TECHNICAL REFERENCE REPORT

Metadata Collection and Scraping

Files, media, documents, public websites and first-party visitor telemetry

PURPOSE

A scenario-by-scenario catalog of what can be observed, extracted, derived, inferred and responsibly retained - with implementation, reliability, privacy and parser-security guidance.

Prepared 31 August 2026

SHAREABLE EDITION

REPORT ORIENTATION

Scope, terminology and reading model

This guide catalogs the information available from locally stored files, transferred files, public web resources and visitors to a server you operate.

Core rule: there is no universal metadata maximum. The technical ceiling is the representation and bytes actually delivered to an authorized client, plus information that can be derived or inferred from those bytes. The responsible ceiling is narrower: collect only what serves a defined purpose, label its reliability, secure it and expire it.

Guide map

1. Foundations

Hashing versus encoding; evidence classes; embedded, filesystem, sidecar and service metadata.

2. Stored files

What Windows/NTFS and other filesystems store around a file, and what survives copying.

3. Network transfer

Protocol envelopes, HTTP fields, server records and the boundary between file data and transport data.

4. Format catalog

Images, video, audio, PDFs, Office files, archives, executables, source code and databases.

5. Public web

DNS, TLS, HTTP, raw HTML, rendered DOM, structured data, public APIs, infrastructure and people.

6. Site visitors

Passive server observations, JavaScript telemetry, permission-gated APIs, user-provided data and limits.

7. Application blueprint

Acquisition, parsing, provenance, storage model, privacy controls, security and inspection commands.

8. References

Authoritative standards, tools and privacy guidance supporting the catalog.

Important terminology

Term	Meaning	Example
Encoding	A reversible representation of data.	Raw pixel bytes displayed as hexadecimal.
Cryptographic hash	A fixed-length, one-way digest calculated from bytes.	SHA-256 of the complete JPEG file.
Perceptual fingerprint	A compact similarity signature derived from normalized visual or audio features.	pHash-like image similarity value.
Metadata	Information describing data, its context, structure, handling or provenance.	EXIF camera model or NTFS modification time.
Telemetry	Measurements produced while a system or user interacts with a service.	Request latency, page view or JavaScript error.
Inference	A conclusion estimated from observations; it may be wrong.	Approximate country from IP or probable duplicate identity.

Hashing correction

A raw pixel map can be represented in hexadecimal, but that is not what a standard file hash does. SHA-256 normally processes every byte in the encoded file - container headers, compressed pixels, metadata, profiles and trailing data. It produces a short digest from which the original cannot be reconstructed. A perceptual hash instead decodes and normalizes media before extracting features. Consequently, editing only EXIF metadata changes an exact file hash while leaving the visible image unchanged; a perceptual fingerprint may remain similar.

EVIDENCE MODEL

Where metadata lives and how much to trust it

Metadata is distributed across several layers. A useful scraper preserves these layers instead of flattening every value into one undifferentiated record.

Layer	Typical fields	Location	Travels with byte-for-byte file?	Reliability
Filesystem	Name, file ID, timestamps, length, owner, ACL, attributes, allocation.	NTFS, ext4, APFS or another filesystem.	Usually no.	Observed locally, but copy operations and clocks affect values.
Embedded	EXIF, XMP, ID3, document author, codec, PDF information dictionary.	Inside the file bytes.	Yes, unless transformed or stripped.	Publisher/device asserted; editable.
Sidecar	XMP edits, captions, project settings, checksums.	Separate companion file or local catalog.	Only if companion data is transferred.	Application asserted.
Extended filesystem	NTFS alternate streams, extended attributes, resource forks.	Attached to filesystem record outside the main stream.	Protocol/filesystem dependent.	Observed locally.
Transport	IP, ports, HTTP headers, TLS parameters, redirects, timing.	Packets, requests, proxies and logs.	No; recreated during transfer.	Observed per transaction; fields can be missing or spoofed.
Service database	Uploader account, source IP, upload time, moderation result, object ID.	Platform database or logs.	No.	Service observed or inferred.
Derived	Hash, dimensions, duration, OCR text, detected MIME.	Calculated by your system.	Recalculable.	Deterministic or model dependent.

Evidence classifications for your database

Class	Definition	Example	Recommended handling
Observed protocol value	Directly measured during a transaction.	HTTP 200 received at 14:03 UTC.	Store raw evidence, observation time and vantage point.
Publisher-asserted value	A value deliberately presented by the publisher.	JSON-LD says the author is Alice.	Preserve source URL and do not present as independently verified.
Embedded assertion	A tag stored in a file by a device or application.	EXIF GPS says Seattle.	Preserve raw tag, namespace and parser version.
Deterministic derivative	A reproducible computation over delivered bytes.	SHA-256, dimensions or exact duration.	Record algorithm and version.
Probabilistic inference	An estimated conclusion with uncertainty.	Two profiles likely represent the same person.	Store method, confidence and evidence; allow correction.

Trust rule

Metadata is evidence, not proof. Author, GPS, software, camera and timestamp fields can be false, stale, copied or edited. Integrity becomes stronger when independently observed server logs, cryptographic hashes, trusted timestamps and digital signatures corroborate one another.

LOCAL STORAGE

What a computer stores with a file

A file is not merely a byte stream. The filesystem maintains a record that locates the bytes and controls how the operating system presents and protects them.

Windows NTFS catalog

Category	Fields available	Notes for collection
Identity and namespace	Filename, extension, parent directory, volume-relative file ID, hard-link names.	Rename does not change content hash; copying usually creates a new file ID.
Size and allocation	Logical length, allocated length, sparse ranges, compression state, block/extents mapping.	Logical and on-disk size can differ substantially.
Timestamps	Creation, last write, last access and metadata/MFT change times.	Copy tools, clock settings and disabled access-time updates affect meaning.
Security	Owner, group/security identifier, discretionary ACL, inheritance and auditing controls.	Permissions typically do not survive ordinary web upload/download.
Attributes	Read-only, hidden, system, archive, temporary, offline, encrypted, compressed, sparse and reparse-point flags.	Attributes belong to the filesystem record, not normally the main data stream.
Named streams	Default data stream plus alternate data streams such as Zone.Identifier.	May be lost on FAT, cloud upload, archives or non-NTFS destinations.
Reparse and link data	Symbolic link, junction, mount-point or cloud placeholder target information.	A scanner must decide whether to follow links and prevent cycles.
Indexing artifacts	Search index properties, thumbnail cache entries and recent-file records.	Often stored elsewhere by Windows rather than attached to the file.

Mark of the Web

Windows may attach a Zone.Identifier alternate data stream to a downloaded file. It can record an internet security zone and sometimes source or referrer information, allowing Windows, SmartScreen or Office to apply additional warnings. This is not embedded in the ordinary file stream and may disappear when transferred through systems that do not preserve NTFS streams.

Hashing boundary: a normal file hash generally covers the unnamed/default data stream only. NTFS alternate data streams are separate content and require separate enumeration and hashing if your goal is a complete forensic identity of the filesystem object.

Filesystem comparison

Filesystem	Metadata structure	Notable capabilities
NTFS	Master File Table records and named attributes.	ACLs, alternate data streams, reparse points, compression, encryption, sparse files.
ext4	Inodes plus directory entries and extended attributes.	Unix ownership/mode, links, xattrs, nanosecond timestamps, extents.
APFS	Object-based metadata and directory records.	Clones, snapshots, encryption, extended attributes, resource forks.
FAT/exFAT	Simpler directory records.	Broad compatibility but fewer permissions, streams and advanced metadata features.

What common operations preserve

Operation	Embedded metadata	Content hash	Filesystem metadata
Rename	Preserved	Unchanged	Name changes; most other fields remain.
Move on same volume	Preserved	Unchanged	File ID often remains; parent/name and metadata-change time can change.
Copy to another location	Preserved if bytes are unchanged	Unchanged	New file ID, owner, permissions and creation context.
Archive and extract	Usually preserved	Usually unchanged	Only the attributes supported and recorded by the archive survive.
Upload/download	Preserved only if service returns original bytes	May change if transformed	Destination creates a new local record.
Re-encode/transcode	Often rewritten or stripped	Changes	New destination record.

NETWORK TRANSFER

What accompanies a file over the internet

Ordinary network transfer sends the file's bytes inside protocol envelopes. Source filesystem ownership, ACLs and file IDs are not inherently transmitted.

Layer	Observable information	Persistence and caveats
Link	Local source/destination MAC addresses, frame type, VLAN and wireless properties where visible.	Replaced hop by hop; MAC addresses do not travel end to end.
IP	Source/destination IP, protocol, packet size, fragmentation and traffic timing.	VPNs, NAT, proxies and CDNs change the apparent endpoints.
TCP/QUIC	Ports, sequence/connection behavior, flags, retransmissions, flow and timing.	Transport metadata is not part of the recovered file.
TLS	Negotiated version/cipher/ALPN, server certificate chain, sizes and timing.	Application content is encrypted in transit; endpoints still see plaintext.
HTTP	Method, URL, status, redirects, Content-Type, length, disposition, validators, caching, cookies and custom headers.	Headers are optional claims. Referer may be absent or truncated.
Application/service	Account ID, upload time, original filename, storage key, hash, scan result and access history.	Stored by the service, not normally embedded in a downloaded copy.

Transfer sequence

```
Source filesystem record -> application reads the main data stream -> HTTP or another application protocol frames the bytes -> TLS encrypts the application data -> transport and network protocols deliver chunks -> destination reconstructs the representation -> destination creates a new filesystem record.
```

What is guaranteed to travel?

Only the bytes actually selected and transmitted are guaranteed to arrive. Embedded metadata travels when those bytes remain unchanged. Sidecars, extended attributes, alternate streams, permissions and local timestamps require an archive or protocol that explicitly preserves them. Social platforms and content delivery services may resize, recompress or transcode media, creating a new file with a new hash and a different metadata set.

Receiving server records

Record	Examples	Where retained
Request envelope	IP, time, route, method, request headers, content length, TLS and proxy context.	Web server, load balancer, CDN or WAF logs.
Object identity	Generated object key, original filename, declared/detected MIME, size, exact hash.	Object store and application database.
Content analysis	Dimensions, duration, codec, document type, malware scan, moderation classification.	Processing pipeline/index.
Account context	Uploader ID, organization, session, permission level and workflow state.	Application database and audit logs.
Usage history	Downloads, shares, edits, deletion requests and retention state.	Analytics and audit systems.

FILE FORMATS

Comprehensive extraction catalog

Different formats define different container structures and metadata namespaces. Every field is optional unless the format requires it, and extensible formats can carry application-defined fields.

Images and graphics

Format	Technical fields	Embedded/descriptive fields	Caveats
JPEG / TIFF / HEIC	Dimensions, bit depth, orientation, color space, compression, thumbnail.	EXIF camera/lens/software/settings, capture timestamps, GPS, owner/comment; IPTC creator/contact/caption/keywords/rights; XMP title, ratings, tool and history; ICC color profile; MakerNotes.	EXIF/IPTC/XMP tags are editable; web derivatives often strip GPS and MakerNotes.
PNG	Dimensions, bit depth, color type, interlace, gamma and physical resolution.	Text chunks, creation software, timestamps, ICC/sRGB profile, XMP and arbitrary ancillary chunks.	Ancillary chunks may be dropped by optimizers.
GIF	Canvas size, frame count, delays, loop behavior, palette and transparency.	Comments, application extensions and creation-tool hints.	Limited standardized descriptive metadata.
WebP / AVIF	Dimensions, animation, alpha, codec/transform properties and color information.	May carry EXIF, XMP and ICC profiles.	Support varies among encoders and platforms.
SVG	XML structure, dimensions/viewBox, paths, styles, IDs, text and linked assets.	Title, description, RDF/XMP, comments, generator and scripts/events.	SVG is active content; never execute scripts during parsing.
RAW camera formats	Sensor dimensions, previews, mosaic/layout, white balance and exposure data.	Extensive vendor MakerNotes, camera/lens serials, focus data, GPS and embedded JPEG previews.	Private fields vary by vendor and parser coverage.
C2PA / Content Credentials	Cryptographically signed provenance manifests when embedded or referenced.	Claim generator, ingredients, edits/actions, assertions, signatures and trust chain.	Presence can strengthen provenance; absence is not evidence of manipulation, and validation depends on trust/signature state.

Audio and video

Format	Technical fields	Embedded/descriptive fields
MP4 / MOV / M4A	Container brand, duration/timescale, tracks, codecs, bitrate, dimensions, rotation, color/HDR, frame rate, sample rate and channels.	Title, author, copyright, creation time, ISO-6709 location, chapters, subtitles, thumbnails and XMP/ID3-like keys.
Matroska / WebM	Segment/track IDs, duration, codecs, languages, video/audio properties and writing/muxing applications.	Title/date/tags, chapters, attachments, cover art, fonts and multilingual metadata.
MP3 / ID3	MPEG version/layer, bitrate, sample rate, channel mode and duration.	Title, album, artists, composer, track/disc, dates, genre, BPM, ISRC, lyrics, artwork, comments, URLs, ratings and private frames.
FLAC / Ogg	Sample rate, channels, bit depth, sample count, codec and PCM MD5 where supplied.	Vorbis comments, pictures, cuesheets, seek points and application-private blocks.
WAV / BWF	Codec, channels, sample rate, byte rate, bit depth and duration.	RIFF INFO tags; Broadcast Wave originator, time reference, UMID, coding history, loudness and XML/ADM.
Streaming manifests	Variant playlists, segment URLs, codecs, bandwidth, resolution, language and encryption declarations.	Program dates, titles, ad markers, subtitles, audio groups and session data.

FILE FORMATS

Documents, archives, code and executables

Documents and office formats

Format	Technical fields	Content/metadata	Caveats
PDF	PDF version, page count/geometry, encryption and permissions, document IDs, fonts, images, color spaces and compression.	Title, author, subject, keywords, creator, producer, creation/modification dates, XMP, annotations, forms, bookmarks, layers, actions/JavaScript, attachments, incremental revisions and digital signatures.	Redaction may leave hidden text or earlier revisions. Active content and attachments are hostile.
DOCX	OOXML package parts, styles, themes, sections, relationships, media and fonts.	Creator, last editor, revision, company, template, comments, tracked changes, hidden text, custom XML, links, embedded objects, macros in DOCM and signatures.	A ZIP-based compound document; external relationships can leak or trigger network access.
XLSX	Worksheets, formulas, cached results, names, tables, charts, pivots, relationships and calculation settings.	Authors, company, comments/notes, hidden sheets/rows/columns, external links, data connections, custom properties and macros in XLSM.	Formula results can be stale; do not evaluate untrusted formulas in a privileged environment.
PPTX	Slides, masters, layouts, themes, media, notes and animations.	Authors, company, comments, hidden slides, speaker notes, external links, embeds, custom properties and macros in PPTM.	Hidden slides and notes often contain unintended disclosures.
ODF	ZIP package structure, XML content/styles/settings and manifests.	Creator, generator, timestamps, comments, revisions, links, embedded objects and custom metadata.	Treat external references and macros as untrusted.
Plain text / Markdown	Byte length, newline convention, probable encoding, BOM and language/syntax inference.	Front matter, comments, headings, links, author fields if explicitly written.	Usually little standardized embedded metadata; most context comes from filesystem or repository.

Archives, executables and structured files

Format	Technical fields	Metadata/content	Caveats
ZIP / 7z / TAR / Gzip	Member paths, sizes, compression, CRC/hash, timestamps, permissions, ownership and comments.	Nested files, archive comment and format-specific extra fields.	Limit recursion and decompressed size; prevent path traversal and archive bombs.
Windows PE EXE / DLL	Architecture, subsystem, sections, entry point, imports/exports, resources, relocations and debug directory.	Version resources, product/company, manifest, Authenticode signature, certificate, PDB path and build hints.	Never load or execute untrusted binaries merely to inspect them.
OBJ / LIB	COFF sections, symbols, relocation records and archive members.	Compiler/linker identifiers, library members and debug references.	Import libraries and static libraries have different semantics.
ELF / Mach-O	Architecture, segments/sections, symbols, dynamic dependencies, loader paths and code signatures.	Build IDs, debug data, comments and toolchain hints.	Platform-specific parsers are required.
Source code	Language, encoding, imports, symbols, complexity and dependency declarations.	Comments, copyright, authorship headers, TODOs, secrets and repository context.	Secrets found in source should be quarantined, not indexed broadly.
JSON / XML / YAML	Schema, keys, types, namespaces, references and validation state.	Arbitrary application metadata and embedded identifiers.	Prevent XML external entity resolution, YAML object construction and oversized nesting.
SQLite / databases	Schema, tables, columns, indexes, triggers, page size and application/user versions.	Rows, deletion remnants, internal metadata tables and write-ahead logs when available.	Opening untrusted databases with extensions or custom functions can be dangerous.
Email EML / MSG	MIME structure, content types, attachment names/hashes and encoding.	From/To/CC, message IDs, routing Received headers, dates, subjects, clients, authentication results and attachments.	Headers can be spoofed; routing and cryptographic results provide corroboration.

Additional structured and specialized formats

Medium	Potential information	Primary concern
Calendar / ICS	UID, organizer, attendees, event times/time zones, recurrence, location, description, alarms, status and attachments.	Invitations and attendee addresses are personal data; recurrence/time-zone interpretation is subtle.
Contacts / vCard	Names, organizations, titles, phones, emails, addresses, URLs, photos, birthdays, notes and stable identifiers.	Highly linkable personal data; apply strict access and retention controls.
Installers / packages	Package name/version, publisher, dependencies, capabilities, manifest, payload list, scripts, target architecture and signature.	Inspect statically; never run install scripts in the parser environment.
Fonts	Family, style, version, foundry, license, naming records, glyph inventory and Unicode coverage.	Embedded-font licensing may restrict extraction or redistribution.
Git / repositories	Commits, authors, emails, timestamps, branches, tags, remotes, signatures, diffs and file history.	Repository history can retain deleted secrets and personal identifiers.
Cloud sync / version history	Provider object ID, sync state, sharing, revision history, conflict copies and retention/trash state.	Usually service metadata available only through an authorized account/API.
Disk images / containers	Partitions, filesystems, labels, boot data, snapshots and nested filesystem records.	Mount read-only or parse offline; enforce recursion and size boundaries.
CAD / GIS	Layers, units, coordinate systems, authors, revisions, object properties, geospatial extents and linked resources.	Can disclose precise facilities and locations; formats are application-specific.
DICOM / medical	Patient/study/series identifiers, acquisition equipment, clinical context, dates and pixel-data properties.	Extremely sensitive regulated data; de-identification requires domain-specific validation.

PUBLIC WEB COLLECTION

Infrastructure, pages, APIs and public people data

A public crawler can collect only the representation returned for its request identity and authorization state. Preserve raw HTTP, post-render DOM and browser network data as separate artifacts.

Infrastructure and protocol surfaces

Surface	Collectible information	Boundary/caveat
DNS	Returned record sets and TTLs: A/AAAA, CNAME, NS, SOA, MX, TXT, SRV, CAA, HTTPS/SVCB and DNSSEC.	DNS is not a directory; ANY is incomplete, split/geo DNS varies and zone transfer works only when authorized.
RDAP / registration	Public or redacted domain, IP and ASN registration data, registrar, status and important dates.	Privacy redaction and proxy registration limit attribution.
TLS	Negotiated TLS/cipher/ALPN; certificate subject/SANs, issuer, serial, validity, public key, policies, SCT/OCSP when presented.	Certificates reveal public material, not private keys or origin topology; CDN certificates may cover shared edges.
HTTP transaction	Requested and final URL, redirect chain, status, protocol, headers/trailers, body, compression and timing.	Headers are optional/untrusted. Same URL can vary by IP, cookies, language, time, A/B test and authentication.
Security policy	CSP, HSTS, CORS, Permissions-Policy, cache policy, cookies and reporting endpoints.	Describes the delivered representation's policy, not the entire server configuration.
Discovery	robots.txt, sitemaps, RSS/Atom, manifests, canonical/alternate links and pagination.	Hints are not completeness guarantees. Robots rules are crawler policy, not access authorization.

Page and application surfaces

Surface	Examples
Raw HTML	Title, meta description, robots directives, canonical/hreflang, link relations, text, comments, IDs/classes, data-* attributes, ARIA, forms, links, media and embedded state.
Structured data	JSON-LD, Microdata, RDFa and custom meta describing people, organizations, products, offers, prices, ratings, articles, events, places and media.
Assets	Scripts, styles, fonts, images, media, source maps when public, manifests, icons and integrity attributes.
Rendered DOM	Post-JavaScript content, shadow DOM visible to automation, lazy-loaded sections, client-generated links and state.
Browser network	Public fetch/XHR/GraphQL responses, SSE/WebSocket messages, redirects, API schemas revealed by use and dynamically requested assets.
Content extraction	Visible text, headings, tables, contacts, addresses, prices, dates, author lines, captions, alt text and accessibility structure.
Derived analysis	Language, OCR, transcription, topic/entity extraction, object/face presence, similarity, sentiment and content classification.

CORS limits what browser JavaScript may read; it is not authorization for your server-side crawler. Public client code may reveal an endpoint or public key without granting permission to use protected responses. Do not bypass login, paywall, CAPTCHA, signed-URL or other access controls.

PUBLIC WEB COLLECTION

People, organizations and public records

Category	Potentially observable fields	Reliability/privacy concerns
Profiles	Display name, handle, bio, profile URL, avatar, public contact fields, claimed location and linked accounts.	Self-asserted; handles and names are not unique identifiers.
Activity	Posts, comments, reactions, timestamps, edits, public memberships and interaction links.	Context can change; deletion and correction should propagate to your index.
Affiliations	Claimed employer, school, organization, role, publications and event participation.	Claims may be stale, parody or impersonation.
Contact data	Public email, phone, business address, contact forms and profile links.	Aggregation substantially increases privacy and abuse risk.
Attached media	Faces, captions, EXIF if retained, document authors, GPS, file names and embedded contact information.	Media derivatives often strip metadata; biometrics and exact GPS are sensitive.
Entity resolution	Likely matches across profiles based on names, handles, links, images and timelines.	Probabilistic, error-prone and potentially high-impact; require confidence and correction mechanisms.
Sensitive inferences	Possible health, religion, politics, ethnicity, sexual orientation, routines or precise locations.	Highest-risk category; block by default without a compelling, lawful and disclosed purpose.

What a public crawler cannot legitimately claim

- Completeness: dynamic, personalized, paginated and effectively infinite URL spaces prevent proof that everything was captured.

- Backend visibility: client responses do not reveal server source, databases, private object storage or original files that were never delivered.

- Identity certainty: IP addresses, names, usernames and fingerprints do not inherently identify one natural person.

- Truth: structured data, file metadata and profile statements are assertions that require corroboration.

- Permission: public accessibility does not automatically authorize reuse, redistribution, profiling or access-control circumvention.

Recommended crawler evidence set

```
1. Raw HTTP response bytes and response fields.
2. Post-render DOM snapshot.
3. Browser network log and public API response bodies.
4. Downloaded artifacts and cryptographic hashes.
5. Extraction output with parser/version.
6. Source URL, crawl time, vantage, request policy, robots decision, license and retention state.
```

FIRST-PARTY TELEMETRY

Information available from visitors to your website

Visitor data expands across four tiers: passive server observations, browser-reported values, permission/user-action data and externally enriched inferences.

Tier 1 - passive server and edge observations

Surface	Available fields	Boundary
Connection	Apparent source IP, connection time, source port, protocol, connection duration and transport behavior.	VPN, NAT, mobile gateways, proxies and CDNs change attribution.
Request	Method, host, path, query, HTTP version, request size and time.	Queries often contain tokens or personal data; redact before logging.
Headers	User-Agent, Accept, Accept-Language, Content-Type/Length, Referer, Origin, Sec-Fetch fields, Client Hints, cookies and GPC.	Many values are absent, privacy-reduced or spoofable; never treat as authenticated identity.
TLS/edge	TLS version/cipher, handshake traits, CDN/WAF bot score, threat/reputation and forwarding context.	What the origin sees depends on where TLS terminates.
Response	Status, response size, cache result, processing time, errors and retries.	Useful for performance, reliability and abuse monitoring.
IP enrichment	Approximate country/region, ASN, ISP/hosting provider, reverse DNS, VPN/proxy/Tor likelihood and reputation.	Derived and approximate; IP does not equal a person or exact physical address.

Cookie visibility

A server receives cookies the browser attaches to the request. Page JavaScript cannot read cookies marked HttpOnly and cannot access unrelated origins' cookies or storage. Retain session values only when necessary; raw cookie and authorization values should be excluded from ordinary analytics logs.

Tier 2 - JavaScript and page interaction telemetry

Category	Examples
Display/layout	Viewport and screen dimensions, device-pixel ratio, orientation, color depth and media-query preferences.
Locale/time	Languages, locale, time zone and formatting behavior.
Platform hints	Reduced browser/platform data, touch/pointer support, hardware concurrency and coarsened memory hints.
Navigation/performance	Navigation type, resource and paint timing, network timing, long tasks, responsiveness and layout shifts.
Behavior	Clicks, pointer movement, scrolling, focus, page visibility, key events on the page, selected controls and workflow progress.
Application state	First-party cookies, local/session storage, IndexedDB values, service-worker state and pseudonymous session ID.
Diagnostics	JavaScript errors, rejected promises, failed resources, feature support and client-generated traces.
Fingerprint surfaces	Canvas, WebGL, audio/rendering, fonts and combinations of configuration values.

Fingerprinting can estimate that a browser has returned without setting an ordinary cookie, but the result is probabilistic. Browsers increasingly reduce entropy and partition storage to discourage covert tracking. Treat fingerprinting as privacy-sensitive and avoid it unless a narrowly defined security purpose genuinely requires it.

FIRST-PARTY TELEMETRY

Permission-gated, submitted and unavailable data

Tier 3 - explicit permission or user action

Capability	Potential information	Boundary
Geolocation	Latitude/longitude, accuracy, altitude/speed where available.	Browser/user permission; collection purpose and retention must be disclosed.
Camera and microphone	Selected audio/video streams and device labels after permission rules allow exposure.	Requires user mediation and visible browser indicators.
Screen capture	User-selected screen, window or tab stream.	Explicit selection; permission is not a blanket right to retain everything shown.
Clipboard	Data involved in authorized clipboard read/write operations.	Browser rules vary and generally require secure context, permission or user activation.
Files/directories	Files the visitor selects or drops, including their contents, names, sizes, types and last-modified claims.	No silent enumeration of the full filesystem.
Device APIs	Selected Bluetooth, USB, serial, MIDI or HID device details in supported browsers.	Requires explicit selection/permission and is highly purpose-specific.
Notifications	Permission state, push subscription endpoint/keys and notification interaction.	Permission can be revoked; subscription identifiers are personal data when linkable.

Tier 4 - information deliberately submitted

Interaction	Potentially collected
Account/login	Email, username, account ID, authentication events, roles, preferences and account history.
Forms	Every field the visitor submits, validation results and workflow state.
Commerce	Cart, purchases, subscriptions, billing/shipping data and transaction references.
Uploads	Complete bytes, filenames, declared/detected types, metadata, hashes, scan results and derived content.
Support/community	Messages, posts, attachments, reports, moderation actions and communications history.
OAuth/federation	Claims the identity provider is configured and authorized to release.

Information a normal website cannot silently obtain

Category	Not reliably available
Direct identity	Real name, email or phone merely from an IP address.
Hardware identity	MAC address, device serial number or arbitrary stable hardware identifier.
Operating-system secrets	Windows username, saved passwords or arbitrary process/memory information.
Other origins	Other sites' cookies, storage, authenticated page content or complete browsing history.
Local storage	Arbitrary local files or directories that the visitor did not select/authorize.
Sensitive sensors	Exact GPS, camera, microphone or screen content without browser/user mediation.
Network origin	The visitor's underlying ISP IP when a correctly configured VPN/proxy conceals it.

APPLICATIONS

What metadata collection can accomplish

Metadata becomes most useful when normalized, indexed and compared across time while retaining provenance and uncertainty.

Use case	What the metadata enables
Search and discovery	Filter by author, date, camera, codec, dimensions, file type, organization, topic, location or document property.
Exact deduplication	Group byte-identical content by SHA-256 regardless of filename or path.
Near-duplicate discovery	Use perceptual fingerprints to identify resized, recompressed, cropped or lightly edited media.
Timeline reconstruction	Correlate filesystem, embedded, server, email and version-control timestamps; surface inconsistencies.
Provenance	Track source URL, retrieval time, hash, signature, parser and transformation lineage.
Privacy auditing	Locate GPS, contacts, usernames, serial numbers, hidden revisions, comments and sensitive embedded data before sharing.
Security	Detect extension/signature mismatch, suspicious imports/macros, unexpected active content, malicious attachments and known hashes.
Inventory/compliance	Catalog licenses, retention class, encryption, signatures, document owners, media codecs and unsupported formats.
Backup optimization	Detect changed content, reuse identical blobs, identify stale files and prioritize large or unique artifacts.
Media operations	Generate thumbnails, transcode unsupported codecs, validate duration/tracks, group camera collections and repair tags.
Web intelligence	Monitor prices, structured data, publications, public changes, technical posture and content availability.
First-party analytics	Measure performance, errors, navigation funnels, feature adoption, conversion and abuse without constructing cross-site dossiers.

Example queries

```
- All documents last edited by a specific account.
- Public images containing retained GPS or camera serial numbers.
- Byte-identical files with different names.
- Videos using unsupported codecs or missing captions.
- Pages whose declared canonical URL conflicts with the fetched URL.
- Visitor errors grouped by route template and browser family.
- Files whose extension disagrees with detected magic bytes.
- Records whose embedded capture time conflicts with server upload time.
```

APPLICATION BLUEPRINT

Recommended crawler and analytics architecture

Public-resource intelligence and first-party visitor analytics should be separate ingestion paths with separate access policies and retention schedules.

Processing pipeline

Stage	Responsibilities	Controls
Policy gate	Determine scope, authorization, robots decision, allowed formats, personal-data classes and retention.	Deny protected destinations, sensitive classes and out-of-scope hosts by default.
Fetcher	HTTP requests, redirects, response headers, raw bytes, conditional GET, rate limiting and retries.	Per-origin queues, time/byte limits, descriptive User-Agent and backoff.
Browser renderer	Execute public JavaScript when required; capture rendered DOM and network events.	Ephemeral profiles, no stored credentials, strict navigation and download policy.
Artifact router	Detect true format by signature and parser; calculate hash and basic dimensions.	Do not trust extension or declared MIME alone.
Sandboxed parsers	Extract metadata/text and enumerate container children.	No network, no macros/scripts, low privilege, CPU/time/memory/decompression/recursion limits.
Normalizer	Map common fields while retaining arbitrary namespace/key/value tags.	Preserve raw value, parser version, evidence class and confidence.
Indexes	Full-text, exact-hash, perceptual similarity, time, geospatial and structured-field indexes.	Role-based access and field-level protection for sensitive values.
Retention/rights	Expiration, deletion, correction, suppression, consent/GPC state and audit history.	Deletion must propagate through raw, normalized, index, cache and backup policies.

Data model

```
ARTIFACTS: Id, SourceUrl, RetrievedAt, DeclaredMime, DetectedMime, Size, Sha256, ParentArtifactId
METADATA: ArtifactId, Namespace, Key, ValueType, RawValue, NormalizedValue, EvidenceClass, Parser, ParserVersion,
Confidence, PersonalDataClass, RetainUntil
FETCHES: ArtifactId, RequestedUrl, FinalUrl, Status, HeadersJson, TimingJson, RobotsDecision, License, Vantage
VISIT_EVENTS: PseudonymousSessionId, Timestamp, RouteTemplate, EventType, PropertiesJson, ConsentState, RetainUntil
SUBJECT_RIGHTS: SubjectKey, RequestType, ReceivedAt, VerifiedAt, Status, Scope, CompletedAt
```

Use normalized columns for common queries and an extensible namespace/key/value or JSON field for arbitrary tags. File formats such as XMP, PDF and OOXML allow custom properties, so a fixed one-column-per-tag schema will always be incomplete.

Recommended extraction stack

Function	Tool/category	Role
Static web acquisition	.NET HttpClient or an equivalent standards-compliant HTTP client	Headers, bodies, redirects, conditional requests and streaming downloads.
Dynamic web acquisition	Playwright	Rendered DOM, browser context, lazy content and network events.
General document parsing	Apache Tika	Unified text and metadata extraction across a broad format catalog.
Media/image metadata	ExifTool	EXIF, XMP, IPTC, ICC, ID3, MakerNotes and many vendor namespaces.
Audio/video probing	ffprobe	Container, streams, codecs, tags, frames, chapters and subtitles.
File identification	Magic/signature detection plus declared MIME	Detect extension mismatches and route to appropriate parser.
Storage	SQLite for local/single-user; PostgreSQL for multi-user/service scale	Relational core plus JSON metadata, full-text and geospatial extensions as needed.

SECURITY AND PRIVACY

Operational guardrails

Maximum technical collection is not the same as maximum lawful or useful collection. A breach exposes everything you chose to retain.

Crawler etiquette and access boundaries

- Honor robots.txt, published rate limits, Retry-After and 429/503 responses.
- Identify the crawler and provide a working contact address.
- Prefer documented APIs and conditional requests using ETag or Last-Modified.
- Do not bypass authentication, paywalls, CAPTCHAs, signed URLs or other technical access controls.
- Review terms, licenses and copyright before storing or redistributing complete works.
- Apply explicit host, URL count, depth, redirect, time, byte and content-type limits.

Hostile-input controls

Threat	Control
SSRF	Resolve and validate destinations; block loopback, link-local, private, metadata-service and disallowed networks; revalidate redirects.
Archive bombs	Limit compressed and expanded bytes, member count, ratio, nesting depth and extraction time.
Path traversal	Normalize archive member paths; reject absolute paths and traversal outside the extraction root.
Parser exploits	Use patched parsers in isolated, low-privilege, no-network workers with hard resource limits.
Active content	Never execute Office macros, PDF JavaScript, SVG scripts, binaries, installers or embedded attachments.
XXE/deserialization	Disable external entities, schema fetching and unsafe object construction.
Secrets/PII	Quarantine and access-restrict credentials, tokens, contact data, GPS, biometrics, health and children's data.
Log leakage	Redact Authorization, Cookie, Set-Cookie, passwords, form bodies and sensitive query parameters before persistence.

Privacy baseline

Principle	Implementation
Purpose	Document why each field is collected and how it is used before ingestion.
Minimization	Collect only fields needed for that purpose; block sensitive categories by default.
Transparency	Provide clear notice covering categories, purposes, recipients, retention and rights.
Consent/choice	Gate nonessential cookies, trackers and permissioned APIs where required; honor Global Privacy Control when applicable.
Retention	Assign a RetainUntil value or review schedule to every personal-data class.
Security	Encrypt in transit/at rest, restrict access, rotate secrets and audit privileged queries.
Rights	Support access, correction, deletion, restriction/objection and suppression where applicable.
Children/sensitive data	Apply heightened controls or exclude entirely unless a specific lawful program requires collection.

Public availability is not a universal privacy or copyright exemption. In the United States, FTC unfair/deceptive-practices enforcement and state privacy laws can apply. CCPA-covered businesses may need notice, proportionality, rights handling and Global Privacy Control support. GDPR can apply to relevant EU personal data and requires a lawful basis, transparency, purpose limitation, minimization, retention limits and security. COPPA adds requirements for child-directed services or actual knowledge of collection from children under 13. This report is a technical reference, not legal advice.

INSPECTION TOOLKIT

PowerShell commands for local exploration

Basic filesystem and hash information

```
$file = Get-Item .\photo.jpg

$file | Select-Object FullName, Length, CreationTimeUtc, `
LastWriteTimeUtc, LastAccessTimeUtc, Attributes

Get-Acl $file.FullName | Format-List Owner, Access
Get-FileHash $file.FullName -Algorithm SHA256
```

NTFS streams and Mark of the Web

```
Get-Item $file.FullName -Stream *

Get-Content $file.FullName -Stream Zone.Identifier -ErrorAction SilentlyContinue
```

Compare copies

```
Get-FileHash .\original.jpg, .\renamed.jpg -Algorithm SHA256 |
Select-Object Path, Hash

# Rename or byte-for-byte copy: hash should match.
# Re-save, re-encode or edit metadata: exact hash should change.
```

External inspection commands

Tool	Example	Purpose
ExifTool	exiftool -G1 -a -s file.jpg	Show duplicate tags, namespaces/groups and raw tag names.
ffprobe	ffprobe -v error -show_format -show_streams -of json video.mp4	Emit container, stream and tag data as JSON.
Apache Tika	java -jar tika-app.jar --metadata document.pdf	Extract generalized document metadata.
PDF info	pdfinfo document.pdf	Display common PDF structure and document properties.
File signature	file --mime-type artifact.bin	Infer content type from signature rather than extension.

Safe test exercise

- Hash an image, rename it and verify the digest remains unchanged.
- Copy it to another directory and compare creation time, file ID, ACL and hash.
- Change one metadata field or re-save it and observe the exact hash change.
- Inspect the original and re-saved copies with ExifTool and compare removed or rewritten tags.
- Upload the file to a test service you control, download the returned representation and compare hashes and metadata.

AUTHORITATIVE REFERENCES

Standards, specifications and official guidance

Reference	Official URL
HTTP Semantics - RFC 9110	https://www.rfc-editor.org/rfc/rfc9110.html
TLS 1.3 - RFC 8446	https://www.rfc-editor.org/rfc/rfc8446.html
Robots Exclusion Protocol - RFC 9309	https://www.rfc-editor.org/rfc/rfc9309.html
HTML Living Standard	https://html.spec.whatwg.org/
Fetch Standard	https://fetch.spec.whatwg.org/
Schema.org data model	https://schema.org/docs/datamodel.html
W3C JSON-LD 1.1	https://www.w3.org/TR/json-ld11/
W3C Permissions	https://www.w3.org/TR/permissions/
W3C Geolocation	https://www.w3.org/TR/geolocation/
W3C Fingerprinting Guidance	https://www.w3.org/TR/fingerprinting-guidance/
W3C File API	https://www.w3.org/TR/FileAPI/
W3C Media Capture and Streams	https://www.w3.org/TR/mediacapture-streams/
W3C Screen Capture	https://www.w3.org/TR/screen-capture/
W3C Clipboard API	https://www.w3.org/TR/clipboard-apis/
CIPA Exif standards	https://www.cipa.jp/e/std/std-sec.html
PNG Specification, Third Edition	https://www.w3.org/TR/png-3/
Adobe XMP specifications	https://developer.adobe.com/xmp/docs/xmp-specifications/
IPTC Photo Metadata	https://www.iptc.org/std/photometadata/specification/IPTC-PhotoMetadata
QuickTime File Format	https://developer.apple.com/documentation/quicktime-file-format
Matroska - RFC 9559	https://www.rfc-editor.org/rfc/rfc9559.html
FLAC - RFC 9639	https://www.rfc-editor.org/rfc/rfc9639.html
PDF 2.0 - ISO 32000-2	https://www.iso.org/standard/75839.html
Office Open XML - ECMA-376	https://ecma-international.org/publications-and-standards/standards/ecma-376/
C2PA Content Credentials	https://spec.c2pa.org/specifications/specifications/2.4/index.html
Microsoft NTFS File Streams	https://learn.microsoft.com/en-us/windows/win32/fileio/file-streams
ExifTool	https://exiftool.org/
Apache Tika	https://tika.apache.org/
ffprobe	https://ffmpeg.org/ffprobe.html
Playwright navigation/rendering	https://playwright.dev/docs/navigations
FTC Privacy and Security	https://www.ftc.gov/business-guidance/privacy-security
California CCPA FAQ	https://cppa.ca.gov/faq.html
California Global Privacy Control	https://oag.ca.gov/privacy/ccpa/gpc
European Commission GDPR principles	https://commission.europa.eu/law/law-topic/data-protection/information-business-and-organisations/principles-gdpr_en
FTC COPPA Rule	https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa
U.S. Copyright Office - protected subject matter	https://www.copyright.gov/help/faq/faq-protect.html

Report note: metadata fields and browser surfaces evolve. Treat this guide as a technical architecture reference, validate parser behavior against representative test files and review applicable law for the jurisdictions and use cases involved.