

2010年到2013年期间，我们主要为澳大利亚(Webscraping)和美国(Vitals)公司提供网站数据采集外包服务，先后成功完成了超过600个项目，涉及网站上千个。这是我们一个重要的成长阶段，为公司累积下来了丰富的Web分析和数据提取实战经验，尤其是在应对各种网站反采集策略上。下面简单介绍一些这些项目：

根据项目内容，大致可以分为如下三种类型：

1. 数据定制抓取。客户最终需要的只是数据。约占70%。
2. 采集器（脚本）定制。客户最终需要的是采集脚本。约占25%。
3. Web API开发。客户最终需要的是一个Web接口。约占5%。

根据项目难度大致可以分为如下四种类型：

1. 简单。网站无防护策略，内容静态展示。约占30%。
2. 中等。网站无防护策略，内容动态加载，例如Ajax技术。约占40%。
3. 较难。网站有封IP策略。约占20%。
4. 非常难。网站封IP很厉害、数据有加密。约占10%。

这些采集项目的开发语言均为Python，运行平台80%为Linux平台（Ubuntu为主）、20%为Windows平台。数据输出格式80%为CSV，10%为JSON和XML，10%为MySQL和SQL Server。

#### 附：部分项目列表。

careerbuilder	laptopown	superpages_lawyers	century21_agents	manta_keywords	merchantcircle
archive	wikipedia_surnames	telelistas	foreign_social_networks	funeral_homes	more_belgian_netw
amazon_best sellers	yahoo_search_results	ethical_junction_directory	lawyer_database	facebook_network	property_owners
mattress_scrapers	britstore	worldcat_api	google_company_rankings	businessdeutschland	busrates
renalsite.com	planningportal	androidzoom	stylesays	uspto	facebook_events
trademarkia	nhl	linkedin_ni	nmls	google_groups	opencorporates
product_importer	facebook_comments	blitz	skattelister_no	facebook_members	merging_books
pinterest	skoreit	google_maps_business_details	genes	amazon_products	bloggers.com
hotels.com	starwoodhostels	upc_product_details	price_comparison	network_emails	eurekalet
fuddruckers	tjsp	tumblr	bluefin_api	android_reviews	facebook_ad_cour
australia_directory	australian_store_locations	yellowpages_keywords	twitter_accounts_stats	tumblr_followers_app	yellowpages.ca
wellsfargo	email_activity	peapod	truequity	flex-craft_images	beats_by_dre
appfigures	resumator	article_counts	yell_uk_professions	mobileroadie	googleplus
wordpress_activity	daft_contact_advertiser	safeway.com	mediapost	product_api	bebe.com
discogs_music	allybank	youtube_insights	cabinetmakers_shopfitters	yellowpages_au_profession	wp_australia
neffco	run_the_elements_download	blog_profiles	facebook_activity	bazaavoice	belgian_65K_rankir
blitz_facebook_events	challenges	charlotterusse.com	linkedin_singapore	number-lab	pinterest_new_form
safecross	tjsp_parsing	youtube_views	zocdoc	linkedin_company_urls	instagram_profiles
garden_directories	usa_names	hotfrog_au	android_pit	manta_listings	hair_cut_stores
esaj	facebook_groups	facebook_ad_count_automation	twitter_messages	nike_facebook_graph	facebook_compan
austlii	recipe_scraper	data_job_listings	australian_retailers	retailer_scrapes	business_google_d
facebook_event_tracking	android_apps	linkedin_custom_search	restaurant_menus	facebook_image_script	crimson_hexagon
itunes_apps	android_iphone_apps	linkedin_search_tool	databazaar	yelp.com	tpb_gov_au
semanticparsing	yelp_be	getlisted	youtube_keyword_tool	newpages	blitz_linkedin
linkedin_companies	PHP_scraping	Klout_API	mobile_app_search	doctoralia	cellartracker
wine_searcher	youngamercards	logogarden	statigr.am	instagram_followers	blitz_twitter_report
webtrends_reports	yahoo_directory	linkedin_connections	blitz_twitter_followers	blitz_foursquare	fsbonline
blitz_weibo	facebook_interests	vivino	eat24hours	themeparks	facebook_user_ids
hotel	linkedin_names	manufacturing_parts	blitz_facebook_movies	amazon_reviews_statistics	german_products
blitz_social_influence	individual_company_profiles	facebook_user_ids_for_pages	mortgage_brokers	belgium_websites_2014	large_business_ma
dashba.com_hooks	amazon_us_products	business_directories	belandia_east_somerset	belandia_company_analyser	amazon_products



**西安鲲之鹏网络信息技术有限公司**

**选择我们，所有数据都是你的！**



公司名称：西安鲲之鹏网络信息技术有限公司

网 址：<http://www.site-digger.com/>

地 址：陕西省西安市雁塔区长安中路99号长安文化综合大厦11821室

邮 编：710061

联系电话：029 - 87553281

手 机：13571845363 齐工

13389148466 周工

客服 QQ: 1649677458 或 312602670

邮箱: [hello@site-digger.com](mailto:hello@site-digger.com)

---