采购中心爬虫需求处理流程规范

文件版本历史

文件版本	修订日期	修订人	审核人	批准人	修订说明
V1.0	2024/09/04	Lennon	Miang	Joel	初版
V1.1	2024/11/27	Lennon	Miang	Joel	明确细节

目录

为提高爬虫需求处理的效率，确保信息传递的准确性及工作流程的规范性，现制定如下流程规范。请采购中心各事业部参照执行，并给予配合。

一、需求整理与提交

1. 各事业部自行爬取：
   - 对于已提供完整代码的简单网站，且事业部内已有爬虫负责人（如原数据部、船长办或其他有爬虫经验的同事），**原则上由各事业部自行爬取**。
   - 如在操作过程中遇到技术问题，可联系 Lennon 提供技术支持。
2. 邮件提交需求：
   - 对于其他较为复杂的网站爬虫需求，请各位自行将需求整理清晰，并将其以 Excel 文件形式作为附件，**发送至 Lennon，同时抄送本部门部长及 Joel**。

二、邮件与附件格式要求

1. 邮件主题：请确保邮件主题包含“【爬虫需求】+部门名称+简要描述”，例如：“【爬虫需求】项目投放部 - Steering Damper”。
2. 邮件正文：
   - 需求描述：简要描述爬虫任务目标及所需数据。
   - 时间要求：明确需求的完成时间或紧急程度。若无紧急要求，**建议预留至少一周时间**；如为紧急需求，**请详细说明项目进度，以便合理安排优先级**。
   - 其他说明：包括文件输出样式、文件体积等特殊要求。
3. Excel 附件格式要求：
   - 附件中的 Excel 文件应简洁明了，包含以下内容：第一列为需求序号；第二列为目标网址。

	A	B
1	No	Url
2	1	https://www.dormanproducts.com/gsearch.aspx?type=keyword&origin=keyword&q=clip
3	2	https://www.dormanproducts.com/gsearch.aspx?type=keyword&origin=keyword&q=insert

   - 请确保每个需求都有明确的网址，避免“爬一下 Dorman 家的卡扣”这种含糊不清的描述。**网址应直接指向需要抓取的具体页面，避免仅提供网站主页地址**。

三、需求处理流程

1. 需求处理：

- Lennon 收到需求后会进行初步评估，并在必要时进一步沟通并细化需求。
- 一旦需求确认，数据部将根据优先级安排任务，并尽力在规定时间内完成。
- 爬虫任务完成后，数据部将通过邮件反馈爬取的结果。

2. 查重处理：

- 如无特殊要求，**请各事业部自行完成查重工作**。
- 若有特殊查重需求，请在邮件正文中详细说明，并将邮件抄送 Nolan。
- 请确保查重需求明确，包括但不限于：需要与哪个品类进行数据库查重、表内重复项是否保留以及相应的保留规则等，**避免如“查重、去重”这种不明确的表述**。

四、其他事项

- 提交需求时，请确保信息准确无误，避免因描述不清导致的处理延误。
- 数据部将对每项爬虫需求进行记录，并定期向采购中心领导汇报进展情况。
- 如有紧急需求或特殊事项，请提前告知 Lennon，以便及时协调处理。