Web Scraping Research Task

Purpose

Conduct systematic web scraping and data collection from publicly available sources to gather comprehensive intelligence on companies, campaigns, and market information for advertising research.

Inputs

- Target company or brand name
- Research objectives and focus areas
- Data points to collect
- Source websites and platforms
- Time period for data collection

Process

1. Research Planning

- Define specific data requirements
- Identify target websites and sources
- Determine scraping approach and tools
- Establish ethical guidelines and boundaries
- Plan data organization structure

2. Source Identification

Primary Sources:
- Company official websites
- Investor relations pages
- Press release archives
- Career pages (for company culture insights)

Secondary Sources:
- News websites and publications
- Industry trade journals
- Award databases (Cannes Lions, Effies, Clios, etc.)
- Creative showcases (Ads of the World, Campaign, etc.)

Social Media Sources:
- LinkedIn (company pages, employee profiles)
- Twitter/X (company accounts, mentions)
- Facebook (company pages, ads library)
- Instagram (brand presence, campaigns)
- YouTube (video campaigns, brand channels)

Database Sources:
- Crunchbase (funding, acquisitions)
- Glassdoor (employee reviews, culture)

- G2/Capterra (client reviews for agencies)
- Industry-specific databases

3. Data Collection

Company Information:
- Company overview and history
- Leadership team and key personnel
- Office locations and global presence
- Company size and employee count
- Mission, vision, and values
- Recent news and press releases

Financial Data:
- Revenue figures (if public)
- Growth rates and trends
- Funding rounds and investors
- Acquisitions and mergers
- Market valuation

Client Information:
- Current client roster
- Past client relationships
- Client testimonials and case studies
- Industry distribution of clients
- Notable client wins and losses

Campaign Data:
- Campaign names and descriptions
- Launch dates and duration
- Campaign objectives and strategies
- Creative executions and assets
- Media channels used
- Performance metrics (if available)
- Awards and recognition

Social Media Metrics:
- Follower counts across platforms
- Engagement rates (likes, comments, shares)
- Content frequency and types
- Audience demographics
- Sentiment analysis
- Influencer partnerships

Competitive Intelligence:
- Market positioning statements
- Unique selling propositions
- Service offerings and capabilities
- Pricing information (if available)
- Partnership and alliance announcements

4. Data Validation

- Cross-reference information across multiple sources
- Verify dates and factual accuracy
- Check for conflicting information
- Assess source credibility and recency
- Flag uncertain or unverified data
- Document data quality issues

5. Data Organization

- Structure data according to research objectives
- Create consistent data schemas
- Tag and categorize information
- Maintain source attribution
- Organize by themes and topics
- Prepare for synthesis and analysis

6. Ethical Compliance

- Respect robots.txt and website terms of service
- Avoid overloading servers with requests
- Use publicly available information only
- Respect copyright and intellectual property
- Maintain data privacy standards
- Document data collection methodology

Tools and Techniques

Web Scraping Tools

- Browser developer tools for manual inspection
- Python libraries (BeautifulSoup, Scrapy, Selenium)
- Web scraping services (when appropriate)
- API access (when available)
- RSS feeds and news aggregators

Data Collection Methods

- Manual research and documentation
- Automated scraping scripts
- API integration
- Social media monitoring tools
- News monitoring services
- Database queries

Data Storage

- Structured databases (for quantitative data)
- Document repositories (for qualitative data)
- Spreadsheets (for organized data sets)
- Note-taking systems (for observations)

- Cloud storage (for collaboration)

Outputs

- Structured data sets organized by category
- Source-attributed information
- Data quality assessments
- Research methodology documentation
- Raw data files for further analysis
- Initial observations and patterns

Success Criteria

- Comprehensive data collection across all focus areas
- High data quality and accuracy
- Proper source attribution
- Ethical compliance maintained
- Data organized for easy synthesis
- Gaps and limitations documented
- Methodology clearly documented

Quality Checks

- [ ] All required data points collected
- [ ] Information verified across multiple sources
- [ ] Sources credible and current
- [ ] Data properly organized and structured
- [ ] Source attribution complete
- [ ] Ethical guidelines followed
- [ ] Data quality issues flagged
- [ ] Methodology documented
- [ ] Ready for synthesis phase

Related

- @agent:company-research-agent
- @agent:client-portfolio-agent
- @template:company-profile-tmpl
- @checklist:advertising-research-quality-checklist
- @data:data-source-directory

Best Practices

- Start with official sources for accuracy
- Use multiple sources to validate information
- Document methodology for reproducibility

- Respect website terms of service
- Maintain organized research files
- Update data regularly as new information emerges
- Flag areas requiring human expertise
- Collaborate with other agents to avoid duplication
- Focus on publicly available information only
- Maintain objectivity and avoid bias