数据清洗工具包

一份可直接照做的清单：去重、合并、标准化与质量检查（示例 PDF）

生成日期：2026-06-06

目标：把“能看”的数据变成“能用”的数据——让筛选、分群、自动化与报表更可靠。

1) 去重：从“相同”到“相似”

先选定你要去重的对象（人 / 公司 / 线索），再确定唯一键与匹配策略。建议从保守开始，逐步放宽。

- 唯一键优先：邮箱、域名、手机号等（若可靠，优先用它）。
- 模糊匹配：名称相似度（忽略大小写、空格、标点、常见后缀）。
- 语音匹配：应对拼写差异（John/Jon），作为补充信号。
- 合并策略：保留最新值 / 最长值 / 来源可信度更高者；冲突值可拼接到备注字段。

2) 合并多表：用“唯一键”建立主列表

多来源数据合并时，最怕的是“合并后又产生重复”。请先统一唯一键，并明确字段覆盖规则。

- 选择一个 稳定且可复用 的唯一键：邮箱/域名/系统 ID（优先）→ 其次是 name+site 组合键。
- 先对各来源做 字段标准化（电话/URL/国家地区/职位），再进行 join。
- 字段覆盖建议：空值才覆盖（先补全）→ 再按“可信来源优先级”覆盖。
- 为每次合并保留 变更日志：来源、时间、覆盖字段、冲突字段。

3) 标准化模板：最常见的 6 个字段

下面这些字段标准化后，你的筛选、去重与自动化会立刻变得更稳定。

- 邮箱：trim 空格 → 全部转小写 → 去除不可见字符。
- 域名/URL：去掉协议/尾部斜杠 → 统一 www 规则 → punycode/大小写规范。
- 电话：去空格与分隔符 → 统一国家码 → 缺省国家码补全（谨慎）。
- 职位：统一大小写 → 映射同义词（VP ≈ Vice President）。
- 国家/地区：ISO 代码或标准中文名二选一（避免混用）。
- 公司名称：移除后缀（Inc/LLC/有限公司）→ 处理特殊符号（& / · / -）。

4) 质量检查：上线前 60 秒自检

- 重复率是否异常上升？（按唯一键统计）
- 关键字段缺失率是否下降？（邮箱/域名/电话）
- 字段类型是否正确？（日期/数字/布尔）
- 采样 20 条记录人工核对：是否合并错人/错公司？
- 是否能用同一套规则 重复执行？（可复用性）

提示：这份 PDF 作为落地页示例下载内容，你可以替换为自己的真实资料（品牌、流程、截图等）。