From SEO to AIO: How Hotels Must Adapt to Search 2.0

Stephen Burns
Web Intelligence Lead
Common Crawl Foundation

October 1, 2025

COMMON CRAWL

Technical Context

Stephen Burns

- Web Intelligence Lead, Common Crawl Foundation
- Enterprise SEO at U.S. Bank
- 25+ years: AOL, Open Directory Project, Topix, Blekko search engine

Common Crawl's Infrastructure:

- 10 Petabyte web dataset, 1000+ downloads/year
- Powers OpenAI, Meta, Amazon, Google training pipelines
- CCBot crawls 3B pages/month from trillion-page index
- **We maintain the largest open training corpus for LLMs**

COMMON
CRAWL

The AI Discovery Shift

The fundamental change in how travelers find hotels:

From decades of optimizing for short, typed queries to travelers having full conversations with AI assistants about their travel needs.

Traditional search behavior:

- "Hotels Bangkok" (2-3 words)
- Ten blue links to compare
- Multiple booking sites to check

AI-powered travel planning:

- *"I'm planning a trip to Bangkok in December, need something boutique near the Grand Palace with a pool, family-friendly, under $200 a night, what are my options?"*
- 20+ words, often spoken
- Single synthesized answer with recommendations

The shift: From ranking in search results to being discoverable by AI travel assistants

Training Data Engineering Reality

Common Crawl's technical impact:

- 10,000+ research papers cite our dataset (exceeds Hubble Space Telescope)
- GPT-3 (June 2020) triggered exponential research using our corpus
- **Every foundation model with web knowledge used our preprocessing pipelines**

The technical imperative:

> "If your hotel data isn't in training corpora, LLMs lack entity awareness for query expansion and retrieval ranking."

Implementation reality:

- Preprocessing filters determine what makes it into training sets
- CDN configurations and robots.txt now affect AI model knowledge
- Your infrastructure choices impact model behavior downstream

COMMON
CRAWL

Technical Architecture Shift - Search 1.0 vs 2.0

Search 1.0 Infrastructure:

- Centralized index (Google's ~50PB web index)
- PageRank + keyword matching algorithms
- Revenue: $300B ads + $80B SEO optimization industry

Search 2.0 Architecture:

- Distributed training: Foundation models + fine-tuning + RAG layers
- Multiple providers with different training corpora
- Revenue models unclear: paid inclusion, ads, affiliate, API licensing

Query processing evolution:

- 1.0: "hotels bangkok" → TF-IDF matching → ranked URLs
- 2.0: "boutique hotels Grand Palace area December family pool under $200" → semantic embedding → entity extraction → multi-step retrieval → synthesized response

Infrastructure implications: Different optimization targets, different failure modes

The Hospital Story - A Cautionary Tale

Real-world example of AI invisibility:

Children's Hospital of Los Angeles (CHLA) - one of America's top pediatric cancer hospitals:

- Parents searching "*where should I take my child with leukemia in LA?*"
- **CHLA doesn't appear in AI responses**
- Why? Hospital's site behind Cloudflare with default AI crawler blocks

The stakes:

- Not just lost website traffic
- Lost patients, potentially lost lives
- World-class hospital invisible to families in crisis

Hotel parallel: Your property could be invisible to AI-powered travel planning

How AI Actually Gets Hotel Information

Stage 1: Foundation Training (The Base Knowledge)

- Trained on Common Crawl's 10PB web dataset
- AI learns "Marriott is a hotel chain" but info may be 6-18 months old
- Your hotel gets baseline recognition but possibly stale details

Stage 2: Fine-Tuning (The Specialization)

- Travel-specific Q&A training data
- "Alice and Bob's Travel Chatbot" gets extra training on hospitality
- **Key insight: Hotels with more Q&A content get better AI representation**

Stage 3: Real-Time Retrieval (The Fresh Data)

- Live web searches during conversations
- AI pulls current information about rates, reviews, availability
- **This is where up-to-date website content matters most**

The integration challenge: AI needs to combine all three sources to give travelers complete, current information about your property

COMMON CRAWL

Interesting Data - What Actually Gets Into AI Training

From our Common Crawl dataset observations:

Foundation Training Data Quality:

- Hotels with better SEO tend to have richer AI representation
- **Stale information problem:** Training data lags 6-18 months behind reality
- Competitor information gets equal weight to your own content
- **Language bias:** 43% of training data is English, smaller languages underrepresented

Fine-Tuning Datasets (The Story Layer):

- Travel-specific Q&A pairs are expensive and limited
- **Key insight:** Hotels that create structured Q&A content get better AI responses
- Generic travel chatbots know less about individual properties than specialized ones

What's Missing from AI Knowledge:

- Real-time rates and availability (training data is static)
- Current promotions and seasonal offerings
- Recent renovations or new amenities
- **The gap:** AI has your basic info but lacks what travelers actually need to book

COMMON
CRAWL

The Opt-Out Crisis Hitting Hotels

What we're seeing at Common Crawl:

- Wave of publishers demanding removal from AI training
- Legal threats and takedown requests
- **Permanent exclusion** - once out, never back in

The hotel industry risk:

- CDNs (like Cloudflare) blocking AI crawlers by default
- Legal departments opting out without understanding consequences
- Marketing unaware their property is invisible to AI

The irony: AI still discusses your hotel (through reviews, forums) but without your authoritative voice

The New Risk - From Ranking to Existence

Traditional SEO risk: Dropping a few positions in search results

AI era risk: Being erased entirely from discovery systems

For hotels:

- Not competing for rank anymore
- **Competing for existence**
- Many properties already invisible without knowing it

The Language Divide Challenge

Training data reality:

- Majority of AI training data is English
- Smaller languages massively underrepresented
- Thai, Portuguese, Arabic content often invisible

Strategy for international hotels:

- Publish content in English AND local language
- English = gateway language into AI systems
- Don't abandon local audience, ensure AI accessibility

Example: Bangkok hotel with only Thai content may be invisible to AI travel planners

Infrastructure Limits Affecting Discovery

The brutal economics of AI:

- Training GPT-4: $78-100 million in compute costs
- Massive energy requirements (nuclear power investments)
- Limited crawling resources

Impact on hotels:

- Not every site crawled equally
- **Higher-value content prioritized**
- Infrastructure now effectively a ranking factor

Implication: Premium properties with better content get more AI visibility

The New Hotel Discovery Funnel - Data-Driven Insights

Traditional funnel: Awareness → Consideration → Booking (across multiple sites)

New AI-powered funnel (based on our training data observations):

1. **Foundation Training** - Is your hotel in Common Crawl's dataset?
   - **43% English bias** - international hotels need English content
   - **6-18 month data lag** - AI knows your old information
2. **Fine-Tuning** - Do travel-specific AI models know about you?
   - Hotels with structured Q&A content get better representation
   - Generic models less accurate than specialized travel AI
3. **Real-Time Retrieval** - Can AI find current info about you?
   - Live web searches supplement training knowledge
   - Fresh content gets prioritized in AI responses
4. **Conversion** - Can travelers actually book?
   - **Current gap:** AI provides recommendations but booking still requires multiple steps

The data insight: Hotels succeed by optimizing for ALL layers, not just traditional SEO

What Hotels Should Focus On - Based on Data Patterns

Priority 1: Foundation Visibility (The Basic Requirement)

- Ensure AI crawlers can access your site (check robots.txt, CDN settings)
- Monitor for AI bot traffic: CCBot, GPTBot, ClaudeBot
- **Data insight:** Hotels that block crawlers become invisible to AI training

Priority 2: Content Strategy (The Quality Factor)

- Create English content even if you serve local markets (43% training data bias)
- Develop FAQ-style content that AI can easily understand
- **Observation:** Hotels with structured Q&A content get more accurate AI responses

Priority 3: Fresh Information (The Currency Problem)

- Keep website content current - training data lags 6-18 months
- Regular content updates help with real-time AI retrieval
- **Key finding:** Stale training data means AI needs fresh web content to supplement

Monitor your AI presence: Test what ChatGPT, Claude, and Perplexity say about your property

COMMON
CRAWL

What Hotels Must Do Now (Continued)

4. Monitor the Political Landscape

- Track AI crawler policies
- Watch for default setting changes
- Stay informed on legal/regulatory shifts

5. Educate Your Teams

- Revenue managers need to understand AI visibility
- Marketing teams must think beyond traditional SEO
- Executive leadership should grasp strategic implications

6. Test Your AI Visibility

- Ask ChatGPT, Claude, Perplexity about hotels in your market
- See if your property appears in recommendations
- Monitor competitor visibility

The Strategic Reality

SEO hasn't died - it's evolved:

- **Old world:** Index and rank
- **New world:** Train and retrieve

For hotels:

- Traditional SEO still matters (for now)
- **AIO (AI Optimization) is becoming critical**
- Training data = new link authority
- Crawl accessibility = new technical SEO

The window is closing - establish AI presence before competitors

Real-World Observations from Common Crawl

What we're seeing in hotel visibility patterns:

- **Discovery paradox:** Some world-class properties invisible to AI because of CDN settings
- **Language gap:** International hotels with only local-language content missing from AI recommendations
- **Freshness problem:** AI knows about hotels that closed years ago, misses new openings

The Children's Hospital case study:

- Children's Hospital of Los Angeles - one of America's top pediatric cancer hospitals
- Parents searching "where should I take my child with leukemia in LA?" don't see CHLA
- **Root cause:** Hospital's CDN blocks AI crawlers by default
- **Hotel implication:** Technical decisions now affect guest discovery

The broader pattern:

- CDN configurations have become a revenue factor
- Default settings often block AI training and retrieval
- Many organizations don't realize they're invisible to AI systems

Early adopter insight: While others accidentally opt out, forward-thinking properties optimize for AI visibility

COMMON CRAWL

The Path Forward - Working with AI Trainers

Based on Common Crawl's partnerships with AI companies:

What major AI companies want from hotels:

- **Provenance:** Clear data licensing and ownership
- **Completeness:** Comprehensive property information
- **Accuracy:** Quality-tested, up-to-date content
- **Trust:** Your data more reliable than random web scraping

The Path Forward - Working with AI Trainers

Two paths to get into AI systems:

Direct partnerships: Work with specific AI travel companies

- "Alice and Bob's Travel Chatbot wants hotel data"
- Custom Q&A datasets about your properties
- Direct revenue sharing or licensing deals

Clearing-house approach: Make data available in industry datasets

- **This is how you get into big models:** OpenAI, Anthropic, Google, Microsoft
- Industry-wide standards and datasets
- Broader reach but less control

COMMON CRAWL

Simple Action Items - Focus on the Fundamentals

This week: Basic health check

- Test what ChatGPT and Claude say about your hotel
- Check if your website allows AI crawlers (look for CCBot, GPTBot in server logs)
- Review robots.txt file and CDN settings

This month: Content audit

- Ensure you have English content even if you serve local markets
- Create FAQ-style content that directly answers guest questions
- Update any stale information on your website

This quarter: Strategic monitoring

- Set up regular AI visibility testing
- Monitor competitor AI presence
- Track how AI describes your property vs. how you want to be described

The goal: Make sure AI systems can find you, understand you, and represent you accurately to potential guests

The Bottom Line

The fundamental shift:

"Training data is the new link graph. If you are not in the crawl, you are not in the model. And if you are not in the model, you are not in the market."

For hotels: The travelers of tomorrow are already asking AI assistants where to stay. Will they find you?

The choice: Be part of the discovery revolution, or become invisible to it.

COMMON
CRAWL

Q&A

Questions?

Contact:

- Stephen Burns
- Web Intelligence Lead, Common Crawl Foundation
- burns@commoncrawl.org
- https://www.linkedin.com/in/burnstephen/

Remember: Your hotel's future discoverability depends on decisions you make today about AI accessibility.