

CÔNG CỤ IMPORT HÀNG LOẠT 1200 BÀI BÁO

Ngày: 9 Tháng 12, 2025

MỤC ĐÍCH

Tạo công cụ import **1200+ bài báo cũ** từ các năm trước vào hệ thống, bao gồm:

- Bài báo đã xuất bản** (PUBLISHED) - Hiển thị công khai
- Bài báo không được duyệt** (REJECTED) - Chỉ admin/editor/tác giả xem
- Upload PDF files** lên AWS S3
- Tự động tạo** tác giả, danh mục, số tạp chí nếu chưa có
- Phân quyền bảo mật** - Bài REJECTED không công khai

THÔNG TIN VỀ DỮ LIỆU

BẠN CÓ GÌ:

- 1200 file PDF
- Thông tin tác giả, tên bài báo trong PDF
- Bìa số tạp chí, tóm tắt trong PDF

CÁCH THỨC:

KHÔNG tự động trích xuất từ PDF (khó, dễ sai)

NÊN điền vào Excel một lần (chính xác 100%)

BẠN PHẢI:

- Điền thông tin vào Excel theo template
- Đặt PDF vào folder `pdf-imports/`
- Chạy script import

CÁC FILE ĐÃ TẠO

1. Script Import Chính

 `scripts/import-articles-from-excel.ts` (350+ dòng)

Chức năng:

- Đọc Excel với 22 cột thông tin
- Match với file PDF theo tên
- Upload PDF lên S3 tự động
- Tự động tạo User (tác giả)

- Tự động tạo Category (danh mục)
- Tự động tạo Issue (số tạp chí)
- Tạo Submission + Article
- Báo cáo kết quả chi tiết

2. Template Excel

 scripts/IMPORT_TEMPLATE.xlsx

Cấu trúc:

- **Sheet 1:** Template chính với 3 dòng dữ liệu mẫu
- **Sheet 2:** Hướng dẫn chi tiết

22 cột quan trọng:

Số	Cột	Bắt buộc	Ví dụ
1	Mã bài báo	✓	BB-2020-001
2	Tiêu đề (VN)	✓	Ứng dụng AI trong y tế
3	Tác giả	✓	Nguyễn Văn A
4	Email tác giả	✓	nguyen-vana@example.com
5	Đơn vị	✓	Đại học Quốc gia Hà Nội
6	Tóm tắt (VN)	✓	Nghiên cứu này tập trung vào...
7	Từ khóa	✓	AI, Machine Learning, Healthcare
8	Danh mục	✓	Công nghệ thông tin
9	Năm xuất bản	✓	2020
10	Số tạp chí	✓	1
11	Tên file PDF	✓	article-001.pdf
12	Trạng thái	✓	PUBLISHED hoặc REJECTED

+ 10 cột khác: Tiêu đề EN, Tóm tắt EN, Tập tạp chí, Trang số, DOI, Ghi chú...

3. Script Tạo Template

 scripts/generate-import-template.ts

Chạy để tạo template:

```
yarn tsx scripts/generate-import-template.ts
```

4. Hướng Dẫn Chi Tiết

 scripts/IMPORT_GUIDE.md (200+ dòng)

Nội dung bao gồm:

- Các bước thực hiện
- Bảng mô tả 22 cột
- Cách chuẩn bị Excel và PDF
- Cách chạy script
- Kiểm tra kết quả
- Xử lý lỗi
- Phân quyền bảo mật
- Thống kê
- Checklist

CÁCH SỬ DỤNG (NHANH)

Bước 1: Tạo Template

```
cd /home/ubuntu/tapchi-hcqs/nextjs_space
yarn tsx scripts/generate-import-template.ts
```

Bước 2: Điền Dữ Liệu

1. Mở file scripts/IMPORT_TEMPLATE.xlsx
2. Điền thông tin 1200 bài báo
3. Lưu lại với tên articles-import.xlsx

Bước 3: Chuẩn Bị PDF

```
cd /home/ubuntu/tapchi-hcqs/nextjs_space/scripts
mkdir -p pdf-imports
# Copy tất cả PDF vào folder này
cp /path/to/your/pdfs/*.pdf ./pdf-imports/
```

Bước 4: Chạy Import

```
cd /home/ubuntu/tapchi-hcqs/nextjs_space
yarn tsx scripts/import-articles-from-excel.ts scripts/articles-import.xlsx
```

BẢO MẬT & PHÂN QUYỀN

Bài PUBLISHED (Đã xuất bản)

PUBLIC - Mọi người xem được:

- Hiển thị trên trang chủ
- Hiển thị trong /archive
- Hiển thị trong /articles
- Có thể search và download PDF

Bài REJECTED (Không duyệt)

PRIVATE - Chỉ những người sau xem được:

1. **SYSADMIN** - Quản trị hệ thống
2. **EIC** - Tổng biên tập
3. **MANAGING_EDITOR** - Biên tập điều hành
4. **SECTION_EDITOR** - Biên tập chuyên mục
5. **TÁC GIẢ** - Chỉ xem bài của chính họ

Không hiển thị trong search công khai, Archive, danh sách bài báo

Có thể tra cứu trong Admin Dashboard

✨ TÍNH NĂNG NỔI BẬT

1. Tự động tạo dữ liệu liên quan

User (Tác giả):

- Tìm theo email
- Nếu chưa có → Tạo mới với role AUTHOR

Category (Danh mục):

- Tìm theo tên hoặc code
- Nếu chưa có → Tạo mới

Volume & Issue (Số tạp chí):

- Tìm theo năm + số
- Nếu chưa có → Tạo mới

2. Upload PDF lên S3

Validation file tồn tại, size, type

Tự động generate S3 key duy nhất

Format: articles/timestamp-code-filename.pdf

3. Tạo Article cho bài PUBLISHED

Nếu status = PUBLISHED:

- Tạo Article record
- Gắn vào Issue tương ứng
- Lưu PDF cloud path
- Set approvalStatus = APPROVED

4. Báo cáo chi tiết

- Real-time progress cho từng bài
 - Summary cuối cùng (success/failed)
 - Chi tiết từng lỗi (dòng + message)
-

⚠ XÓ LÝ LỖI THƯỜNG GẶP

Lỗi 1: File PDF không tồn tại

Nguyên nhân: Tên file trong Excel không khớp

Giải pháp: Kiểm tra tên file, phân biệt HOA/thường

Lỗi 2: Email không hợp lệ

Nguyên nhân: Sai format email

Giải pháp: Sửa email trong Excel

Lỗi 3: Upload S3 thất bại

Nguyên nhân: Thiếu cấu hình AWS

Giải pháp: Kiểm tra `.env` file

```
AWS_ACCESS_KEY_ID=your-key
AWS_SECRET_ACCESS_KEY=your-secret
AWS_REGION=ap-southeast-1
AWS_BUCKET_NAME=your-bucket
```

THỐNG KÊ SAU IMPORT

Truy vấn thống kê:

```
-- Tổng số bài
SELECT
  COUNT(*) FILTER (WHERE status = 'PUBLISHED') AS published,
  COUNT(*) FILTER (WHERE status = 'REJECTED') AS rejected,
  COUNT(*) AS total
FROM "Submission";

-- Thống kê theo danh mục
SELECT
  c.name,
  COUNT(*) AS total_articles,
  SUM(CASE WHEN s.status = 'PUBLISHED' THEN 1 ELSE 0 END) AS published
FROM "Submission" s
JOIN "Category" c ON s."categoryId" = c.id
GROUP BY c.name
ORDER BY total_articles DESC;
```



MẸO HỮU ÍCH

1. Test với số lượng nhỏ trước

Trước khi import 1200 bài, hãy test với 10-20 bài:

1. Tạo file Excel với 10-20 dòng
2. Copy 10-20 PDF vào `pdf-imports/`
3. Chạy import
4. Kiểm tra kết quả
5. Nếu OK, import toàn bộ

2. Backup database trước khi import

```
cd /home/ubuntu/tapchi-hcqs/nextjs_space
bash scripts/backup-db.sh
```

3. Đặt tên file PDF có quy tắc

Ví dụ:

- `2020-01-001.pdf` (Năm-Số-STT)
- `BB-2020-001.pdf` (Prefix-Năm-STT)

4. Sử dụng Excel Formula

Tự động tạo mã bài báo:

```
= "BB-" & L2 & "-" & TEXT(A2, "000")
// Kết quả: BB-2020-001
```

Tự động tạo tên file PDF:

```
= B2 & ".pdf"
// Kết quả: BB-2020-001.pdf
```



PERFORMANCE

Ước tính thời gian:

- Đọc Excel: ~5-10 giây
- Upload 1 PDF (2-5MB): ~2-3 giây
- Tạo records: ~0.5 giây/bài

Tổng: 1200 bài × 3 giây = **~60 phút** (1 giờ)

SOS FILE STRUCTURE

```

scripts/
├── import-articles-from-excel.ts      ← Script chính
├── generate-import-template.ts        ← Tạo template
├── IMPORT_TEMPLATE.xlsx              ← Template mẫu
├── IMPORT_GUIDE.md                  ← Hướng dẫn
├── articles-import.xlsx             ← File của bạn
└── pdf-imports/
    ├── article-001.pdf            ← Folder PDF
    ├── article-002.pdf
    ...

```

KẾT LUẬN

Đã hoàn thành:

1.  Script import đầy đủ tính năng (350+ dòng)
2.  Template Excel chuẩn, dễ sử dụng
3.  Script generate template tự động
4.  Hướng dẫn chi tiết (200+ dòng)
5.  Phân quyền bảo mật PUBLISHED/REJECTED
6.  Tự động tạo user, category, issue
7.  Upload PDF lên S3 tự động
8.  Báo cáo chi tiết success/failed
9.  Xử lý lỗi graceful
10.  Documentation đầy đủ

Checklist cuối cùng:

- [x] Tạo script import
- [x] Tạo template Excel
- [x] Tạo hướng dẫn sử dụng
- [x] Test validation
- [x] Xử lý PUBLISHED/REJECTED
- [x] Documentation đầy đủ
- [x] Error handling
- [x] Progress reporting
- [x] Ready for 1200 articles!

HỖ TRỢ

Các file quan trọng:

-  scripts/import-articles-from-excel.ts - Script chính
-  scripts/IMPORT_TEMPLATE.xlsx - Template Excel
-  scripts/IMPORT_GUIDE.md - Hướng dẫn chi tiết
-  scripts/generate-import-template.ts - Tạo template

 CHÚC BẠN IMPORT THÀNH CÔNG 1200 BÀI BÁO!

Build Status:  Ready to use

Created: December 9, 2025

Author: DeepAgent - Abacus.AI