项目总结2023.6

爬取时间为6月5-9日, 分析时间为6月12-13 日

涉及单位

目前爬取并分析了阵地类型为官方网站、微信公众号两种类型的数据，其中官网网站69个（成功爬取63个），微信公众号102个。
其中官方网站爬取了同域名下所有链接地址，微信公众号爬取了历史所有文章。
共爬取网页数量27876页，公众号文章4153篇

分析结果

通过对爬取结果进行分析并与标准文档比对，分别在27876页网页中发现错误100处，在4153篇公众号中发现错误33处，具体见结果表

存在问题

目前存在部分网站因反爬措施或无法访问或技术原因，未获取到数据，见下表

单位	可能原因
中国建筑材料科学研究总院有限公司_http://www.cbma.com	网址错误
对比服务平台_http://www.ctc-online.cn/companyLogin?company	网站需登录
中国建材检验认证集团江苏有限公司_http://www.ctcjs.com	不能访问
乌鲁木齐京诚检测技术有限公司_http://www.wlmqjc.cn/	网站域名过期
中材江西电瓷电气有限公司_http://www.sinoma-insulator.com	不能访问
中国新型建材设计研究院有限公司_http://www.cnhdi.com/	不能访问