CRSTBL Inc.

WHITE PAPER

The New Discoverability Stack:

How OKF and MCP Are Building the New Rails for AI-Mediated Discovery

July 2026 • v2.0

Prepared by CRSTBL Inc. — Conversational Discovery for Local Business

Contents

1. Executive Summary

The way consumers discover businesses is undergoing a structural shift. Traditional search engine results pages, built on keyword matching and link authority, are giving way to AI-generated answers that synthesize information from structured data, authoritative content, and real-time tool access. This transition from ranked links to direct answers has created an entirely new optimization discipline, one that favors structured, machine-readable knowledge over keyword density and backlink volume.

Three layers of open standards now define this shift. First, schema.org and JSON-LD provide the structural foundation that search engines and AI systems already consume today. Second, Google's Open Knowledge Format (OKF) introduces a vendor-neutral specification for packaging organizational knowledge as portable Markdown bundles that AI agents can read directly. Third, the Model Context Protocol (MCP) standardizes how AI agents discover and invoke tools against external data sources in real time. Together, these three layers form a discoverability stack: schema.org structures what a business is, OKF packages what it knows, and MCP makes that knowledge queryable.

This paper examines how these standards facilitate semantic search, why they matter for businesses pursuing Answer Engine Optimization (AEO) and Generative Search Optimization (GEO/GSO), and how verified provenance addresses the critical trust gap that stands between structured content and AI citation.

2. The Business Discoverability Crisis

2.1 From Blue Links to AI Answers

In 2024, 56% of news-related Google searches resolved without a click to any website. By mid-2025, that figure had risen to 69%. Gartner projected that traditional search engine volume would decline 25% by 2026 as consumers shifted to AI chatbots and virtual agents. ChatGPT reported over 900 million weekly active users by early 2026, roughly double the prior year.

The implication for businesses is stark: consumers are increasingly receiving synthesized answers rather than lists of links. The brand that appears in an AI-generated answer captures the intent; the brands that don't appear are invisible. This is not a refinement of SEO. It is a category change in how attention is allocated.

2.2 The SMB Disadvantage

Large enterprises already maintain structured data layers, knowledge graphs, API integrations, and schema.org markup that make them legible to AI systems. More than 99% of U.S. businesses are small to medium-sized, and the vast majority have none of this infrastructure. Research shows that 76% of local searches result in same-day store visits and 80% of consumers search for local businesses weekly, yet only 45% of brands leading in traditional local search also appear among the most recommended businesses in AI-generated results.

This 55% visibility gap represents businesses that are findable on Google but invisible to the AI assistants increasingly mediating consumer decisions. The gap exists because AI answer engines are not ranking

pages. They are evaluating confidence: whether a business's information is structured enough to extract, consistent enough to trust, and authoritative enough to cite.

3. The Three-Layer Discoverability Stack

The emerging discoverability infrastructure is not a single standard but a stack of three complementary layers, each solving a different problem. Conflating them, or building on the wrong one first, leads to infrastructure that is technically sound but practically invisible.

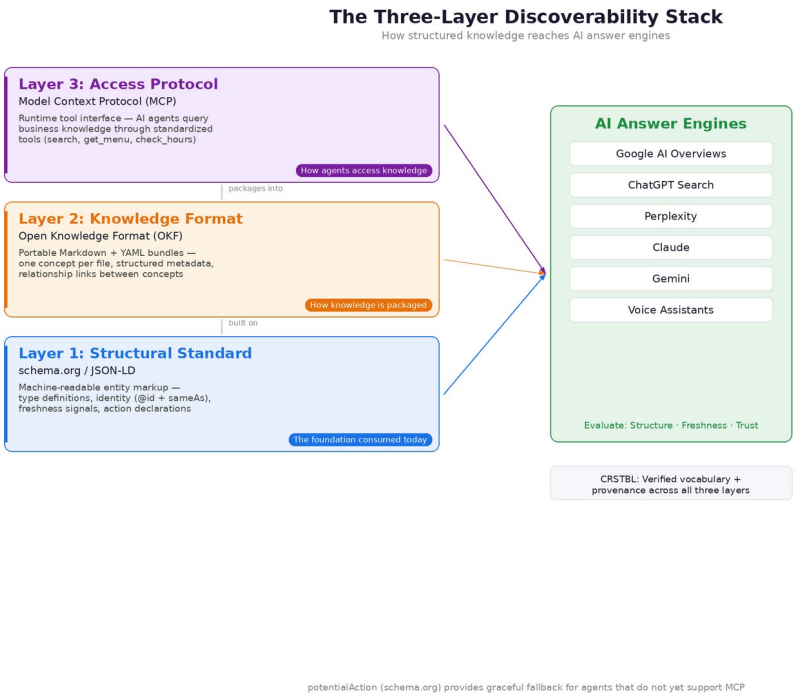

Figure 1: The three-layer discoverability stack and its relationship to AI answer engines

3.1 Layer 1: schema.org / JSON-LD — The Foundation Consumed Today

schema.org provides the machine-readable type system that search engines and AI systems actively consume. JSON-LD is Google's explicitly recommended serialization format. The installed base of consumers is enormous: every major search engine, every AI answer engine performing retrieval-augmented generation, and a growing number of voice assistants parse schema.org markup as a primary input.

For local businesses, the relevant types are well-defined: **Restaurant** (or the broader **LocalBusiness**), **Menu**, **MenuItem**, **OpeningHoursSpecification**, **Offer**, **FAQPage**, and **AggregateRating**. Four properties carry disproportionate weight for discoverability:

@id and sameAs form the identity layer. A business's `sameAs` array pointing to its Google Business Profile, Yelp listing, and Apple Maps entry makes the cross-directory fragmentation problem machine-readable and solvable.

dateModified is the freshness signal AI systems weight for recency-sensitive queries. An audit trail of changes is what makes a freshness bump defensible rather than manipulative.

Offer.validFrom and validThrough give promotions machine-readable expiry, eliminating stale-offer bleed in AI-generated recommendations.

potentialAction (`ReserveAction`, `OrderAction`) with an `EntryPoint` is the declarative ancestor of MCP tool-calling. These action declarations serve as the graceful fallback for AI agents that do not yet support MCP, ensuring businesses are actionable even when the newer protocol layer is unavailable.

3.2 Layer 2: Open Knowledge Format (OKF) — Portable Knowledge for Agents

Published by Google Cloud in June 2026 as version 0.1 and updated to v0.2 in July 2026, the Open Knowledge Format is an open specification for representing knowledge as a directory of Markdown files with YAML frontmatter. Each file represents one concept. The format requires only a single field (type) and supports optional metadata including title, description, tags, timestamps, and relationship links between concepts.

OKF addresses a problem schema.org does not: how to package a body of knowledge that goes beyond entity markup. A restaurant's schema.org markup declares what the business is; its OKF bundle describes everything the business knows, including FAQ content, policy details, seasonal changes, and the relationships between these concepts. An AI agent reasoning about whether a restaurant can accommodate a gluten-free group dinner on a Saturday night needs more than type markup. It needs traversable knowledge.

OKF is explicitly **not a search ranking signal**. Google's own documentation is clear on this point. But the distinction between "not a ranking signal" and "not relevant to discoverability" misses the trajectory. As answer engines increasingly pull from structured, agent-readable sources, OKF-packaged content becomes more citable and retrievable, not through traditional crawling, but through direct ingestion or MCP-exposed tool access.

3.3 Layer 3: Model Context Protocol (MCP) — Runtime Access

The Model Context Protocol, introduced by Anthropic in November 2024 and donated to the Linux Foundation's Agentic AI Foundation in December 2025, is an open JSON-RPC 2.0 protocol that standardizes how AI model clients discover and invoke tools on external servers. The Python and TypeScript SDKs see approximately 97 million monthly downloads. The protocol has been adopted by OpenAI, Google DeepMind, Microsoft, and thousands of development teams.

MCP's July 2026 specification release candidate introduced a stateless protocol core, enabling MCP servers to scale horizontally on ordinary HTTP infrastructure. The practical architecture is straightforward: an MCP server can read an OKF bundle, parse its structured content, and expose tools that allow any connected AI agent to query a business's knowledge base through a standardized interface.

Where schema.org declares what a business is, and OKF packages what it knows, MCP provides the live interface through which AI agents access that knowledge at query time. This is the layer that enables real-time answers, not cached crawl data.

3.4 How the Layers Interact

The three layers are complementary, not competing. schema.org is the structural foundation: the type system and identity graph that today's consumers already parse. OKF packages knowledge *built on* that foundation into a format AI agents can traverse concept by concept. MCP provides the runtime protocol through which agents access both layers at query time.

The practical implication is that schema.org should be shipped first because it has demonstrated consumption today. OKF should be shipped alongside it as the forward bet. MCP provides the access layer for both. And potentialAction declarations in schema.org serve as the graceful fallback for agents that do not yet support the MCP protocol.

4. Semantic Search for Local Business Discovery

4.1 The Structured Knowledge Gap

Consider a consumer asking an AI assistant: "What's a good family-friendly Italian restaurant near downtown with outdoor seating and gluten-free options?" Answering this query requires structured knowledge about cuisine type, service model, dietary accommodations, ambiance, location, and hours. Today, this information is scattered across Yelp listings, Google Business Profiles, the restaurant's website (if it exists), and various third-party directories, often inconsistent and rarely machine-readable in a unified format.

The three-layer stack offers a path forward. A restaurant's knowledge can be structured as schema.org entity markup for identity and type, packaged as an OKF bundle for traversable depth (one concept file for the business entity, one for the menu, one for hours, one for policies, one for each frequently asked question), and exposed through an MCP interface for real-time agent access.

4.2 Demand-Responsive Vocabularies

Effective semantic search depends on consistent vocabularies: a controlled set of tags and type definitions that allow similar concepts to cluster together and reduce retrieval noise. For the restaurant and local business vertical, this means standardizing terms for cuisine types, service models, dietary accommodations, price tiers, ambiance, and occasion context.

Crucially, schema.org standardizes the *container* but not the *contents* for most of these dimensions. `servesCuisine` is an unconstrained string. `amenityFeature` is a name/value pattern with no controlled values. The one exception is `RestrictedDiet`, which provides a closed enumeration (GlutenFreeDiet, VeganDiet, HalalDiet, KosherDiet, etc.) that implementations should align to 1:1 for lossless projection. For everything else, the gap between "we have a field for this" and "we have a consistent vocabulary for what goes in the field" is precisely where vertical-specific controlled vocabularies create value.

These vocabularies must be built from consumer behavior, not category theory. Canonical labels should reflect how consumers actually search: "date-night" rather than "romantic-dining-experience," "kid-friendly" rather than "family-oriented-establishment." And they must be continuously responsive to real demand, not constructed once and left static. As new questions emerge from actual consumer

interactions, the vocabulary evolves to capture them, ensuring that the knowledge base reflects how people actually talk about businesses rather than how someone predicted they would.

When these vocabularies carry relational structure, not just labels, the benefits compound. **Broader, narrower, and related** relationships between terms allow an AI agent to expand context beyond the immediate query: a guest asking about group dining gets the private event policy surfaced without explicit linking. When this relational structure is rich enough, the information architecture of the knowledge surface can be derived from the ontology itself, ensuring that navigation reflects how consumers actually think about a business.

4.3 MCP as the Access Layer

Once a business's knowledge is structured, vocabularized, and packaged, an MCP server exposes it to any connected AI agent. The server can offer tools for semantic search across the business's knowledge base, menu retrieval, hours and availability checks, and policy lookup. The AI agent never needs to scrape a website, parse unstructured HTML, or reconcile conflicting directory listings. It gets structured, authoritative, current answers through a standardized protocol.

This architecture inverts the traditional discovery model. Instead of relying on a search engine to crawl, index, and rank a website, the business pushes its knowledge into a structured format and exposes it through a tool interface that answer engines can query directly. Importantly, certain classes of business intelligence, such as which questions consumers ask most frequently and how demand patterns shift over time, remain private behind the MCP interface. The protocol exposes answers; it does not expose analytics. This separation preserves the business's competitive insight while maximizing its discoverability.

4.4 Agent-First, Human-Readable: The New Content Surface

The three-layer stack implies a class of content that did not exist in the traditional web: pages whose primary audience is AI agents rather than human browsers. This is a structural novelty. The web was built on the assumption that every page's audience is a person. Crawlers were secondary consumers, extracting signals from content designed for human eyes. The emerging discoverability infrastructure inverts this: a business's knowledge surface can be purpose-built for agent consumption, structured from the ground up to be parsed, queried, and cited by AI systems, while remaining human-readable to satisfy both the visible-content constraints of search engine guidelines and the merchant's own need to audit what is published on their behalf.

Existing standards do not yet formally represent this distinction. schema.org's `Audience` type and `audienceType` property were designed for human audience segments: `PeopleAudience`, `BusinessAudience`, `EducationalAudience`. There is no `AgentAudience` or `MachineAudience`. In practice, the signal that content is agent-optimized is expressed through other mechanisms: `llms.txt` files that index key pages for LLM consumption, clean semantic HTML with comprehensive JSON-LD, and MCP tool interfaces that only agents can invoke.

This gap points to a broader infrastructure reality: the agentic future requires building new operational rails for web content, not just optimizing existing ones. Traditional website operations were designed around a browser rendering a page for a human. Agent-mediated discovery demands a parallel set of concerns: Is the content structured for extraction, not just display? Is the knowledge graph traversable,

not just navigable? Does the page declare its entities, relationships, and freshness in machine-readable form, not just in prose? Can an agent act on the information through a standardized protocol, not just read it?

These are not SEO refinements. They are infrastructure requirements for a channel that operates on fundamentally different mechanics. Businesses and platforms that build these rails, treating agent readability as a first-class operational discipline alongside human usability, will be structurally advantaged as AI-mediated discovery scales. Those that treat agent optimization as a layer applied after the fact to content built for humans will find the retrofit increasingly inadequate.

5. The Trust Gap: Provenance in Agent-Mediated Discovery

Structured content and tool access solve the format and access problems. They do not solve the trust problem. When an answer engine assembles a response, it must determine which sources to cite. Traditional search uses domain authority, backlinks, and E-E-A-T signals. schema.org provides identity and publisher markup. OKF's v0.2 specification introduced provenance metadata and Attested Computation concepts for verifiable claims. But the ecosystem has not yet established how an answer engine should determine which knowledge bundle to trust when multiple sources offer conflicting information about the same business.

This trust gap is especially acute for local businesses. Anyone can create an OKF bundle or emit schema.org markup claiming to represent a restaurant. Without a verification layer, structured knowledge is just structured noise. The provenance question, where did this claim come from, and was it produced by someone with authority to make it, is the unsolved problem in the structured-knowledge-to-AI-citation pipeline.

The core principle is that a verified knowledge platform must **enrich the merchant's canonical entity, never create a competing one**. Two nodes claiming to be the same business is precisely the conflicting-source problem that structured knowledge was supposed to solve. A platform that creates it in its own infrastructure is self-defeating.

6. CRSTBL's Verified Knowledge Framework

CRSTBL's conversational discovery platform addresses the trust gap through a four-tier provenance model that establishes and maintains the chain of authority from business owner to AI citation.

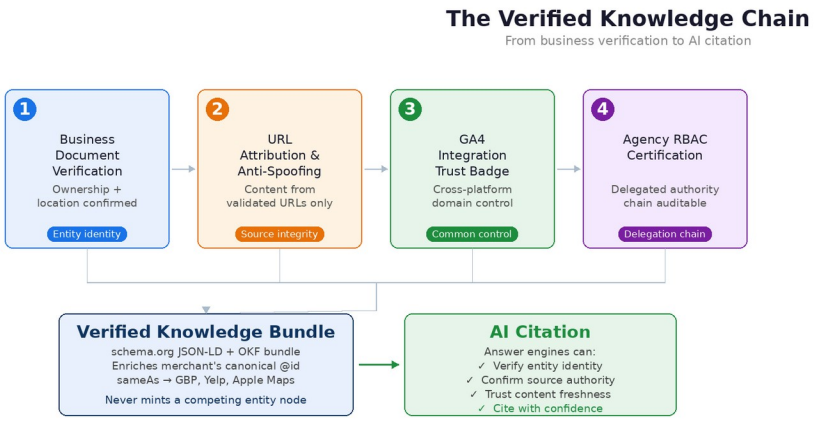

Figure 2: The four-tier provenance chain from business verification to AI citation

6.1 Tier 1: Business Document Verification

Every CRSTBL account begins with the submission of a business document that confirms ownership and physical location. This is not a self-asserted claim; it is a verified source of truth. The business entity's identity, location, and operational status are validated against primary documentation before any content enters the system. This baseline verification ensures that the knowledge generated for a business is tied to a confirmed, real-world entity.

6.2 Tier 2: URL Attribution and Anti-Spoofing

CRSTBL attributes conversational content only from validated URLs belonging to the verified entity. This prevents a critical attack vector in agent-mediated discovery: content spoofing. By binding the knowledge extraction process to authenticated web properties, CRSTBL ensures that the content feeding the knowledge base originates from the business itself, not from third-party sites that may misrepresent the business's offerings, hours, or policies.

6.3 Tier 3: GA4 Integration and Common Control

Businesses that connect their Google Analytics 4 accounts to the CRSTBL platform receive additional trust verification. This integration serves as the strongest available common-control signal: it confirms that the entity controlling the CRSTBL account also controls the web analytics for the associated domain. A shared GA4 property across a business's primary domain and its knowledge surface provides a genuine, cross-platform evidence of unified ownership that requires no new infrastructure to establish.

6.4 Tier 4: Agency-Managed Account Certification

For larger organizations whose digital presence is managed by marketing agencies, CRSTBL's role-based access control system includes a business relationship certification layer. The agency's authority to act on behalf of the client is formally attested within the platform, creating an auditable chain of delegation from business owner to managing agent to published content.

6.5 Outward Projection: From Platform Trust to Machine-Readable Identity

Internal verification is necessary but not sufficient. The trust established through these four tiers must be projected outward in a format that AI systems can evaluate without knowing anything about the CRSTBL platform. This is where schema.org's identity layer becomes the bridge.

A verified business's `@id` resolves to its own primary domain, not to a platform URL. Its `sameAs` array references its Google Business Profile, Yelp listing, and Apple Maps entry, making the fragmentation problem from Section 2 machine-readable and solvable. `dateModified` updates are backed by an auditable change-reason ledger. And the knowledge surface is hosted on the merchant's own domain, not on the platform's, so every signal reinforces the merchant's canonical entity rather than creating a competing node.

Together, these mechanisms transform structured knowledge from "content from an unknown source" to "verified knowledge from a confirmed business entity with authenticated web properties and cross-platform validation," expressed in a format that any AI system can evaluate independently.

7. Implications for AEO and GEO/GSO Strategy

7.1 Answer Engine Optimization (AEO)

AEO is the practice of structuring content so AI platforms can extract, trust, and cite it as a direct answer. Research from AirOps' 2026 State of AI Search Report found that for commercial and evaluation-stage queries, 83% of AI citations came from pages updated within the past 12 months, with over 60% refreshed within six months.

The three-layer stack directly addresses each of the three variables AEO practitioners can control. Structure: schema.org provides the type system AI systems already parse, and OKF extends it into traversable knowledge. Freshness: MCP enables real-time tool access to current business data rather than cached crawl data, and auditable change-reason tracking makes freshness signals defensible. Credible sourcing: verified provenance provides the chain of authority that establishes source trust, projected outward through schema.org identity properties that any system can evaluate.

7.2 Generative Search Optimization (GEO/GSO)

GEO is the broader strategic discipline of ensuring a brand appears accurately and favorably inside AI-generated answers across platforms. While AEO focuses on individual content selection, GEO encompasses the entire source ecosystem that shapes how an AI system perceives and represents a business.

For local businesses, GEO has historically required maintaining consistency across Google Business Profile, Yelp, Apple Maps, social media, and the business's own website, a fragmented and labor-intensive process. The three-layer stack offers a unifying architecture: schema.org identity properties (sameAs, @id) formally declare the entity across directories, the OKF bundle serves as the portable knowledge layer that feeds multiple serialization targets, and MCP provides the live interface for real-time queries. Changes propagate through the stack, ensuring that every AI agent querying the business gets the same current, authoritative information regardless of which layer it accesses.

7.3 The Compounding Advantage

The strategic implication is that AEO and GEO/GSO are converging from optimization tactics into infrastructure problems. The businesses and platforms that build the structured knowledge layer, the tool access layer, and the trust layer first will establish the default sources that answer engines cite. Early movers in schema.org adoption gained disproportionate visibility in featured snippets and knowledge panels. The same dynamic is emerging with the full stack: structured, verified, agent-accessible knowledge will compound its citation advantage over unstructured alternatives as AI-mediated discovery grows.

8. Conclusion

The transition from search engines to answer engines represents the most significant structural change in business discoverability since the rise of Google. Traditional SEO is not going away, but it is no longer sufficient. Businesses that want to be found by the AI systems increasingly mediating consumer decisions need three things: structured knowledge that machines can parse without guessing, a standardized access protocol that allows AI agents to query that knowledge directly, and a trust layer that gives answer engines confidence in the source.

schema.org provides the structural foundation consumed today. OKF provides the portable knowledge format for the agent-mediated future. MCP provides the runtime access protocol. And verified provenance, projected outward through machine-readable identity, provides the trust layer.

Critically, this transition demands new operational infrastructure, not just new optimization tactics. The agentic web requires content surfaces built for machine extraction alongside human readability, knowledge graphs that agents can traverse rather than just pages that browsers can render, and standardized protocols through which AI systems can query and act on business information in real time. These are new rails, not refinements to existing ones, and the platforms and businesses that build them first will define the default sources that answer engines learn to trust.

The cost of adoption is low: schema.org is a well-documented standard, OKF is plain Markdown, MCP is an open protocol, and verification is a one-time setup process. The cost of inaction is high: invisibility in the fastest-growing channel for consumer intent. For local businesses, this is not an optimization exercise. It is an existential one.

References

Google Cloud. “How the Open Knowledge Format Can Improve Data Sharing.” June 2026.

Google Cloud Platform. “OKF Specification v0.2 (SPEC.md).” GitHub, July 2026.

Anthropic / Agentic AI Foundation. “Model Context Protocol Specification, 2026-07-28 Release Candidate.” May 2026.

WorkOS. “Everything Your Team Needs to Know About MCP in 2026.” March 2026.

schema.org. “Release 30.0.” March 2026.

AirOps. “2026 State of AI Search Report.” 2026.

Similarweb. “Generative Engine Optimization: The Complete 2026 Guide.” March 2026.

Duda. “AEO-Ready SMB Sites Earn 3.2X More Traffic.” June 2026.

SOCi. “2026 Local Visibility Index.” 2026.

Aggarwal, P. et al. “GEO: Generative Engine Optimization.” 2023.

Karpathy, A. “LLM Wiki Gist.” April 2026.
