Permanent Stewardship.** Once accepted, materials become part of the Trust's permanent holdings. Removal is permissible only under extraordinary, transparent, and procedurally governed circumstances. Contributors and external parties cannot demand withdrawal except where required by verified legal obligations. The permanence of accepted materials is a defining feature of the Trust.
- **Discoverability and Best-Effort Intelligibility.** The Trust commits to best-effort preservation of context: descriptive metadata, schemas, documentation, software artifacts, and other interpretative aids needed to understand preserved materials. The Trust acknowledges that long-term intelligibility cannot be guaranteed, but will pursue reasonable strategies (migration, emulation, documentation) to support future researchers.
- **Tiered Access.** Holdings may be classified as open, delayed, restricted, or confidential. Each restriction must be explicitly justified in relation to risk, law, or harm mitigation. Restrictions should be time-bounded when feasible and must undergo periodic review for possible declassification.
- **No Single Points of Failure.** Holdings must be sufficiently replicated and distributed, so that no single organization, jurisdiction, or failure mode is allowed to jeopardize the historical record.
- **Fairness, Reciprocity, and Incentives.** Access frameworks (whether open, restricted, or commercial) must not disproportionately advantage any member or contributor. Revenue

derived from commercial uses should be reinvested into preservation, infrastructure, and operations. Access models should avoid disincentivizing contributions or undermining contributors' own livelihood or mission.

- **Transparent Governance and Financial Operations.** The Trust shall maintain visible, reviewable governance processes, decision-making, and financial operations. Transparency is necessary to preserve the institution's legitimacy and protect it from capture, mission drift, or undue influence.
- **Century-Scale Continuity.** To endure across generations, the Trust must periodically re-examine its principles, update its technical and legal frameworks, and adapt to new understandings of "public interest." Such renewal shall occur through stable procedural mechanisms that maintain continuity while allowing appropriate evolution.

In the following sections, we'll examine the initial steps of organizing such a Trust according to these principles, starting with an overview of the preservation priorities, and then providing a responsive roadmap for the Trust's legal organization, governance, and sustainability.

Technical and Operational Foundations

Prioritization Strategy

In a perfect world, an Internet Data Trust would accept any set of Internet measurements ever performed and preserve them for the ages, on the off chance that some far-future researcher would find buried within them the missing pieces of an as-yet unimagined puzzle. In practice, of course, the Trust will be limited by its available resources, and decisions about what to accept in trust must be weighed carefully.

In the popular imagination, data archivists never have enough space: not enough disks, tapes, petabytes and exabytes. But in truth, preservation is more often a race against time: time to identify endangered measurement sets before they are lost, time to make contact with the custodians and negotiate accession agreements, time to piece together and preserve enough working context (metadata, software, documentation) to make the data legible to future researchers before that, too, is lost in a disaster or through neglect.

The Internet Data Trust therefore articulates a set of acquisition priorities, in part to guide preservation efforts, but also as a distress signal to the technical community to jog memories and bring more of these endangered datasets into the light before they are lost.

Initial datasets that are considered top priority for preservation are those that give us the basic map of the Internet at each moment of history, as it evolved. It's important that these have point-in-time provenance; that is, so that when the Trust accessions a snapshot, it's clear when it was collected, from which vantage point, so that it can be integrated into the global picture.

Top priority: foundational datasets

These datasets provide a map of the evolving Internet from the 1990s to present day, from the assignment of globally unique identifiers, to the evolving interconnection of autonomous systems, to the end-to-end performance across the resulting Internet, along with some indication of where all those machines were and what they were doing at the time. These are all primary sources that were created at a given moment in time, rather than inferred reductions or derivations of primary sources.

- **Number resource allocations:** RIR databases, including historical data for legacy number resources that were delegated before the introduction of an RIR membership structure, or indeed, before the creation of the RIRs, along with his-

torical delegated stats files, summarizing allocations and assignments of number resources by RIRs

- **WHOIS and RDAP** historical snapshots
- **Internet Routing Registry (IRR) datasets** that document providers' intended configurations and routing policy metadata
- **Border Gateway Protocol (BGP) routing tables, and BGP update streams,** which describe the best paths that existed to various network prefixes from the perspective of many autonomous systems around the world
- **RPKI records** that document the efforts of autonomous systems to validate the origins of their announced BGP routes over time
- **PeeringDB snapshots, IXP membership data,** and other documentation of the peering and transit relationships among autonomous systems and their presence at IXPs worldwide
- **End-to-end performance measurements** of latency, loss, and throughput across the Internet **(ping, traceroute)** from as many vantage points as possible, with special attention to technology change (eg., traces through decommissioned networks; traces through emerging mobile and satellite networks)
- **Netflow datasets,** capturing traffic volumes and protocol composition as seen on a particular provider's backbone network at a moment in time
- **DNS zone files,** showing the mapping between names and numbers, and between names and the hosting of authoritative servers, at various moments through history, with particular focus on root zone and ccTLD operations
- **DNS logs and packet captures,** collected server-side during operations or client-side as part of name resolution for other purposes (e.g., as part of web crawling, or as part of a periodic research initiative such as DNS OARC's DITL), providing insight into what kinds of services hosts were providing and which parts of the Internet were using them
- **Geofeeds,** documenting provider-level ground truth about IP geolocation (not inference; inferred geolocation sources are also important, but are of secondary importance because they can be reconstructed from foundational sources)
- **TLS/PKI certificate transparency logs and historical CA datasets** that record trust relationships over time for content
- **Blacklists and spamlists** that purport to identify domains and number resources associated with various Internet behaviors over the years

Some of these datasets contain potentially sensitive information (particularly in the WHOIS databases, which contain sensitive registrant information, and DNS packet captures and log files, which may contain user activity), and the Trust will need to provide appropriate guarantees of tiered access that accomplish the primary preservation mission, while still supporting far-future research.

Secondary priority: derived datasets, background radiation datasets

The foundational datasets are large enough, and noisy enough, that over the years researchers have created many elaborate reductions, interpretations, and filtered products from the raw data. These are important in their own right; in many cases they would be of more use in replicating published work than the original datasets from which they were derived. Along with the versioned tools that generated them, these datasets are worthy of preservation even though in theory they can be regenerated from the original data on demand. To give a few examples:

- **Unifying datasets** that make it easy for a researcher to identify, for example, which of many RIRs had relevant information for a given prefix at a given moment in history, and who might have been originating it in BGP at that same moment
- **ASN topology datasets,** which construct point-in-time graphs in which ASNs are nodes and their interconnection relationships are edges, along with metrics that assess graph properties such as centrality and connectedness, often on a national level
- **ASN to Organization Mappings** between autonomous systems and organizations (grouping number resources together after business events such as M&A–driven consolidation, to allow organization identity to be used as a key for information reduction)
- **ASN-ASN Relationships** (peering versus transit relationship inference at various moments throughout history, to aid in BGP path interpretation)
- **Geolocation inference resources** that purport to describe the locations on earth where number resources were physically deployed, using inference from DNS/RDNS, WHOIS records, latency measurements, and other techniques

The Trust also regards extremely large **network telescope datasets** as being of secondary priority, not because they are uninteresting or unendangered, but because they are extremely large by comparison to the other foundational datasets. Network telescopes record unsolicited packet-level traffic arriving to an otherwise unused region of IP address space, to understand the kind of background radiation that all Internet hosts experience (and, typically, filter and reject). It may be desirable to implement a selective sampling policy, or to bias toward preserving older telescope data, rather than attempting to preserve complete and continuous historical collections.

Tertiary priority: things other than datasets that provide context

These might constitute mailing lists, newsgroups, user communications, scraped Web resources, oral histories, or other kinds of historical materials that provide additional context to future historians to help interpret the structure and function of the Internet as revealed by the foundational datasets. This might also include out-of-band information about the physical connectivity of the Internet, such as submarine cable maps or historical information about physical facility presence by operators.

The Trust explicitly deprioritizes these sources because other efforts have already sought to preserve them, and because they require more intensive, different forms of curation than are envisioned as the Trust's core competency. Still, they form an important bridge between the public's perception of Internet history and the Trust's more data-driven approach to preservation. If the Trust can provide even limited support to rescue these important contextual materials from destruction, it should do so, given their small storage requirements relative to the larger preserved datasets.

Creating Placeholders for Rediscovery

Finally, the Trust can help organize the Internet measurement community in the search for lost or "stranded" data assets. Many important datasets are known to scholars only because early papers referenced them, perhaps in summary: for example, original packet captures from the early 1990s, regional measurement campaigns that vanished with institutional churn, or files that existed on FTP sites long gone. These datasets may still live on in a disk or tape stored in someone's garage, somewhere in the world, waiting for a spark of recognition from someone who hears about the Trust's mission.

Like an ancient manuscript that is known only

through its passing mentions in a later writer's surviving commentary, the Trust's collections will include placeholder identifiers for known-but-lost datasets, as a way of communicating to the world the need to keep searching for these lost treasures that fill the gaps in the Internet's historical record.

Preservation Collection: Century-Scale Challenges

The long-term preservation of Internet measurement and infrastructure data is not, fundamentally, a technical mystery. Storage technologies, replication strategies, fixity verification, open formats, and distributed ingest pipelines are all solvable with today's tools and tomorrow's incremental improvements. What is genuinely difficult, and what the Internet Data Trust is designed to confront, is the organizational, legal, funding, and community architecture required to make preservation reliable for a century or more.

Accordingly, it's important to avoid locking the Trust into premature technology commitments. Instead, it's useful to try to articulate a values-driven technical posture that allows the Trust to make technology choices over time that will remain faithful to the core operational principles: durability, intelligibility, openness, distributed stewardship, and service to the global public interest.

Distributed Federated Replication

Unlike specific hardware choices, distributed federated replication is a durable commitment, because it expresses an institutional (rather than technological) principle: no single organization, jurisdiction, or failure mode can be allowed to jeopardize the historical record.

The three key attributes of the global preservation collection, then, are:

- Distributed: Preservation copies live across multiple continents and sovereign institutions.
- Federated: Hosting partners maintain autonomy while adhering to common standards for metadata, fixity, and refresh cycles.
- Replicated: Each accession receives multiple independently verifiable preservation copies.

While the Trust is not prescribing specific software platforms at this stage, it should explore the value of mature, open-source, community-governed repository frameworks that already support large-scale scientific preservation.

A key example is Invenio[1], the platform developed at CERN, which demonstrates several properties that align strongly with the Trust's values-driven preservation doctrine:

- Open formats and extensibility. Invenio emphasizes open metadata standards (JSONSchema, JSON-LD, DOIs, PID systems) and modular components. This echoes the Trust's commitment to intelligibility and long-term maintainability.
- Federated deployment model. While often run centrally (e.g., by CERN), Invenio is explicitly designed for multi-institution deployments, allowing autonomous nodes to interoperate through shared protocols. This mirrors the Trust's intended model of distributed federated replication.
- Integration of data, code, and metadata. Platforms built on Invenio (e.g., Zenodo) can ingest heterogeneous artifacts (datasets, software packages, documentation, workflows) and bind them via persistent identifiers. This aligns with the Trust's requirement to preserve the entire research ecosystem, not simply the raw datasets.
- Proven relevance for large, heterogeneous scientific communities. The success of Zenodo demonstrates that Invenio-style architectures can support multi-disciplinary users, complex provenance structures, mixed access policies, and long-lived archival commitments.

- Public-interest stewardship ethos. CERN's governance model for Invenio provides a template for how open-source preservation software can be sustained collaboratively without dependency on any single commercial vendor—an essential requirement for century-scale trustworthiness.

The Trust does not commit itself to adopting Invenio or any similar platform outright. Rather, these infrastructures serve as reference implementations illustrating how the Trust's organizational principles—openness, federation, verifiability, and institutional pluralism—can be technically instantiated using today's tools.

Institutional Partners as Co-Stewards

Finally, libraries, archives, research infrastructures, and national memory institutions are envisaged not as passive mirrors but as co-equal stewards, shaping the federated replication environment. The Trust's distributed model for preservation may deliberately mirror key aspects of collaborative efforts such as Hasdai[2] and SLICES-RI[3], recognizing that long-term safety lies in institutional pluralism and jurisdictional diversity.

Research Collection: A Living, Open, Actionable Commons

While the Preservation Collection ensures that foundational Internet datasets survive intact for future generations, the Research Collection makes them useful today. It will constitute a live, continually updated scientific environment that draws from the preservation substrate to create a coherent, discoverable, richly annotated archive of Internet measurement history. Its purpose is to accelerate research, support evidence-driven public policy, and enable cross-dataset synthesis that has previously been impossible.

The Research Collection is where the Trust's long-term commitments meet the immediate needs of scholars, engineers, policymakers, and communities that rely on open Internet measurement. It sits as a "warm" operational layer above the cold preservation pool.

While the Preservation Collection ensures the survival of raw data, the Research Collection provides:

- normalized, schema-aligned extracts from heterogeneous measurement sources;
- searchable indexes, summaries, and references that can be regenerated whenever tools improve; stable access endpoints for query, analysis, and cross-referencing;
- curated snapshots for teaching, reproducibility, and method development.

In short, the Research Collection builds a regenerable working archive atop a century-scale preservation repository. As scientific practice and data storage technologies advance, the Trust simply reprocesses and regenerates the same preserved inputs to rebuild the research layer at higher fidelity.

Metadata APIs, Provenance Services, and Cross-Indexing

Central to the Research Collection is a suite of metadata and provenance APIs that allow researchers to interrogate the archive in structured ways. These services can include:

- rich per-dataset and per-record metadata, aligned with open schemas (e.g., JSON-LD, schema.org, DataCite) and resolvable via persistent identifiers;
- cross-indexing between Trust holdings and original institutional sources (e.g., RouteViews, RIPE NCC, CAIDA, M-Lab, DNS OARC);
- adherence to the FAIR principles for scientific data management and stewardship[4]
- automated DOI assignment for curated bundles, analyses, and reproducible workflows; and

- machine-verifiable provenance chains showing how each derived object was generated.

Preserving and Amplifying Community Know-How

Old Internet datasets are notoriously difficult to interpret. They contain quirks of instrumentation, shifts in operational practice, undocumented outages, and format drift. Historically, only a small group of expert researchers retained the "tribal knowledge" needed to work with them correctly. Without intervention, much of this expertise risks being lost.

The Research Collection therefore institutionalizes this knowledge by building a living corpus of know-how: a kind of "scholarly memory layer" that preserves decades of methodological insight for future generations. The Trust does not maintain open-source software, but it should accept and preserve the libraries of contributing institutions. And as researchers gravitate toward the Research Collection, their research outputs can become suitably careful in referencing the archival materials they draw on, linking to DOIs, provenance metadata, and the precise preserved datasets used.

Done correctly, the Research Collection becomes both a pedagogical resource and a scientific accelerator, enabling rapid onboarding of new researchers and validation of existing work, with full credit and feedback to the institutions that contributed the historical data.

Governance Architecture

We envision:

- a legal trust structure for preservation of contributed datasets, metadata, documentation, and other contextual materials (the **Internet Data Trust, or IDT**), defined by a charter of principles, with a set of Trustees and rules for their election and succession, domiciled in an appropriate jurisdiction with strong history of charitable trust law
- a supporting nonprofit organization (jurisdiction TBD) that serves to operationalize the Trust and use its resources according to its charter to support research and preservation activities in the public interest (the **Internet Data Initiative, or IDI**), with a board of directors and an executive management team guided by a voting, but non-shareholding, membership.

Because the IDI serves to balance the concerns among a set of existing institutions, while leaving the door open to new institutions that may arise in the next century, IDI membership is institutional, rather than individual (though under the operation of these entities, individuals should still be considered users of the data). The IDI should solicit representation from custodial institutions, research institutions, service providers (including commercial services), and Internet advocacy NGOs, and may offer observer status to representatives of states.

In order to sustain the IDT's activities, the IDI's funding model would ideally draw on a diverse range of funding sources: member contributions such as annual payments, in-kind services, data contributions, commercial use payments, service payments or grant funding. An important principle underpinning the model would be to ensure that no single member benefits disproportionately.

As data contribution agreements are negotiated with the founding organizations, the Trustees should aim for uniformity where possible, to minimize the costs of both data acquisition and downstream licensing, under the guiding principles of the Trust Charter (neutrality, transparency, service of the public interest, primacy of preservation). This is critical not only for the operation of the trust, but for ensuring the simplicity of licensing for use of the data and for building trust and understanding among the membership community and broader public.

The IDI should then establish advisory groups from

among the members (technical, legal, preservation, community) to contribute feedback on the activities and developments, provide recommendations to the trustees, and ensure that the trust and associated support organizations remain accountable for the delivery of the original mission and purpose. Furthermore, it should establish an independent Ombudsman's process by which members (and potentially the public) can raise concerns about member actions (such as inappropriate downstream use), and have those concerns explored to establish facts before the Trustees take disciplinary or legal action.

Legal Structure Considerations

Underpinning effective governance and operation of the preservation and research collections is the organizational Trust structure: a clear and robust governance regime within which data assets are preserved and accessed for future centuries.

But before settling on the charitable trust structure as the best vehicle for a mission centered on the permanent stewardship of non-substitutable public assets, it's worth considering the alternatives and understanding their benefits and limitations.

- A simple international consortium could be formed among the custodial institutions and key research institutions, using nothing more than a Memorandum of Understanding for flexibility. This approach would solve the data sharing aspect, but does not necessarily address the long term preservation needs and therefore wouldn't achieve the primary mission goals: protecting the data assets for the long term and providing a legal framework under which to guide use of the assets in the public interest for integrative research. The vulnerability lies in the ease whereby the consortium could be dissolved if the founding institutions elected to withdraw.

  Example: Europeana[5], a 19-member consortium led by the non-profit Europeana Foundation, maintaining a common European dataspace for cultural heritage. The Foundation serves only to coordinate and maintain metadata; the collections remain with the 19 member institutions.

- A nonprofit corporation (e.g., a US 501(c)(3)) would be governed by a board and chartered to serve a public-benefit purpose. But a nonprofit can be dissolved or repurposed by board vote, weakening the continuity of its guarantees of perpetual protection for the assets, which are owned outright by the corporation.

  Example: the Filecoin Foundation for the Decentralized Web[6] is a US-based 501(c)(3) created by the Filecoin Foundation[7] to study the specific mission of promoting a decentralized Web. It has distinct governance and legal status, and preserves cultural and scientific datasets on the Filecoin network.

- A Dutch foundation / Stichting governed by a board of directors would provide flexibility, offer strong autonomy in managing endowed assets, and be able to make grants and operate projects in ways that would be compatible with EU funding and tax structures, potentially with lighter regulatory oversight than a common law trust. But because fiduciary duties are defined by an amendable charter, not by centuries of trust law, enforcement of purpose would be subject to drift over generations, even though the initial charter might anticipate perpetual management.

  Example: Stichting Internet Archive[8] is the Dutch non-profit research library whose primary project, Internet Archive Europe, seeks to preserve European digital materials to safeguard collective history.

In short, not all trust-like structures offer the same protections as a bona fide charitable trust, and the differences may be significant. For example, the OpenCorporates Trust[9] emerged in a 2018 reorganization as the holder of the OpenCorporate

database of company data. While not organized as a trust, this division nonetheless sought to clarify the long-term independence of the dataset as an open resource maintained in the public interest, with a distinct "board of trustees" drawn from their community. Similarly, the IETF Trust[10] is organized as a Delaware nonprofit corporation. Whether the distinction will be material over the course of a century is, of course, unknown.

By contrast, a charitable trust (for example, under the law of England and Wales) is not a corporation but a fiduciary relationship, in which trustees hold assets solely for the advancement of specifically defined charitable purposes — here, the preservation of Internet measurement data and its use in education and research. Article 13b of the Charities Act, in particular, highlights the Charitable purpose of Education (including research organizations).

The trust legal structure is simple, durable, and regulated (in the UK) by the Charity Commission, which legally enforces trustee duties of loyalty, prudence, and transparency — and critically, these duties survive changes in membership and leadership. Best of all, because the trust has no shareholders or owners, its entrusted assets can only be used to further its stated purpose — forever.

In order to ensure that the Trust and its Trustees steward the data according to its core mission and purpose, a companion non-profit organization can be considered to manage user community access and operational support. This important step can help ensure that no undue influence is brought to bear by members, users and external organizations.

For shorthand, in what follows we will distinguish the Internet Data Trust (IDT) from an Internet Data Initiative (IDI), with the latter identifying a to-be-named not-for-profit entity (either in the same country as the IDT or elsewhere in the world) that ultimately provides operational services to the IDT and its trustees on behalf of a membership community drawn from among the full diversity of stakeholders.

The IDT would draw on a true charitable trust structure to provide the best odds of true longevity for the preservation of entrusted assets, and their transmission to future generations. Some examples of successful organizations making use of analogous structures have included:

- **The PLACE Trust**[11] is an endowment data trust in the UK that holds licensed copies of government data in perpetuity for the public interest. It's paired with three nonprofit entities: ThisisPLACE UK (a company limited by guarantee that handles basic operations in the UK), ThisisPLACE Trading Ltd, a wholly-owned trading company that only handles commercial licensing, and ThisisPLACE Foundation, a US 501(c)3 that manages the membership, data collection, processing and storage. They all work together with a shared mission via a Framework Agreement.
- **The UK Biobank**[12] was established in 2007 to hold de-identified biomedical data for 500,000 UK study participants in perpetuity. A charitable company limited by guarantee serves as the operational partner; the charitable company's Board members also serve as charity trustees.

"The trustees might elect to set aside support for Preservation Support Grants, which help member institutions defray the cost of data transfer, metadata curation, or staff training costs."

Financial Model; Sustainability

We think of the financial model of the Trust (and its operating partner company, the Initiative) as being organized around a particular design goal: ensuring that the Trust's core preservation obligations are insulated from year-to-year funding volatility, on a path that can sustain century-scale operation.

Launch capital requirements are derived from founding contributions and early-stage income. It's important to differentiate between data-rich and capital-rich founding member institutions; both contribute to the launch of the Trust, through different channels. Founding member contributions may be in kind (datasets, storage, staff time) or financial (grant support), and both kinds of commitments are equally valuable in the establishment and operations of the Trust.

We start by considering the kinds of capital endowment resources that might exist to generate predictable yield or to provide buffers against shocks, in order to promote very long-timescale preservation. Maintaining these resources, we then consider the sources of operational income (annual budgets, grants, membership fees, service revenues including commercial use) that create sustainability by refreshing the endowment.

The goal at launch would be to secure committed "catalyst funding" sufficient to support a 5-year runway, enough to acquire core datasets and establish minimal viable operations, including a technical pipeline for dataset acquisition and preservation, community convening, first-phase research collection programs that will help the Trust tell its public interest story, and the development of products or services (not competitive with members) that help to generate sustaining revenue.

- **A core preservation endowment** would be managed under the Trust's investment policy, with income used to pay for accession costs, data storage, and global replication of preservation media in the first 5 years: in other words, only the core costs of the associated Initiative (where 'core' costs are those administrative overheads and operational expenses that ensure the execution of the Trust's primary preservation mission, as initially scoped).
- **Strategic reserve funds** would set aside additional funds for purposes that can reasonably be anticipated over a 5 to 10 year horizon: e.g., infrastructure refresh, tech upgrades, and legal contingencies, with drawdowns approved by the Trustees.
- A set of **special projects funds** would solicit and segregate foundation or agency grants intended to support specific data rescues beyond

the initial core, with a distinct reporting track to support compliance with grant terms.

This 5 year startup phase allows the Trust to find its way in a changing funding landscape, at the end of which we might anticipate having stabilized one or more of the following kinds of sustaining income streams:

- **Short-term grants** (or low-interest loans) that supplement endowment proceeds in support of Initiative operations
- **Membership fees** paid annually by founding member institutions (or accepted in kind through the donation of services and labor)
- **Service fees** paid by commercial users and other non-members; for example, non-exclusive, fee-based access to curated historical data in support of commercial analytics, with strict reciprocity (e.g., funds reinvested into the preservation endowment).

All cost recovery and licensing policies would be regularly reviewed for consistency with the Trust's principles; in particular, with its fiduciary responsibility to administer Trust assets in the public interest for long-term preservation.

Funding streams need not only be inbound; they may also be outbound, as part of the community's emphasis on mutual aid for data preservation efforts. For example, the trustees might elect to set aside support for Preservation Support Grants, which help member institutions defray the cost of data transfer, metadata curation, or staff training costs, or establish a Rapid Response Fund for acquiring seriously at-risk datasets (analogous to cultural heritage "rescue" grants).

To administer these goals, Trustees would adopt (and direct the operating company to administer) an Investment and Spending Policy to protect and grow the preservation endowment, targeting an inflation-protected 4% to 5% drawdown.

Beyond the initial 5 years, the Trustees would explore sustainability pathways that might include a larger, stabilized endowment; an expanded membership base that includes academic, NREN, policymaking, and corporate members; and periodic capital campaigns for specific infrastructure refreshes.

The key insight is that to last a century, preservation continuity must not depend solely on short-term grants and annual fundraising. It's critically important to establish buffered pools of capital that can be refreshed by diversified income flows.

An estimate for startup costs and the necessary long-term capital requirements can be found in appendix A.

Roadmap to Launch

The following roadmap outlines the primary milestones en route to a public launch of the Internet Data Trust (IDT). It spans approximately 12–18 months and culminates in the deployment of minimal viable infrastructure for both the Preservation Collection and the Research Collection. Each phase concludes with concrete institutional, legal, or operational deliverables that enable progression to the next stage.

Phase 1: Convening and Institutional Design

Outcome: a shared institutional vision and a committed founding coalition.

- Draft and refine foundational principles through a convening of primary contributing institutions (Barcelona workshop) [Complete]
- Develop a nonbinding Internet Data Trust Memorandum of Understanding (MoU), articulating core principles, definitions, and scope
- Circulate the MoU and secure expressions of commitment from founding participants and aligned institutions (e.g., ISOC, research libraries)
- Conduct targeted governance workshops and

outreach to underrepresented or missing stakeholder communities

- Converge on a draft IDT mission statement, charter outline, and governing principles

Phase 2: Foundational Commitments

Outcome: legal standing and financial viability established.

- Agree on a preferred governance framework and legal structure for the Trust
- Establish a multi-year, sustainable financial model, including endowment strategy and operating assumptions
- Complete an initial inventory of preservation-relevant datasets available from contributing partners
- Legally constitute the Internet Data Trust as an independent fiduciary entity
- Incorporate the Internet Data Initiative as the Trust's operational and research-facing counterpart
- Establish an initial preservation endowment to support long-term stewardship obligations

Phase 3: Technical and Operational Readiness

Outcome: core operational capacity demonstrated.

- Recruit and onboard a core operational team
- Develop initial operational practice documents, including accessioning, risk assessment, and governance procedures
- Design and implement ingestion pipelines and metadata standards for preserved and research-accessible datasets
- Conduct capacity-building activities, including small grants and pilot engagements with contributing institutions
- Execute a first demonstration ingest and accession workflow
- Operationally demonstrate minimal viable, tape-based preservation infrastructure

Phase 4: Operational Launch

Outcome: the Trust enters public operation.

- Finalize and adopt access, use, and disclosure policies for preserved and research-accessible data
- Complete first production dataset accessions into the Preservation Collection
- Release an initial public Research Collection, including curated datasets with persistent identifiers and documentation suitable for researcher use
- Conduct a public launch of the Internet Data Trust and Internet Data Initiative, accompanied by the release of founding documents

Notes

1. https://inveniosoftware.org/
2. https://www.hasdai.org/
3. https://www.slices-ri.eu/
4. https://www.go-fair.org/fair-principles/
5. https://pro.europeana.eu/about-us/foundation
6. https://ffdweb.org/
7. https://fil.org/
8. https://www.stichtinginternetarchive.nl/
9. https://opencorporates.com/companies/gb/11329152
10. https://trustee.ietf.org/
11. https://thisisplace.org
12. https://www.ukbiobank.ac.uk/

Appendix A. Budget Estimate

This Appendix provides an indicative estimate of the resources required to establish and operate the Internet Data Trust (IDT) and its operational partner (the Internet Data Initiative, IDI) during the first five years. Numbers are deliberately conservative: they reflect a minimal viable organization capable of accessioning foundational datasets, launching the Preservation Collection and Research Collection, establishing legal and governance structures, executing distributed replication, and supporting a small global member community.

Figures are expressed in USD and are intended as order-of-magnitude planning benchmarks. Actual budgets will vary by jurisdiction, staffing model, hosting arrangements, and in-kind contributions from founding members.

To summarize the key budget drivers:

- The Preservation Collection is cheap in terms of storage, but expensive in terms of organization and curatorial labor.
- The Research Collection is a huge public-benefit multiplier, but it requires expensive ongoing engineering effort.
- Distributed federated replication is central to the century-scale mission but introduces its own costs of coordination and maintenance.
- Building an endowment is probably mandatory for the Trust's century-scale survival.
- The first five years are the window of heaviest investment, to support institution building and get standardized patterns in motion.

Year 1 (Launch Year)

A plausible Year 1 operating budget for a lean but globally credible institution is $4.75M.

- Personnel ($2.5M fully loaded) include 9 FTE, including an Executive Director, a Technical Director (preservation), a Senior Archivist / Preservation Lead, 2 distributed systems engineers, a Data Curator to establish the research collection, a Program Manager to oversee member relations and governance support, and fractional coverage for Legal/Policy Counsel and Admin/Finance/Ops.
- Technical Infrastructure ($1.5M) sufficient to establish 5-10PB of an initial replicated footprint, with some storage donated by partners, across a cold archive plus warm ingest cluster. Bandwidth and hosting fees, tape refresh cycles, and some limited cloud archiving capacity. Initial ingest tooling and Invenio-style prototyping are accelerated by hiring contractors, with the potential for some in-kind contributions from members.
- Legal Formation, Governance, Trust Establishment ($500k) includes funding for UK trust law counsel and cross-jurisdictional review, the parallel establishment of the IDT and the IDI (either on its own, or within an existing nonprofit partner willing to serve in the operational role), and the execution of a first set of multi-institution accession agreements.
- Convening, Community, and Communications ($250k) would support travel, workshops/meetings,

and public interest positioning.

First Five Years: Scaling Up

With a gradual ramp-up in accession volume, the expansion of the Research Collection, and the deployment of three federated replication partners by Year 3 (plausibly, research universities on different continents), annual costs would increase to $6.0M by year 5, for an estimated 5-year startup operating cost of approximately $25M.

Sustaining for a Century

Beyond the costs of formation, a long-term financial model might assume three complementary components:

- Catalyst Funding (Years 0-–5). Multiyear commitments that underwrite launch activities, early hires, initial replication infrastructure, legal formation, and ingestion pipelines.
- Preservation Endowment. A capital pool whose annual 4-5% drawdown covers the Trust's long-term core preservation obligations (storage, replication, fixity validation, accessioning support).
- Operational Income (Recurring). Annual membership fees, grants, and optional commercial access fees (under reciprocity and public-interest constraints) to support the IDI's operating expenses.

Ideally, to support century-scale operations, the Trust would build an endowment of $35–40M, as follows:

Fund	Year-5 Target	Purpose
Core Preservation Endowment	$25M–30M	4–5% annual draw (~$1.0M–1.5M) dedicated to preservation obligations.
Strategic Reserve	$5M	Major refresh cycles, legal contingencies, very large dataset migrations.
Special Projects / Data Rescue Fund	$3M–5M	Opportunistic rescue of endangered datasets, emergency ingest support.

These long-term capital pools would be complemented by diverse kinds of program income, reaching a steady state by year 5:

Source	Annual Amount	Notes
Membership fees (institutional, tiered)	$1.5M	Mix of cash + in-kind services.
Grants (foundations, agencies)	$1.0–1.5M	Research-layer improvements, pedagogy, endangered-data rescue.
Commercial reciprocity fees	$0.5–1.0M	Strictly compliant with access principles; reinvested in endowment.
Endowment draw (4–5%)	$1.0–1.5M	Covers core preservation operations.

About This Report

This report was prepared with the support of the Internet Society and informed by a community workshop held during ISOC Pulse Measurement Week in Barcelona in December 2025. The convening brought together researchers, network operators, archivists, and institutional stakeholders to share experiences and concerns related to the long-term preservation of Internet measurement data.

The report was authored by Jim Cowie (cowie@internethistoryinitiative.org). The author thanks Denise McKenzie (PLACE Trust), Amreesh Phokeer, Mat Ford, and Michuki Mwangi (Internet Society), and all of the Barcelona workshop participants for their feedback, encouragement, and constructive critique. Any remaining errors or omissions are the responsibility of the author alone.