UCSF-JHU Opioid Industry Documents Archive
Digital Preservation Plan

Introduction

The purpose of the UCSF-JHU Opioid Industry Documents Archive (OIDA) Digital Preservation Plan is to: 1) Define the scope of OIDA's digital preservation activities, including the amount of, and types, of digital content; and 2) Describe specific strategies and processes that are currently in place to ensure the long-term preservation of its digital content.

Mission

The Opioid Industry Documents Archive collects, organizes, preserves, and provides free online access to millions of previously internal documents made public through legal settlements to enable multiple audiences to explore and investigate information which shines a light on the opioid crisis.

OIDA is a collaborative undertaking between the University of California, San Francisco (UCSF) and Johns Hopkins University (JHU). UCSF's Industry Documents Library (IDL) hosts OIDA within its technical infrastructure and makes the OIDA collections freely cross-searchable with other major industry documents collections hosted by the IDL, including the Truth Tobacco Industry Documents Library.[1]

Standards

OIDA aims to achieve archival good practices and will continue to apply up-to-date preservation standards related to the maintenance and storage of digital content and metadata as determined by international, national, consortial, and industry governing bodies. To that end, this plan is heavily informed by the National Digital Stewardship Alliance (NDSA) levels of preservation, the Open Archival Information System (OAIS) Reference Model, and the Digital Preservation Coalition's Rapid Assessment Model (RAM).

Further, OIDA will continue to advance its digital preservation goals and objectives by developing and maintaining the necessary hardware, software, expertise, and protocols to ensure long-term access.

[1] The Industry Documents Library (IDL) is a digital archive of documents created by industries which influence public health, hosted by the University of California, San Francisco (UCSF) Library. Originally established in 2002 to house the millions of documents publicly disclosed in litigation against the tobacco industry in the 1990s, the IDL has expanded to include documents from the opioid, drug, chemical, food, and fossil fuel industries to preserve open access to this information and to support research on the commercial determinants of public health.

Content

As of September 2024, OIDA includes 18 collections totaling more than 3.5 million publicly available documents (more than 16 million pages). This number is expected to more than double by 2025-2026.

The documents in OIDA collections are primarily internal corporate records publicly disclosed from ongoing opioid litigation brought by local and state governments and tribal communities against opioid manufacturers, wholesalers, distributors, and pharmacies. The documents reveal the many ways opioid litigation defendants sought to increase sales of drugs they knew to be addictive and deadly.

OIDA collections also include plaintiff and defendant trial exhibits and transcripts, as well as deposition testimony submitted during the course of opioid litigation, primarily from the federal multi-district National Prescription Opiate Litigation (MDL 2804). OIDA also collects public documents obtained by law firms, individual states, journalists, researchers, and other individuals through legal proceedings, court records, or Freedom of Information Act (FOIA) requests.

The files are received, stored, and managed by IDL personnel within the IDL storage environment. Original file formats include textual documents (.doc, .docx, .pdf), data files (.xslx, .csv, .json, .xml), images (.jpg), email (.pst, .eml), audiovisual materials (.mp3), and code (.html). Copies of many of these original files are converted to Portable Document Format (PDF) to aid online viewing and full-text searching.

The OAIS Model

The Reference Model for an Open Archival Information System (OAIS) provides useful terminology to distinguish data at different stages of the digital preservation lifecycle. The OAIS model was approved as ISO International Standard 14721 in 2002 and has been adopted by digital preservationists worldwide.

Submission Information Package (SIP)
In the OAIS model, the Submission Information Package (SIP) is the set of files originally submitted or transferred to a repository. The SIP contains “raw” data which generally must be processed for public access and long-term preservation.

At minimum, any SIP received by OIDA must be screened to check for sensitive information such as Personally Identifying Information (PII) and Protected Health Information (PHI).

Archival Information Package (AIP)
In the OAIS model, the Archival Information Package (AIP) is the set of files that has all the qualities needed for permanent or indefinite long-term preservation of its content. As defined by a Digital Preservation Coalition Technology Watch Report, “the AIP consists

of the information that is the focus of preservation, accompanied by a complete set of metadata sufficient to support the OAIS's preservation and access services."[2]

OIDA may maintain two versions of the AIP: 1) an unredacted copy, containing all original files (which may still contain PII and PHI) and 2) a redacted copy, containing a mixture of original files which do not contain PII and PHI, along with copies of original files which have been processed to redact PII and PHI.

In most cases, OIDA receives documents which have been redacted by the defendant under the terms of legal settlement agreements. OIDA is not in a position to review or remove these original redactions. Where appropriate, OIDA will retain unredacted AIPs for at least 5 years to allow any additional OIDA-applied redactions to be reviewed and, if appropriate, to be rolled back. This enables OIDA to investigate and address any over-redactions which may result from OIDA's large-scale automation of redaction processes. At the end of the retention period, OIDA Project Leadership will consult with the OIDA Redaction and Preservation Workgroups, key stakeholders, and university counsel to determine if the unredacted AIP can be deaccessioned. OIDA does not plan to preserve unredacted AIPs indefinitely, due to data security risks and the cost of secure storage.

Dissemination Information Package (DIP)
In an OAIS, the Dissemination Information Package (DIP) is the version of the content which is delivered to the public.

In OIDA's case, the DIP contains the data accessible on the IDL's public servers, including: processed copies of OIDA content files, which may also have been converted to PDF and/or redacted; metadata, including that originally supplied by a company and any added or enhanced metadata supplied by OIDA; Optical Character Recognition (OCR) text, which is added to PDFs to enable full-text searching within documents; and thumbnail images used for document display and online viewing.

Current Preservation Activities and Processes

This section provides a description of the current state of OIDA's digital preservation activities mapped to the categories evaluated by the Digital Preservation Coalition's Rapid Assessment Model (DPC RAM). Assessment is on a scale of 0 to 4, where Level 0 indicates minimal awareness and Level 4 represents optimized activity.

Organizational Capacities

A. Organizational viability

This refers to the governance, organizational structure, staffing, and resourcing of digital preservation activities.

[2] Lavoie, B. (2014). The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). DPC Technology Watch Report 14-02 October 2014. The Digital Preservation Coalition. https://www.dpconline.org/docs/technology-watch-reports/1359-dpctw14-02/file

After completing the DPC RAM, OIDA's assessed level of organizational viability for digital preservation is Level 1 (Awareness).

OIDA currently employs several staff with knowledge and experience of digital preservation best practices and has created a Preservation Workgroup within the OIDA organizational structure. OIDA also has access to in-house experts at its host institutions (UCSF and JHU), including digital archivists and IT professionals. In addition, OIDA has set aside funds to specifically dedicate to long-term preservation.

B. Policy and strategy

This refers to policies, strategies, and procedures which govern the operation and management of a digital archive.

After completing the DPC RAM, OIDA's assessed level of policy and strategy is Level 1 (Awareness).

OIDA has conducted a Digital Preservation Assessment, prepared a Digital Preservation Plan (this document) and established a Disaster Recovery Plan. The Preservation Plan and Disaster Recovery Plan will be reviewed and updated at least every two years.

Relatedly, OIDA also has an awareness of the environmental impacts of the energy- and carbon-intensive processes involved in storing, maintaining, and providing access to multiple terabytes of data. Two factors which OIDA has considered specifically to date include energy consumption by checksum type (OIDA uses MD5 checksums, one of the least energy-consumptive) and cloud storage provider (OIDA uses Amazon Web Services, which claims to have a 72% reduction of carbon emissions from their data centers when compared to other enterprise data centers according to some digital library experts[3]). Additional information will be gathered to continually inform recommendations and revisions to current practices to minimize OIDA's environmental impact as much as possible.

C. Legal basis

This refers to the management of legal rights and responsibilities, compliance with relevant regulation, and adherence to ethical codes related to acquiring, preserving, and providing access to digital content.

After completing the DPC RAM, OIDA's assessed level of legal basis is Level 3 (Managed).

[3] Kinnaman, A., and Munshower, A. (2022). "Green Goes with Anything: Decreasing Environmental Impact of Digital Libraries at Virginia Tech." In *Proceedings of the 18th International Conference on Digital Preservation*. Retrieved from https://www.dpconline.org/docs/miscellaneous/events/2022-events/2791-ipres-2022-proceedings/file.

OIDA has executed legal agreements (Memoranda of Understanding and/or Mutual Letters of Understanding) outlining the acquisition, processing, and preservation of major collections and receives regular guidance from university counsel at UCSF and JHU. OIDA has implemented workflows and procedures for managing sensitive data, including the provision of secure storage, a set of Redaction Protocols which governs OIDA's compliance with privacy law, and a Privacy and Takedown Policy. The OIDA website meets basic legal requirements for accessibility and further improvements are in active development.

D. IT capability
This refers to information technology capabilities for supporting digital preservation activities.

After completing the DPC RAM, OIDA's assessed level of IT capability is Level 3 (Managed).

OIDA is supported by a dedicated team of software developers. IT systems relating to OIDA are actively monitored and are regularly patched and updated. New tools and systems are deployed when required and there is comprehensive documentation of the entire application infrastructure. Contracts and services with third-party suppliers (e.g., Amazon Web Services) are well-managed and documented.

E. Continuous improvement
This refers to processes for the assessment of current digital preservation capabilities, the definition of goals, and the monitoring of progress.

After completing the DPC RAM, OIDA's assessed level of continuous improvement is Level 1 (Awareness).

OIDA has completed an initial digital preservation assessment, and has defined goals and how progress will be monitored by drafting this Preservation Plan.

F. Community
This refers to engagement with and contribution to the wider digital preservation community.

After completing the DPC RAM, OIDA's assessed level of community is Level 2 (Basic).

OIDA staff in relevant roles participate in professional networks and have some digital preservation contacts and peer support (e.g., the University of California (UC) Born-Digital Common Knowledge Group, the UC Digital Preservation Working Group, and the Digital Preservation Coalition). OIDA staff also attend community events, trainings, and conferences (including the Digital Library

Forum and International Preservation (iPRES) meetings) and are committed to building knowledge across the OIDA team, through activities such as World Digital Preservation Day.

Service Capacities

G. Acquisition, transfer, and ingest

This refers to processes to acquire or transfer content and ingest it into a digital archive.

After completing the DPC RAM, OIDA's assessed level of acquisition, transfer, and ingest is Level 3 (Managed).

OIDA maintains relationships with contributing individuals and organizations throughout the acquisition and transfer process and archival appraisal is a standard part of the workflow towards ingest. Transfer agreements (often MOU/MLOUs) are executed for the majority of collections when there is a known contributor. Other documentation is created if collections are obtained directly by OIDA from existing public sources where there is no known contributor.

OIDA can accept transfer of digital content via the shipment of physical hard drives or via electronic transfer (FTP or download). Ingest of files into OIDA storage (utilizing the IDL's servers and databases) is largely automated.

H. Bitstream preservation

This refers to processes to ensure the storage and integrity of content to be preserved.

After completing the DPC RAM, OIDA's assessed level of bitstream preservation is Level 2 (Basic).

OIDA utilizes dedicated cloud storage through Amazon Web Services (AWS) Simple Storage Service (S3) which also offers redundancy services. Maintaining multiple (redundant) copies of data in diverse storage locations reduces the risks of data loss or damage due to threats such as a natural disaster or human error. Preserving multiple copies also provides opportunities for repairing, or replacing lost or damaged data.

Files are also checked for viruses and malware during processing. OIDA activities at UCSF are managed on machines using Symantec Endpoint Protection which is designed to (1) detect, remove and prevent the spread of viruses, spyware and other security risks and (2) provide Windows, Mac and Linux computers with anti-virus (AV) and anti-spyware protection. In addition, the majority of OIDA documents are currently processed using the Everlaw e-

discovery platform which can detect and flag files that might contain malware when processing native data.

OIDA has implemented a basic process for bitstream preservation which includes backup regimes and the calculation of MD5 checksums. Checksums serve as a unique alphanumeric identifier for a specific file in a specific state. If anything about the file's condition changes (e.g., if the file is edited, or any underlying bits change) then the checksum will be different the next time it is calculated. Checksums help to validate data integrity and provide information to detect errors introduced in the course of data transmission or storage. Checksums can provide verification that no changes in the files have occurred and that the integrity of digital content remains intact.

I. Content preservation

This refers to processes to preserve the meaning or functionality of the digital content and ensure its continued accessibility and usability over time.

After completing the DPC RAM, OIDA's assessed level of content preservation is Level 2 (Basic).

Upon ingest to OIDA, staff perform checks to validate files. The content files have often undergone some level of conversion prior to ingest since many have been previously housed in an e-discovery system such as Everlaw or Relativity. Specific procedures are followed to ensure that all incoming file formats are identified and that any quality issues, such as incomplete or missing content, or encrypted, broken, or invalid files are identified.

OIDA staff continually work to understand the needs of current and future users, and decisions regarding content and preservation actions are informed by this awareness.

J. Metadata management

This refers to processes to create and maintain sufficient metadata to support preservation, discovery, and use of preserved digital content.

After completing the DPC RAM, OIDA's assessed level of metadata management is Level 3 (Managed).

OIDA has implemented a minimum requirement for descriptive metadata to ensure that basic fields are always complete. It applies Dublin Core metadata standards where possible and appropriate, and when necessary, creates additional preservation and administrative metadata for objects to support other archival activities. Preservation metadata encompasses additional descriptive terms to describe digital objects by capturing information which enables preservation actions. It often includes rights management information, and

technical information (such as original file format) that maintain the ability to use the full value of the files.

OIDA also assigns Archival Resource Keys (ARKs) for each document. ARKs are very similar to Digital Object Identifiers (DOIs) and serve as persistent identifiers that enable long-lasting online access to digital assets, even if a file's location changes. For example, if the URL of an OIDA document changes, the ARK would allow a user to find that document's new location. The ARKs created for OIDA documents are registered with the ARK Alliance, an open global community supporting the ARK infrastructure on behalf of research and scholarship. The ARK Alliance maintains a link resolver which translates an ARK identifier into its current URL location.

K. Discovery and access
This refers to processes to enable discovery of digital content and provide access to users.

After completing the DPC RAM, OIDA's assessed level of discovery and access is Level 3 (Managed).

OIDA provides resource discovery for all digital content through its public website and full text search is available for all digital content. Information about copyright and re-use of the documents is available. OIDA has created subject guides for each collection to help users navigate the documents. Basic reports can be generated by staff about user access to digital content via web analytics, and the access system (web user interface) is currently being updated to reflect feedback and questions from users. The OIDA website meets basic accessibility requirements, and a Disaster Recovery Plan for OIDA content is currently in place under the Industry Documents Library's existing plan. In the event of a disaster or website crash, public access to content can generally be restored within a few hours, or in a worst-case scenario, within three days.

Planned Improvements

Capability	**Current Level**	**Target Level**	**Goals**
A. Organizational viability	1	3	1. Identify staff with digital preservation responsibilities and update job descriptions with at least 5% effort dedicated to digital preservation 2. Ensure staff working on digital preservation activities have time and funding to attend at least 1 related training event per year

Capability	Current Level	Target Level	Goals
B. Policy and strategy	1	3	1. Create a basic digital preservation policy, with focus on environmentally sustainable digital preservation 2. Ensure there is current and accurate documentation of procedures for managing and providing access to digital content
C. Legal basis	3	4	1. Implement a storage plan and retention schedule for unredacted AIPs 2. Document OIDA processes for managing formerly-privileged documents 3. Seek further expert feedback on Redaction Protocols with the aim of sharing them publicly and with the larger digital preservation community
D. IT capability	3	4	1. Explicitly define activities performed by software developers that support digital preservation 2. Provide additional training to developers on archival digital preservation topics 3. Include digital preservation as a factor in future technical decision-making
E. Continuous improvement	1	3	1. Identify gaps in digital preservation activities 2. Determine how OIDA's overall digital preservation capability compares to NDSA Levels of Preservation, and to peer institutions
F. Community	2	3	1. Further develop peer contacts and networks for digital preservation advice 2. Facilitate OIDA staff attendance at digital preservation training/webinars 3. Continue raising awareness of digital preservation through events such as World Digital Preservation Day 4. Ensure OIDA is involved in at least one digital preservation event or conference each year (through attendance, presentations, or publishing in relevant blogs/journals)

Capability	Current Level	Target Level	Goals
G. Acquisition, transfer, and ingest	3	3	1. No activities planned at this time
H. Bitstream preservation	2	3	1. Establish formal workflow to guide frequency of file integrity checks and outline how repairs to files would be made
I. Content preservation	2	4	1. Document how changes to digital content and metadata are recorded (when, what, how, why and who)
J. Metadata management	3	4	1. Create documentation outlining what metadata fields are used to record preservation information 2. Create documentation of an “exit strategy” detailing how documents and their metadata would be extracted in standardized content packages (AIPs and DIPs) and moved to a new data management system if that is required due to failure or obsolescence of the current system
K. Discovery and access	3	4	1. Monitor web analytics to evaluate use of current website 2. Implement updates to user interface 3. Review disaster recovery plan at least every two years

Progress Monitoring

The OIDA Preservation Workgroup will track these goals in its Preservation Project Activities tracker spreadsheet. This will serve as a “road map” for the next 1-2 years. The Workgroup will categorize the goals by priority and set target dates for the goals to be achieved. Workgroup members will be assigned to specific activities and take responsibility for completing tasks. The Workgroup will meet at least monthly to review progress, adjust timelines and dates as needed, identify any problems which may delay or prevent work, and make recommendations for how new workflows and policies will be implemented.

Last Revised

October 28, 2024