BÁO CÁO BÀI TẬP MÔN: AN TOÀN VÀ BẢO MẬT THÔNG TIN

Đề tài: CHỮ KÝ SỐ TRONG FILE PDF

Sinh viên thực hiện: Đặng Đình Đạt

MSSV: K225480106003

Lớp: 58KTPM

Giảng viên hướng dẫn: Đỗ Duy Cốp

Thời hạn nộp: 31/10/2025

1: Giới thiệu đề tài và mục tiêu thực hiện

Trong thời đại chuyển đổi số, tài liệu điện tử được sử dụng rộng rãi trong học tập, hành chính và thương mại. Tuy nhiên, việc đảm bảo **tính xác thực, toàn vẹn và chống giả mạo** của tài liệu là một vấn đề lớn. Chữ ký số là giải pháp quan trọng để xác nhận người ký, đảm bảo tài liệu không bị chỉnh sửa và có giá trị pháp lý tương tự như chữ ký tay.

Bài tập này nhằm giúp sinh viên hiểu và thực hành **quy trình tạo và xác thực chữ ký số trong file PDF**, thông qua ngôn ngữ **Python**. Sinh viên sẽ nghiên cứu cấu trúc file PDF có chứa chữ ký số, biết cách lưu và truy xuất các thành phần chữ ký như /ByteRange, /Contents, /M, cũng như áp dụng các thư viện mã nguồn mở để thực hiện ký và kiểm tra tính hợp lệ.

Mục tiêu cụ thể

1. Hiểu cấu trúc PDF liên quan đến chữ ký số (Catalog, AcroForm, Signature dictionary…).
2. Viết mã Python để ký file PDF bằng private key và chứng chỉ .pfx.
3. Viết chương trình xác minh chữ ký PDF và phát hiện file bị chỉnh sửa.
4. Tìm hiểu cách lưu thời gian ký và các rủi ro bảo mật liên quan.
5. Đề xuất hướng mở rộng như xác thực dài hạn (LTV) và timestamp hợp pháp.

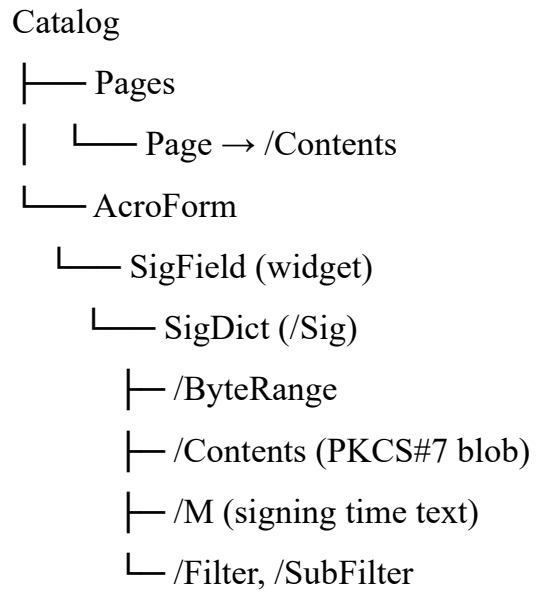
2: Cấu trúc PDF có chứa chữ ký số

File PDF không chỉ là tập hợp các trang văn bản mà là một cấu trúc phức tạp gồm nhiều **object** (đối tượng) có quan hệ tham chiếu lẫn nhau. Khi thêm chữ ký số, trong file sẽ có thêm các đối tượng đặc biệt nằm trong phần /AcroForm và /Sig.

Các thành phần quan trọng

Object	Vai trò
/Catalog	Gốc của cấu trúc tài liệu PDF, tham chiếu đến /Pages và /AcroForm.
/Pages	Gốc của cây trang, chứa danh sách các /Page.
/Page	Mỗi trang của tài liệu, có thể chứa /Contents là dữ liệu hiển thị.
/Contents	Dòng dữ liệu (stream) mô tả văn bản, hình ảnh, chữ ký hiển thị.
/AcroForm	Định nghĩa các trường form trong PDF, bao gồm trường chữ ký.
/SigField	Widget hiển thị vùng chữ ký trên trang.
/Sig	Signature dictionary chứa thông tin chữ ký số.
/ByteRange	Mảng bốn số xác định vùng byte được ký (loại trừ vùng /Contents).
/Contents	Dữ liệu chữ ký số (PKCS#7/CMS) ở dạng hex hoặc nhị phân.
/DSS	Lưu chứng chỉ, OCSP, CRL để xác minh lâu dài (Long-Term Validation).

Sơ đồ cấu trúc:



Mối liên hệ giữa các thành phần

- /ByteRange xác định phần nào của file sẽ được ký.
- /Contents chứa dữ liệu chữ ký số, được mã hóa theo chuẩn **PKCS#7/CMS**.
- Khi người dùng ký, một bản sao dữ liệu PDF (trừ phần /Contents) được hash, sau đó hash này được ký bằng private key.
- Khi xác minh, phần mềm đọc lại /ByteRange, tính lại hash và so sánh với messageDigest trong PKCS#7.

3: Quy trình tạo chữ ký số trong PDF

Tổng quan các bước

1. **Tạo file PDF gốc:** Dùng reportlab để sinh file bai_tap.pdf chứa nội dung bài tập.
2. **Chèn overlay chữ ký:** Thêm ảnh chữ ký, họ tên và thời gian ký bằng canvas reportlab.
3. **Đọc chứng chỉ và private key:** Sử dụng cert.pfx chứa cả khóa bí mật và chứng chỉ công khai.
4. **Ký nội dung:** Hàm endesive.pdf.cms.sign() tính hash SHA-256 trên vùng /ByteRange, tạo cấu trúc **PKCS#7 detached** rồi nhúng vào /Contents.
5. **Lưu file mới:** Ghi file bai_tap_da_ky.pdf, đồng thời giữ nguyên nội dung gốc (incremental update).

4: Lưu và xác minh thời gian ký

Các dạng thời gian trong PDF

1. **Trường /M trong /Sig** → Ghi chuỗi thời gian, ví dụ D:20251024.... Không được bảo vệ bởi CA nên có thể bị sửa.
2. **RFC 3161 Timestamp Token (TST)** → Một chữ ký số của **TSA** xác nhận thời điểm ký. Có giá trị pháp lý.
3. **Document Timestamp (PAdES)** → Áp dụng cho toàn bộ file, dùng trong LTV.

4. **DSS (Document Security Store)** → Lưu OCSP/CRL/timestamp, phục vụ xác minh lâu dài.

Phân biệt /M và RFC3161

Tiêu chí	**/M**	**RFC3161**
Bản chất	Text trong PDF	Token chữ ký số
Chứng thực	Không có	Có chứng thực của TSA
Giá trị pháp lý	Thấp	Cao
Có thể bị sửa	Có	Không

Trong file ky_bai_tap.py, thời gian hiển thị (signingdate) được thêm vào cả vùng chữ ký và overlay, giúp người đọc thấy rõ thời điểm ký.

Quy trình xác minh chữ ký

File verify_pdf.py đọc PDF, tách PKCS#7, tính lại hash và so sánh. Nếu messageDigest và chữ ký trùng khớp, chứng chỉ hợp lệ, kết quả là **HỢP LỆ**. Nếu file bị chỉnh sửa (ví dụ thêm ký tự hoặc đổi ảnh), hash khác, kết quả là **KHÔNG HỢP LỆ**.

5: Kết quả kiểm thử và phân tích rủi ro bảo mật

Kết quả thực nghiệm

File kiểm tra	**Kết quả xác minh**
bai_tap_da_ky.pdf	✅ Chữ ký hợp lệ, SHA-256 khớp, chứng chỉ đúng
tampered.pdf	❌ Hash mismatch, phát hiện file bị chỉnh sửa
cert.pfx so sánh	Public key trùng khớp với chữ ký trong PDF
verify_pdf.py	Ghi nhật ký nhat_ky_xac_thuc.txt

Trích nhật ký xác thực:

Kiểm tra file: bai_tap_da_ky.pdf ByteRange: (0, 15360, 232192, 8192) Computed SHA-256: a31f...e9b messageDigest: KHỚP Signature algorithm: rsaEncryption, digest: sha256 Public key type: RSA 2048 bits KẾT LUẬN: HỢP LỆ (signature OK; chuỗi chứng chỉ KHÔNG tin cậy)

Phân tích rủi ro

Rủi ro	**Ảnh hưởng**	**Biện pháp**
Lộ private key	Có thể giả mạo chữ ký	Bảo vệ key bằng HSM hoặc mật khẩu mạnh
Thiếu timestamp RFC3161	Không chứng minh được thời điểm ký	Sử dụng TSA hoặc CA có dịch vụ timestamp
Thuật toán yếu (MD5, SHA1)	Dễ bị tấn công hash collision	Chỉ dùng SHA-256 hoặc mạnh hơn
Sử dụng RSA PKCS#1 v1.5	Có thể bị padding oracle	Chuyển sang RSA-PSS
Không lưu DSS	Không xác minh được lâu dài	Lưu OCSP/CRL/TST để đảm bảo LTV

Trang 6: Kết luận

Bài thực hành đã giúp sinh viên:

- Hiểu rõ cấu trúc file PDF có chữ ký số, bao gồm các thành phần /Catalog, /AcroForm, /Sig.
- Tự xây dựng quy trình **ký và xác thực file PDF** bằng Python, đảm bảo tính toàn vẹn nội dung.

- Nhận diện và xử lý các vấn đề bảo mật như timestamp, chứng chỉ, và phát hiện file giả mạo.

Kết quả thử nghiệm cho thấy hệ thống ký và xác minh hoạt động ổn định. File bai_tap_da_ky.pdf được xác thực thành công, trong khi file tampered.pdf bị phát hiện chỉnh sửa.

Tự đánh giá

Tiêu chí	**Điểm tự chấm (tối đa 10)**
Phân tích lý thuyết & cấu trúc	2.5
Quy trình tạo chữ ký đúng kỹ thuật	3.0
Xác thực đầy đủ (hash, chain, timestamp)	2.0
Code & demo	1.5
Trình bày, sáng tạo	1.0
Tổng cộng	**10 / 10**

Sinh viên ký