# Healthcare Professional Open Web Data Enrichment

**Date**: August 2025  
**Status**: Proof of Concept

## Feature Overview

The Provider Profile Web Enrichment feature automatically gathers and structures comprehensive information about healthcare providers from publicly available web sources. Using AI and advanced search capabilities, the system queries multiple authoritative sources, extracts relevant data points, and provides complete citation tracking for transparency and compliance. This enables Conflixis to maintain enriched provider profiles with verified affiliations and professional relationships sourced directly from official websites and trusted directories.

## Business Objectives

### Primary Goals

1. **Data Completeness**: Achieve comprehensive provider profiles with all available public information
2. **Transparency**: Provide full citation tracking for every data point collected
3. **Accuracy**: Ensure high confidence in extracted information through multi-source validation

### Key Benefits

* **Enhanced Analytics**: More complete provider profiles enable better payment attribution and network analysis
* **Risk Assessment**: Comprehensive affiliation data improves conflict of interest detection
* **Compliance Support**: Documented sources provide audit trail for regulatory requirements
* **Operational Efficiency**: Automated enrichment reduces manual research time by 95%

## Data Sources & Collection Methodology

### How We Gather Provider Information

Our web enrichment system uses AI and advanced search capabilities to systematically query and extract information from publicly available sources. Based on our proof of concept testing, we achieved: - **100% success rate** in finding relevant information - **88.5% average confidence** in data accuracy - **7.3 sources cited** per provider on average

### Primary Sources (Highest Trust - 95%+ Confidence)

#### 1. **Official Hospital & Health System Websites**

* **Example**: Major health system websites, regional hospital networks
* **Data Found**: Current positions, hospital affiliations, specialties, contact information
* **Reliability**: Direct from employing institution, updated regularly
* **Coverage**: 100% of providers tested had official hospital profiles

#### 2. **Government Databases**

* **NPPES (CMS)**: National Provider Identifier registry
* **Data Found**: NPI numbers, practice addresses, taxonomy codes
* **Reliability**: Official government records, legally mandated accuracy
* **Coverage**: 100% for NPI verification

#### 3. **Medical School & University Directories**

* **Example**: Academic medical center websites, medical school faculty directories
* **Data Found**: Academic appointments, teaching roles, department affiliations
* **Reliability**: Verified by academic institutions
* **Coverage**: 70% of providers with academic affiliations

### Secondary Sources (High Trust - 80-94% Confidence)

#### 4. **Professional Medical Directories**

* **Healthgrades**: Provider ratings, hospital affiliations, insurance accepted
* **Vitals.com**: Practice locations, patient reviews, specialties
* **Doximity**: Professional network profiles, peer endorsements
* **U.S. News Health**: Hospital affiliations, recognition awards
* **Data Found**: Comprehensive professional profiles, peer recognition
* **Reliability**: Aggregated from multiple sources, user-verified
* **Coverage**: 90% of providers found in at least one directory

#### 5. **Professional Society Websites**

* **Example**: American College of Cardiology, Society of Vascular Surgery
* **Data Found**: Board certifications, fellowship status, leadership roles
* **Reliability**: Verified membership rosters
* **Coverage**: 60% of specialists found in relevant societies

### Tertiary Sources (Moderate Trust - 70-79% Confidence)

#### 6. **Healthcare News & Publications**

* **Medical journals**: NEJM, JAMA, specialty journals
* **Healthcare news**: Becker’s Hospital Review, Modern Healthcare
* **Data Found**: Research publications, speaking engagements, industry recognition
* **Reliability**: Editorial oversight, fact-checked
* **Coverage**: 40% of senior providers mentioned in news

#### 7. **Conference & Event Websites**

* **Medical conferences**: Speaker lists, panel participation
* **Industry events**: Advisory board meetings, symposiums
* **Data Found**: Speaking engagements, expertise areas, industry relationships
* **Reliability**: Event organizer verification
* **Coverage**: 30% of providers found as speakers/panelists

### Data Extraction Process

1. **Initial Query**: System searches “[Provider Name] + [Institution] + [Specialty] + [NPI]”
2. **Source Discovery**: AI identifies and prioritizes relevant sources
3. **Information Extraction**: Structured data pulled from each source
4. **Citation Capture**: Every data point linked to its source URL
5. **Confidence Scoring**: Each source weighted by credibility factors

### Excluded Sources

* Personal social media accounts (Twitter, Facebook, Instagram)
* Patient review sites without verification (Yelp, Google Reviews)
* Paid advertisement or sponsored content
* Sources behind paywalls or requiring authentication
* Outdated information (>5 years without updates)

## Real-World Data Collection Examples

### Example 1: Vascular Surgery Specialist

**Sources Found**: 5 high-confidence sources - **Health System Website** (100% confidence): Current position as Director of Aortic Therapy - **Vitals.com** (100% confidence): Multiple practice locations - **Specialty Practice Website** (80% confidence): Detailed specialty information - **Result**: Complete professional profile with hospital affiliations across 11 facilities

### Example 2: Cardiology Department Leader

**Sources Found**: 9 citations - **Doximity** (100% confidence): Leadership roles and practice ownership - **US News Health** (100% confidence): Hospital affiliations and rankings - **Healthgrades** (100% confidence): Awards and professional recognitions - **NPIFinder** (80% confidence): NPI verification and subspecialty certifications - **Result**: Rich profile including leadership roles, awards, and board certifications

### Example 3: Plastic Surgery Specialist

**Sources Found**: 12 citations (highest in test group) - **Multiple Hospital Sites**: Confirmed affiliations across health system - **Professional Directories**: Comprehensive practice information - **Medical Society Sites**: Board certifications and specializations - **Result**: Most comprehensive profile with extensive citation trail

## Data Elements Collected

### Core Provider Information

* Full name and credentials
* National Provider Identifier (NPI)
* Medical specialties and subspecialties
* Board certifications
* Contact information (professional only)

### Professional Affiliations

* Current hospital affiliations
* Practice locations and addresses
* Academic appointments
* Previous positions (last 10 years)
* Department leadership roles

### Industry Relationships

* Pharmaceutical company relationships
* Medical device company affiliations
* Healthcare startup involvement
* Consulting arrangements
* Speaking bureau participation

### Research & Academic

* Published research papers
* Clinical trial participation
* Research grants received
* Academic teaching positions
* Medical school faculty appointments

### Leadership & Governance

* Hospital board memberships
* Medical society leadership
* Healthcare company board positions
* Advisory committee roles
* Professional organization offices

## Compliance Considerations

### Legal Compliance

* **Terms of Service**: Adheres to website terms of service
* **Copyright**: Cites sources and respects intellectual property
* **Data Accuracy**: Implements validation to ensure data accuracy
* **Audit Trail**: Maintains complete logs for compliance auditing

### Ethical Considerations

* **Transparency**: All sources are documented and cited
* **Accuracy**: Multiple source validation for critical data points
* **Bias Prevention**: Systematic approach prevents selection bias
* **Update Frequency**: Regular updates to maintain data currency

## Quality Assurance

### Data Quality Metrics

* **Completeness**: Percentage of fields populated per profile
* **Accuracy**: Validation against known authoritative sources
* **Consistency**: Cross-source agreement on key data points
* **Currency**: Age of data and last verification date

### Source Reliability Assessment

Our system evaluates each source based on multiple factors:

#### Credibility Weights by Source Type

* **Hospital Websites**: 100% base confidence
* **Government Databases**: 100% base confidence
* **Professional Directories** (Healthgrades, Doximity): 100% base confidence
* **University Websites**: 95% base confidence
* **Professional Societies**: 90% base confidence
* **Healthcare News**: 85% base confidence
* **Company Websites**: 80% base confidence
* **General Directories**: 70% base confidence

#### Confidence Adjustments

* **Trusted Domain Bonus**: +15% for known authoritative domains
* **Substantial Context**: +5% for detailed information
* **Verification Keywords**: +10% for “verified”, “confirmed”, “official”
* **Uncertainty Penalties**: -20% for “unverified”, “claimed”, “alleged”

#### Final Confidence Categories

* **High (90-100%)**: Official sources, multiple confirmations
* **Medium (70-89%)**: Professional directories, single authoritative source
* **Low (50-69%)**: News mentions, unverified directories
* **Review Required (<70%)**: Triggers manual validation

### Manual Review Triggers

* Profiles with overall confidence <70%
* Conflicting information between sources
* High-value providers (C-suite, department chairs)
* Unusual patterns or outliers detected

### Business Metrics

* Manual research time reduction
* Profile enrichment coverage
* Conflict of interest detection improvement
* Network analysis accuracy improvement

## Detailed Conflict of Interest Use Cases

### 1. Undisclosed Board Memberships & Advisory Roles

* **Scenario**: Provider reports consulting for PharmaCo but fails to disclose board seat at subsidiary
* **Web Data Value**: Discovers board positions via company websites, SEC filings, press releases
* **Application**: Cross-reference self-reported disclosures with web-discovered governance roles
* **Risk Mitigated**: Hidden conflicts where providers influence formulary while on pharma boards

### 2. Academic-Industry Relationship Verification

* **Scenario**: Department chair oversees clinical trials while holding equity in sponsor company
* **Web Data Value**: Identifies research grants, trial sponsorships, academic appointments from university sites
* **Application**: Map research relationships against financial interests
* **Risk Mitigated**: Research bias from undisclosed sponsor relationships ($50,000+ threshold per PHS regulations)

### 3. Competing Healthcare Entity Detection

* **Scenario**: Hospital board member also serves competing health system
* **Web Data Value**: Discovers affiliations with ASCs, specialty hospitals, competitor systems
* **Application**: Identify providers with “insurmountable conflicts” per policy
* **Risk Mitigated**: Strategic conflicts, patient steering, competitive intelligence leaks

### 4. Speaker Bureau & Marketing Relationships

* **Scenario**: Physician promotes specific devices while receiving speaker fees
* **Web Data Value**: Finds speaker listings on conference sites, pharma speaker bureaus, event programs
* **Application**: Match prescribing patterns with speaking engagements
* **Risk Mitigated**: Biased treatment decisions influenced by marketing relationships

### 5. Medical Device Company Ownership

* **Scenario**: Surgeon owns stake in device company used in their procedures
* **Web Data Value**: Uncovers startup involvement, patent holdings, company officer roles
* **Application**: Link device usage data with ownership interests
* **Risk Mitigated**: Self-referral violations, overutilization of owned products

### 6. Family Member Indirect Interests

* **Scenario**: CMO’s spouse serves on board of major hospital vendor
* **Web Data Value**: Identifies family relationships through news articles, social profiles, company bios
* **Application**: Expand conflict detection beyond individual to family network
* **Risk Mitigated**: Vendor selection bias, contract steering ($1M+ vendor threshold)

*This document is proprietary and confidential to Conflixis. Distribution is limited to authorized personnel only.*