# Data Management Plan Report

## 1. Datasets Overview

为实现营养信息的高效管理与智能推荐，本项目使用了多个公开数据源：

### 1.1 Open Data Sources

**1.1.1 RecipeNLG – A dataset of cooking recipes (CSV)**

* 来源: [https://recipenlg.cs.put.poznan.pl/](https://recipenlg.cs.put.poznan.pl/?utm_source=chatgpt.com)
* 内容: ~2M 条食谱，字段包括 title, ingredients, directions, source, link, NER
* 更新频率: 静态数据集，不定期更新
* 许可: 公开研究用途

**1.1.2 Open Nutrition (TSV → CSV)**

* 来源: [https://www.opennutrition.app](https://www.opennutrition.app?utm_source=chatgpt.com)
* 内容: ~300k 食材营养数据，字段包括 id, name, alternate\_names, description, nutrition\_100g, ingredients, labels
* 更新频率: 定期社区更新
* 许可: 开放许可，研究使用

**1.1.3 Nutritional intake recommendations (CSV/JSON)**

* 来源: 国际营养机构与公开研究标准
* 内容: 不同年龄、性别、饮食习惯的推荐摄入值
* 用途: 与食谱数据结合，计算合理膳食比例

**1.1.4 Health Mentor Dataset (JSON)**

* 来源: [https://github.com/Hannan2004/AI-Health-Mentor-Dataset/blob/main/nutrition.json](https://github.com/Hannan2004/AI-Health-Mentor-Dataset/blob/main/nutrition.json?utm_source=chatgpt.com)
* 内容: 健康饮食与膳食相关元数据
* 用途: 用作 nutrition-match-api 的辅助参考

## 2. Iteration 1

### 2.1 Data Usage

* **Recipes API**: 提供食谱查询、筛选、分页功能。
* **Foods API**: 支持根据食材名称/前缀搜索对应的营养数据。
* **Nutrition-Match API**: 输入一组食材及数量，返回标准化匹配结果与营养总量（per\_recipe / per\_serving）。
* 用户可通过 RESTful API 访问所有接口，前端应用基于结果进行展示与分析。

### 2.2 Data Preparation (Cleaning & Wrangling)

* **RecipeNLG**: 增加唯一主键 recipe\_id，统一 UTF-8 编码。
* **Open Nutrition**: 转换为 CSV，标准化字段格式。
* **index.js 数据处理**:
  + 清洗、去重与格式化字段；
  + 增加输入校验（必填项检查、参数类型检测）；
  + 错误处理与日志输出，方便调试。
* **nutrition-match-api**:
  + 进行小写化、去空格、同义词映射与模糊匹配；
  + 按克数加权汇总营养值；
  + 输出 standardized JSON 响应。

### 2.3 Data Storage

* **Amazon S3**: 存放原始 RecipeNLG / OpenNutrition 数据集 (CSV/TSV/JSON)，作为 DynamoDB 导入源及备份。
* **DynamoDB**:
  + **Recipes\_v2** 表: 存储 ~2M 条食谱数据，按 recipe\_id 主键检索；
  + **Foods** 表: 存储 ~300k 食材营养信息，支持前缀搜索与模糊匹配。
* **备份机制**: 定期启用 DynamoDB PITR（Point-in-Time Recovery），并保留原始 S3 文件。

### 2.4 Database Design

* **Recipes\_v2**
  + 主键: recipe\_id (String)
  + GSI: gsi\_title\_prefix (title 前缀索引)
  + 字段: recipe\_id, title, title\_lc, title\_lc\_first1, ingredients, directions, source, link, NER
* **Foods**
  + 主键: id (String)
  + GSI: gsi\_name\_prefix (食材前缀搜索)
  + 字段: id, name, name\_lc, name\_lc\_first1, alternate\_names, description, nutrition\_100g, labels, serving
* **API 交互**
  + GET /recipes?recipe\_id=...
  + GET /recipes?title\_prefix=...
  + GET /foods?id=...
  + GET /foods?name\_prefix=...
  + POST /match （nutrition-match-api）

### 2.5 Data Analytics

#### 2.5.1 Hindsight Insights

* 使用 RecipeNLG 与 Open Nutrition 历史数据，生成饮食习惯与健康标签 (vegan, gluten\_free, low\_sugar)。
* 分析用户常用食材的营养分布，识别潜在健康风险。

#### 2.5.2 Foresight Insights

* nutrition-match-api 实现个性化营养预测：
  + 提前估算每日营养摄入与不足；
  + 提供膳食组合优化建议；
  + 模拟未来饮食趋势，辅助用户制定饮食计划。

#### 2.5.3 Insights

* API 的标准化输出结构便于前端应用、研究人员和营养师直接利用。
* 结合推荐摄入标准与个人输入参数，系统能提供个性化饮食建议。