**Title: RAG Test Corpus — ACME Manufacturing (Synthetic)  
Doc-ID: RAG-TEST-ACME-001  
Version: 1.0.0  
Last-Updated: 2025-10-22  
Classification: Public / Synthetic (no real PII)  
Languages: vi, en, zh-TW**

**0) Mục đích**

Bộ tài liệu tổng hợp *giả lập* để kiểm thử hệ thống RAG: đánh giá **chunking**, **metadata filtering**, **hybrid search**, **reranking**, **context compression**, và **citation**. Bao gồm nhiều kiểu dữ liệu (bảng, JSON, CSV, mã nguồn, quy trình, FAQ, từ điển thuật ngữ đa ngôn ngữ).

Gợi ý cắt đoạn (chunking): chunk\_size=800~1200 tokens, chunk\_overlap=100~150. Gán metadata: {doc\_id: "RAG-TEST-ACME-001", section, lang, tags}.

**1) Tổng quan công ty (vi)**

**ACME Manufacturing** là nhà sản xuất pin công nghiệp. Các xưởng chính: **BenLuc**, **LongAn**. Hệ thống nội bộ gồm ERP (*Oracle TIPTOP*), hệ đo lường (*Hioki/Keyence*), và phần mềm SPC nội bộ (*SPC\_CAN\_ACID*, *BatteryMonitor*).

* Mã khách hàng chính: CUST-SEA-001 (APAC), CUST-EU-017 (EU).
* Ca làm việc: **Ca1** (07:00–19:00), **Ca2** (19:00–07:00 hôm sau).
* Kỳ chốt bảo dưỡng: **ngày 03 hàng tháng**.

**Quy tắc điểm thi đua bảo dưỡng (rút gọn):**

* Nếu sau *ngày 03* mà đơn vị **chưa nhập** ngày *dự tính bảo dưỡng* tháng trước → bị trừ điểm trong tháng sau.
* Chỉ tính **số ngày trễ dài nhất** trong các khoảng trễ (tránh đếm trùng ngày).

Câu hỏi mẫu: *"Khi nào bị trừ điểm nếu quên nhập dự tính bảo dưỡng?"*

**2) Danh mục sản phẩm (Product Catalog)**

| **SKU** | **Tên** | **Dải V (V)** | **Dải R (mΩ)** | **Ngưỡng Cảnh Báo** | **Xưởng** |
| --- | --- | --- | --- | --- | --- |
| **BAT-12V-100Ah** | Bình 12V 100Ah | 11.8–12.8 | 3.5–7.0 | V<11.8 hoặc R>7.0 | BenLuc |
| **BAT-48V-50Ah** | Bình 48V 50Ah | 47.5–54.5 | 12.0–18.0 | V<47.5 hoặc R>18.0 | LongAn |
| **BAT-24V-80Ah** | Bình 24V 80Ah | 23.6–27.4 | 6.0–12.0 | V<23.6 hoặc R>12.0 | BenLuc |

Lưu ý: **BAT-12V-100Ah** có ngưỡng R **7.0 mΩ**, còn **BAT-24V-80Ah** là **12.0 mΩ** (dễ nhầm khi truy vấn).

**3) SOP Bảo dưỡng (Maintenance SOP)**

**Mã quy trình:** SOP-MNT-ACID-2025

1. Nhập *ngày dự tính bảo dưỡng* trước **23:59 ngày 03** mỗi tháng.
2. Với **Ca2** (qua 00:00), ngày báo cáo tính cho **ngày hôm trước**.
3. **Audit**: lưu *Audit Log* gồm {user\_id, role, timestamp, action, record\_id}.
4. **RBAC**: Vai trò tech, supervisor, auditor.

**Ví dụ tính ngày Ca2**:

* Thời điểm: 2025-10-21 00:00:00, Ca2 → quy chiếu **ngày 2025-10-20**.

**4) ERP & Mã Nghiệp vụ (TIPTOP)**

| **Mã** | **Diễn giải** | **Ghi chú** |
| --- | --- | --- |
| tc\_sth\_file | Bảng lịch bảo dưỡng | Khóa: tc\_sth001 |
| tc\_tao\_file | Lệnh sản xuất | Chuỗi liên kết workorder |
| cglt200 | Phiếu nhập | Tồn kho |
| axci100 | BOM | Định mức |

**Khóa chính mẫu:** TC\_STH\_001 = BL-2025-10-001.

**5) Danh bạ liên hệ nội bộ**

| **Bộ phận** | **Họ tên** | **Email** | **Điện thoại** |
| --- | --- | --- | --- |
| QA | **Nguyen Thi Lan** | [qa.lan@acme.example](mailto:qa.lan@acme.example) | +84 28 3899 0001 |
| IT/ERP | **Tran Minh Khoa** | [it.khoa@acme.example](mailto:it.khoa@acme.example) | +84 28 3899 0002 |
| Sản xuất | **Le Hoang** | [prod.hoang@acme.example](mailto:prod.hoang@acme.example) | +84 28 3899 0003 |

Email là *giả lập*, dùng để test *regex* và lọc metadata.

**6) Chính sách & Ngoại lệ**

* Chính sách POL-MNT-001: nếu **trễ** nhập ngày dự tính sau mốc chốt (**03 hàng tháng**), ghi nhận **vi phạm**.
* Ngoại lệ: khi hệ thống *downtime* > 2 giờ, cho phép nhập bù kèm *ticket* INC-YYYYMM-####.

**Ví dụ ticket:** INC-202510-0421 (downtime 2h30 ngày 2025-10-03).

**7) FAQ (vi/en)**

**Q1 (vi):** *Qua 0h Ca2 thì ngày tính là ngày nào?*  
**A1:** Ngày **hôm trước** theo SOP (xem §3).

**Q2 (en):** *When is the maintenance plan considered late?*  
**A2:** After **23:59 on the 3rd** of the month (see §1 & §6).

**Q3 (vi):** *Ngưỡng R tối đa của BAT-12V-100Ah?*  
**A3:** **7.0 mΩ** (xem §2).

**8) Phụ lục dữ liệu có cấu trúc**

**8.1 JSON API mẫu**

{

"doc\_id": "RAG-TEST-ACME-001",

"endpoint": "/api/v1/plan/maintenance",

"fields": ["workorder", "planned\_date", "shift", "factory"],

"example": {

"workorder": "WO-BL-241015-001",

"planned\_date": "2025-10-20",

"shift": "2",

"factory": "BenLuc"

}

}

**8.2 CSV records**

workorder,factory,shift,datetime,v,r,sku

WO-BL-241015-001,BenLuc,1,2025-10-20 08:15:00,12.45,4.2,BAT-12V-100Ah

WO-BL-241015-002,BenLuc,2,2025-10-21 00:05:00,12.10,6.9,BAT-12V-100Ah

WO-LA-241015-003,LongAn,2,2025-10-21 01:10:00,47.8,17.5,BAT-48V-50Ah

**8.3 Mã nguồn (C#) — quy tắc Ca2 qua 0h**

if (shift == "2" && time.TimeOfDay >= TimeSpan.Zero)

{

time = time.AddDays(-1); // quy chiếu về ngày hôm trước

}

**8.4 Công thức & ký hiệu**

* Ký hiệu SPC: ( \bar{X} ), (\sigma), **UCL/LCL**.
* Quy đổi mΩ → Ω: mΩ / 1000.

**9) Thuật ngữ đa ngôn ngữ (vi/en/zh-TW)**

* **dự tính bảo dưỡng** / maintenance plan / 維護預計
* **chốt (ngày 03)** / monthly cutoff (day 03) / 月結（每月3號）
* **Ca2** / Night shift / 夜班
* **điểm thi đua** / performance score / 績效分

**10) Trường hợp dễ gây nhầm lẫn (for RAG stress test)**

1. **BAT-12V-100Ah** và **BAT-24V-80Ah** có **ngưỡng R khác nhau** (7.0 vs 12.0 mΩ).
2. **Cutoff** là ngày **03**, không phải ngày 01.
3. Bản ghi **Ca2 00:05** ngày 2025-10-21 quy chiếu **2025-10-20**.

**11) Truy vấn kiểm thử gợi ý**

* "Qua 0h Ca2 thì lấy ngày nào để chấm điểm? Trích dẫn điều khoản."
* "Ngưỡng điện trở tối đa của BAT-12V-100Ah là bao nhiêu và nguồn ở đâu?"
* "Khi nào bị tính trễ nhập ngày dự tính bảo dưỡng?"
* "Liệt kê email liên hệ QA và IT."
* "Đếm số bản ghi Ca2 trong CSV và cho biết SKU."
* "Sự khác nhau giữa ngưỡng R của 12V-100Ah và 24V-80Ah?"

**12) Metadata gợi ý**

parent\_doc: RAG-TEST-ACME-001

sections:

- id: s1

title: Tong quan cong ty

tags: [overview, vi]

- id: s2

title: Product Catalog

tags: [table, thresholds]

- id: s3

title: SOP Bao duong

tags: [policy, shift]

- id: s8

title: Structured Data

tags: [json, csv, code]

**13) Điều khoản đánh giá (Evaluation Notes)**

* **Grounding/Citation**: bắt buộc dẫn § mục chứa số liệu.
* **Reranker** cần ưu tiên bảng §2 cho câu hỏi về ngưỡng.
* **Normalization** ngày Ca2: test bộ biến đổi truy vấn (query transform) theo SOP §3.

Đây là tập dữ liệu *synthetic* cho mục đích thử nghiệm, có thể mở rộng bằng phụ lục riêng.