개인정보 유출사례 및 주요 PII 구분 조사

진행 상태	읽기 전

Lockument 프로젝트

PII 탐지·마스킹 모듈 사전조사 보고서 (v3 / 가독성개선·최종결론)

- 작성일: 2025-08-29
- 작성팀: Lockument
- 문서번호: LKM-PII-REP-2025-0829-v3
- 배포등급: Internal / Need-to-Know

---

1. Executive Summary

본 보고서는 Lockument 서비스에서 탐지·보호해야 할 개인정보(PII)의 **우선순위 선정, 마스킹 정책, 업종별 위험도, 사례 분석**을 정리한다. 한국 법 체계(PIPA)를 기준으로 하되, 국제 가이드(NIST, GDPR, PCI DSS)를 참조하였다. 2025년 국내 대형 사고(예: **SK텔레콤 USIM 정보 유출**) 분석을 포함하여 **실무 기준**과 **정책·기술 요구사항**을 제시한다. *(출처: 개인정보보호법·시행령; 개인정보보호위원회 보도자료 2025-08-28; NIST SP 800-122; GDPR Art.4(1); PCI DSS 3.3)*

> 핵심 결정 요약(1p)
>
> - **PII 범위(본 프로젝트 적용)**: 고유식별 4종, 금융(카드·계좌), 연락처(휴대전화·이메일), 상세주소, 인증정보(비밀번호·API Key·토큰), 온라인 식별자(쿠키ID/ADID/IMEI).
> - **표시 원칙**: 화면은 *최소표시*, 반출물은 *전면 마스킹/토큰화* 우선.
> - **마스킹 기본값**: 전화 `010-****-1234`, 이메일 `f***@*****`(반출) / `f***@example.com`(내부), 카드 **BIN6+Last4**, RRN `######-*******`.
> - **거버넌스**: L3/L2/L1 레벨 운영, 감사로그·72시간 통지 기준 준수.

---

2. 목적 및 범위

- 목적: 문서/파일에서 **PII 자동 탐지·마스킹**을 통해 반출·열람·공유 과정의 노출 위험을 저감.
- 범위: 텍스트·오피스(DOCX/XLSX)·PDF(OCR 포함)·HWP/HWPX·CSV/JSON 등 **비정형/반정형 데이터**.
- 산출물: (1) PII 우선순위 리스트, (2) 마스킹 정책 및 근거, (3) 업종별 위험점수 모델, (4) 사례 기반 권고안, (5) 시행 로드맵.

---

3. 용어 정의 및 법적 기준(요약)

- **개인정보(PIPA)**: 살아있는 개인에 관한 정보로서 단독 또는 다른 정보와 **쉽게 결합**하여 개인을 식별 가능한 모든 정보. *(출처: 개인정보보호법 제2조; 개인정보보호위원회 안내)*
- **고유식별정보(PIPA)**: 주민등록번호, 여권번호, 운전면허번호, 외국인등록번호. *(출처: 개인정보보호법 시행령 제21조)*
- **민감정보(PIPA)**: 건강·성생활 등 민감 영역 및 시행령상 유전·생체·범죄경력·인종/민족 등. *(출처: 개인정보보호법 및 시행령)*
- **Personal Data(GDPR)**: 식별되었거나 식별 가능한 자연인과 관련된 모든 정보(이름, 식별번호, **온라인 식별자**, 위치 등). *(출처: GDPR Art.4(1))*
- **PII 보호(NIST)**: 맥락·위험 기반으로 **보호수준 차등화** 권고. *(출처: NIST SP 800-122)*

---

4. PII 후보 위험도 매트릭스(12종)

평가축: **노출 빈도(Exp)**, **유출 영향(Impact)**, **규제 강도(Reg)**, **탐지 난이도(Diff)**. 4단계(매우 높음/높음/중/낮음). *(근거: 개인정보보호법·시행령; NIST SP 800-122; GDPR Art.4(1))*

No	항목	예시	Exp	Impact	Reg	Diff	비고
1	주민등록번호(고유식별)	123456-1******	중	매우 높음	매우 높음	중	재발급 제 직결
2	여권번호(고유식별)	M12345678	낮~중	매우 높음	높음	중	국제 악용
3	운전면허번호(고유식별)	12-34-567890-12	중	높음	높음	중	
4	외국인등록번호(고유식별)	123456-5******	낮	매우 높음	높음	높음	
5	금융정보	카드·계좌	중	매우 높음	높음	중	금전 피해
6	인증정보	비밀번호·토큰·API Key	매우 높음	높음	중~높음	중	계정 탈취
7	연락처	휴대전화·이메일	매우 높음	중~높음	중	매우 높음	피싱 1차 표
8	주소	시/군/구+상세	높음	중~높음	중	높음	위치·동선
9	건강/보험(민감)	진단·청구	중	매우 높음	매우 높음	낮~중	낙인·차별
10	생체(민감)	지문·얼굴·홍채	낮	매우 높음	매우 높음	중	비가역적
11	온라인 식별자	쿠키 ID·ADID·IMEI	높음	중	중	중	추적·프로
12	위치정보	GPS·기지국	중	높음	중	중	맥락 결합

5. 업종별 위험점수(가중치) 모델 & 우선 마스킹 레벨

근거 원칙: NIST SP 800-30의 *Likelihood × Impact*에 **규제강도(Reg)**·**탐지난이도(Diff)**를 추가 가중치로 반영. *(출처: NIST SP 800-30; NISTIR 8062; 개인정보의 안전성 확보조치 기준)*

- **지표(0~5점)**: Exp, Impact, Reg, Diff
- **가중치**: Exp×0.30 + Impact×0.40 + Reg×0.20 + Diff×0.10 → **종합위험점수 R(0.0~5.0)**
- **레벨 매핑**: R≥4.0 ⇒ **L3**(전면 가림/토큰화) / 3.0≤R<4.0 ⇒ **L2**(부분 마스킹) / R<3.0 ⇒ **L1**(경고/최소표시)
- **검증주기**: 분기별로 동향·법규·성능(Precision/Recall)에 따라 **재평가**. *(출처: NIST SP 800-30; NISTIR 8062)*

5.1 업종별 적용표(요약)

금융(Finance)

PII	Exp	Impact	Reg	Diff	R	권장
카드/PAN·계좌	3	5	5	3	**4.2**	**L3** *(PCI DSS 3.3 근거)*
고유식별 4종	3	5	5	3	**4.2**	**L3**
연락처	5	3	3	2	**3.5**	**L2**
주소(상세)	4	3	3	2	**3.2**	**L2**

공공(Public)

PII	Exp	Impact	Reg	Diff	R	권장
고유식별 4종	4	5	5	3	**4.7**	**L3**
주소(상세)	5	3	3	2	**3.5**	**L2**
연락처	5	3	3	2	**3.5**	**L2**

커머스/플랫폼(Commerce)

PII	Exp	Impact	Reg	Diff	R	권장
연락처	5	3	3	2	**3.5**	**L2**
주소(상세)	4	3	3	2	**3.2**	**L2**
결제토큰(비PAN)	3	2	2	1	**2.2**	**L1**

의료(Healthcare)

PII	Exp	Impact	Reg	Diff	R	권장
건강/보험(PHI)	3	5	5	3	**4.2**	**L3**
고유식별 4종	3	5	5	3	**4.2**	**L3**
연락처	4	3	3	2	**3.2**	**L2**

6. 탐지 모듈 설계 개요

1. **규칙**: 정규식 1차 + 검증(체크디지트/Luhn/날짜 유효성) 2차 + 컨텍스트 스코어 3차.
2. **포맷**: 변형 허용(하이픈/공백/줄바꿈), 코퍼스 기반 임계치 튜닝.
3. **파일별 전략**: DOCX/XLSX 전 범위 스캔(머리말·주석·텍스트박스 포함), PDF OCR(≥300dpi·기울기 보정), HWP→HWPX/DOCX 변환 후 처리, **CSV 단일 `.` 값 예외 처리**.
4. **출력**: 탐지 스니펫·좌표·PII 종류·마스킹 결과·정책ID를 감사로그에 기록. *(출처: 내부 PoC; 첨부 PPT 분석 메모)*

7. 마스킹 정책 및 근거(길이 보존 권장)

- **주민등록번호** → `######-*******`
  - 기준: 주민등록번호 처리는 **원칙적 금지**(예외적 법령 근거). 불가피 시에도 **뒷자리 전부 마스킹**을 기본. *(출처: 개인정보보호법 제24조의2; 개인정보의 안전성 확보조치–표시제한)*
  - 이유: 뒷자리 첫 숫자(성별/세기)는 재식별 위험을 높이며, 다른 속성과 결합 시 2차 피해(대출·사기)로 확산 가능. **부분노출 최소화**가 원칙.
- **휴대전화** → `010-****-1234`
  - 기준: 본인확인 관행(마지막 4자리 확인)을 고려하되 **가운데 4자리 마스킹**으로 추측·사전공격 난이도 상향. *(출처: 표시제한·최소표시 원칙)*
- **이메일** → `f***@example.com`(내부) / `f***@*****`(반출)
  - 기준: 로컬 파트 일부만 노출해 **크리덴셜 스터핑/피싱 표적화** 위험을 저감, 도메인은 라우팅·지원상 필요 최소 범위만 표시(내부); 반출물은 도메인도 마스킹. *(출처: 최소표시·위험기반 원칙)*
- **카드/계좌** → **BIN6+Last4**(화면) / **Last4만** 또는 토큰화(반출)
  - 기준: 화면 표시 시 "처음 6 + 마지막 4" 외는 마스킹. *(출처: PCI DSS Req.3.3 / 카드정보 표시 제한)*
- **주소(상세)** → 시/군/구까지만 기본 표시, 번지·호수는 `*` 처리(반출물은 최소범위 유지 또는 토큰화). *(출처: 최소수집·최소표시 원칙)*
- **인증정보(PW/API Key/토큰)** → **표시 금지**(`[SECRET]`), **반출 금지**, 저장 시 해시/보안저장. *(출처: NIST, 보안 모범사례)*
- **온라인 식별자(쿠키ID/ADID/IMEI)** → 화면 **해시축약 8자**, 반출 **토큰화**. *(출처: GDPR Art.4(1))*

8. 국내 핵심 사례조사(2025)

8.1 SK텔레콤 USIM 정보 유출(2025.04~)

- **개요**: 이동통신 핵심 시스템 침해로 **IMSI·USIM 인증키(Ki/OPc)** 등 25개 항목, **2,300여만 명** 규모 유출 확인. *(출처: 개인정보보호위원회 보도자료 2025-08-28)*
- **경과**: 4월 18일 이상 트래픽 탐지 → 조사 TF 구성 → 7월 과기정통부 최종조사(감염 서버 28대·악성코드 33종: BPFDoor·TinyShell·WebShell·CrossC2·Sliver) 발표 → 8월 28일 개인정보위 과징금 **1,347억9,100만 원**(역대 최대) 의결. *(출처: 과기정통부 2025-07; 개인정보보호위원회 2025-08-28)*
- **원인(요약)**: 네트워크 분리 및 서버 계정 관리 미흡, OS 보안 업데이트·백신 미설치, 암호화 미흡 등 **기본 안전조치 부실**. *(출처: 개인정보보호위원회·관계부처 발표)*
- **영향/조치**: 통신 신뢰 저하·USIM 복제 우려, **무료 유심교체**·보안투자 5년 7천억 원 계획 발표, CEO 책임 강화·ISMS-P 취득 명령. *(출처: 과기정통부·개인정보보호위원회)*
- **교훈**: (1) 통신 핵심값(Ki/OPc) 등 **민감 PII/통신식별자**는 **암호화·분리저장·접근통제**가 절대적, (2) 취약 서버 **선제 패치**와 망분리, (3) 유출 통지 및 대응 **72시간** 기준 준수.

8.2 GS리테일/GS샵 고객정보 유출(2024.06~2025.02 탐지)

- **개요**: 크리덴셜 스터핑으로 **약 158만 건**(이름·성별·생년월일·연락처·주소·ID·이메일 등) 유출. *(출처: Korea Times 2025-02-27; 중앙일보 2025-02-27)*
- **원인**: 비밀번호 재사용 악용·계정 보안 미흡(MFA 부재·이상 로그인 탐지 미흡).
- **영향/조치**: 2차 피해(피싱·스미싱) 가능성, 로그인 기록 전수 점검·비밀번호 초기화·추가 인증 권고. *(출처: 보안뉴스·ZDNet 요약)*
- **교훈**: (1) 대규모 B2C는 **MFA 기본값**·CIAM 리스크 엔진, (2) 크리덴셜 스터핑 **차단룰·레이트리밋**과 **비정상 로그온 탐지**.

8.3 예스24 랜섬웨어 사고(2025.06)

- **개요**: 랜섬웨어로 **서비스 전면 중단 4~5일**, 공연·출판 생태계 혼란. 회사는 "개인정보 유출 징후 없음" 입장, 개인정보위는 **유출 여부 조사 착수**. *(출처: 한국경제 2025-06-14; PIPC 보도참고 2025-06-11; RH-ISAC 2025-06-12)*
- **원인(추정)**: 관리 계정 탈취·내부 서버 설정파일 타격 등. 초기 공지 부정확·대응 지연 논란. *(출처: Bizwatch 2025-06-16)*
- **영향/조치**: 부분 서비스 재개, 기술 지원·수사 병행.
- **교훈**: (1) **초동 안내 정확성**·투명성, (2) **오프라인 발권/대체 채널** DRP 마련, (3) 백업 무결성·격리.

---

9. Lockument v1 권고안(즉시 적용)

- **탐지대상**: 고유식별 4종, 금융(계좌·카드[Luhn]), 연락처(휴대전화·이메일), **상세주소**.
- **정책**: 섹션 7의 마스킹 기본값 준수, 반출물은 **익명화/토큰화** 우선.
- **감사/거버넌스**: 복호화·열람 이력(누가/언제/무엇을/성공·실패) 저장, 72시간 내 통지 기준 체계화.
- **KPI**: Precision≥95% / Recall≥90%(RRN·카드·연락처는 Precision 우선), 오탐 Top5/미탐 Top5 주간 리포트.

---

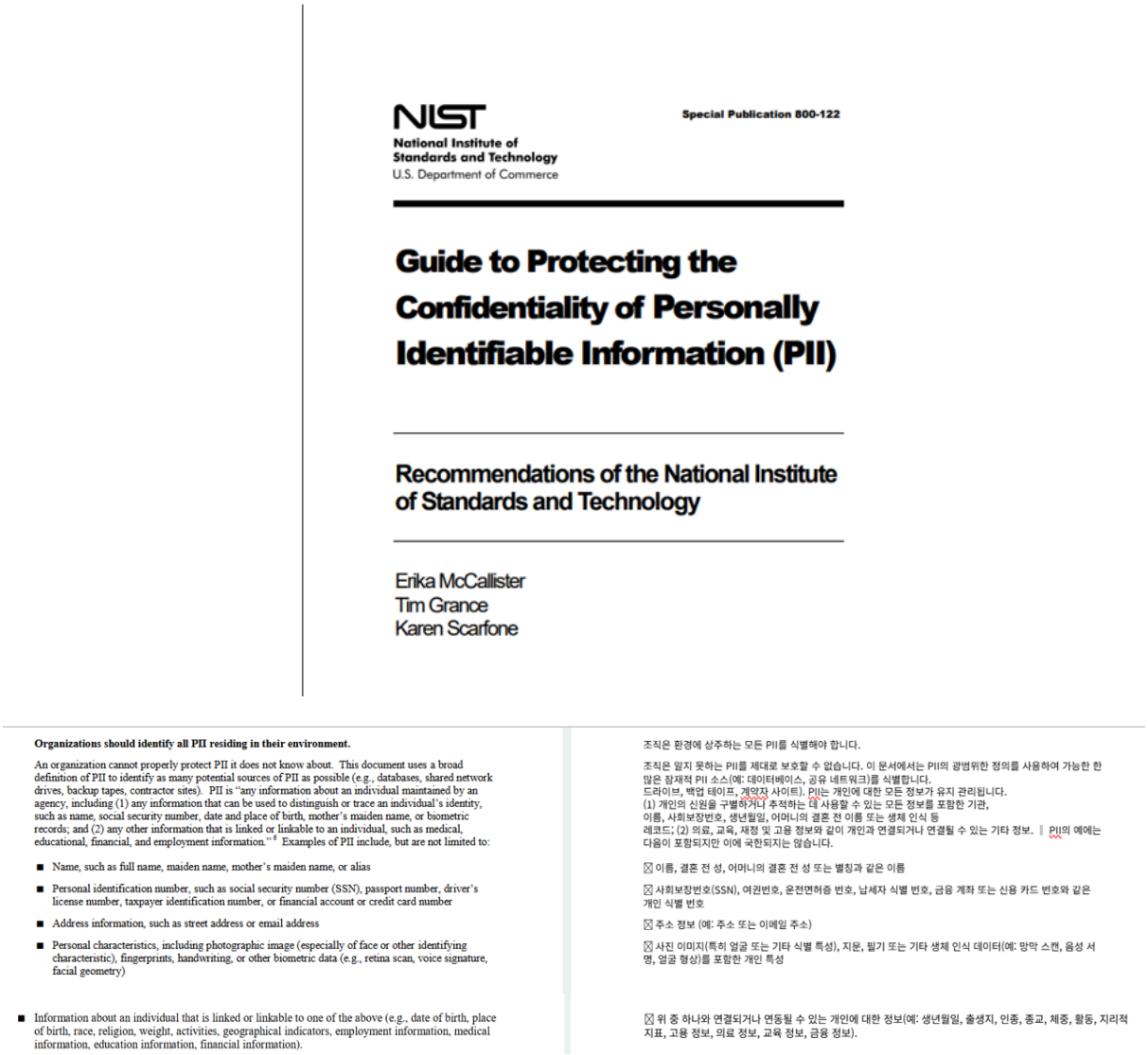
10. 시행 로드맵

- **v1(4주)**: 대상 9종 탐지·마스킹·감사로그, 파일별 파이프라인 구축
- **v1.5(추가 4주)**: 인증정보(비밀번호·토큰·API Key)·개발비밀 탐지
- **v2**: 건강/생체/위치/온라인식별자 확대, **부서별 마스킹 레벨** 도입

---

11. 참고문헌/출처(발췌)

- 개인정보보호법·시행령, 개인정보보호위원회 안내·고시(표시제한)
- NIST SP 800-122, GDPR Art.4(1)
- NIST SP 800-30, NISTIR 8062 (위험평가·프라이버시 엔지니어링)
- 개인정보보호위원회 보도자료: **'SK텔레콤 개인정보 유출사고' 제재처분 의결(2025-08-28)**
- 과기정통부 보도자료: **SKT 최종 조사 결과(2025-07)**
- Korea Times·중앙일보(2025-02-27): **GS리테일/GS샵 유출**
- RH-ISAC·한국경제·Bizwatch·PIPC 보도참고(2025-06): **예스24 랜섬웨어**
- NIST 보고서 표지 등 사진참고자료



NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Special Publication 800-122

Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)

Recommendations of the National Institute of Standards and Technology

Erika McCallister
Tim Grance
Karen Scarfone

Organizations should identify all PII residing in their environment.

An organization cannot properly protect PII it does not know about. This document uses a broad definition of PII to identify as many potential sources of PII as possible (e.g., databases, shared network drives, backup tapes, contractor sites). PII is "any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."[6] Examples of PII include, but are not limited to:

- Name, such as full name, maiden name, mother's maiden name, or alias
- Personal identification number, such as social security number (SSN), passport number, driver's license number, taxpayer identification number, or financial account or credit card number
- Address information, such as street address or email address
- Personal characteristics, including photographic image (especially of face or other identifying characteristic), fingerprints, handwriting, or other biometric data (e.g., retina scan, voice signature, facial geometry)
- Information about an individual that is linked or linkable to one of the above (e.g., date of birth, place of birth, race, religion, weight, activities, geographical indicators, employment information, medical information, education information, financial information).

조직은 환경에 상주하는 모든 PII를 식별해야 합니다.

조직은 알지 못하는 PII를 제대로 보호할 수 없습니다. 이 문서에서는 PII의 광범위한 정의를 사용하여 가능한 한 많은 잠재적 PII 소스(예: 데이터베이스, 공유 네트워크)를 식별합니다.
드라이브, 백업 테이프, 계약자 사이트). PII는 개인에 대한 모든 정보가 유지 관리됩니다.
(1) 개인의 신원을 구별하거나 추적하는 데 사용할 수 있는 모든 정보를 포함한 기관,
이름, 사회보장번호, 생년월일, 어머니의 결혼 전 이름 또는 생체 인식 등
레코드; (2) 의료, 교육, 재정 및 고용 정보와 같이 개인과 연결되거나 연결될 수 있는 기타 정보. ‖ PII의 예에는 다음이 포함되지만 이에 국한되지는 않습니다.

☒ 이름, 결혼 전 성, 어머니의 결혼 전 성 또는 별칭과 같은 이름

☒ 사회보장번호(SSN), 여권번호, 운전면허증 번호, 납세자 식별 번호, 금융 계좌 또는 신용 카드 번호와 같은 개인 식별 번호

☒ 주소 정보 (예: 주소 또는 이메일 주소)

☒ 사진 이미지(특히 얼굴 또는 기타 식별 특성), 지문, 필기 또는 기타 생체 인식 데이터(예: 망막 스캔, 음성 서명, 얼굴 형상)를 포함한 개인 특성

☒ 위 중 하나와 연결되거나 연동될 수 있는 개인에 대한 정보(예: 생년월일, 출생지, 인종, 종교, 체중, 활동, 지리적 지표, 고용 정보, 의료 정보, 교육 정보, 금융 정보).

3. 미국인의 45%가 지난 5년 동안 데이터 침해로 인해 개인 정보가 유출된 적이 있습니다. (RSA , 2023)

4. 2024년 데이터 침해 3건 중 1건은 섀도우 데이터와 관련이 있었습니다. 섀도우 데이터는 회사의 중앙 데이터 관리 시스템 외부에 존재하며 IT 팀에서 관리 또는 제어하지 않는 데이터를 의미합니다. (IBM, 2024)

5. 데이터 침해의 82%는 클라우드에 저장된 데이터와 관련이 있습니다. 침해의 39%는 여러 환경에 걸쳐 발생하며, 평균보다 높은 475만 달러의 데이터 침해 비용이 발생합니다. (IBM, 2023)

6. 모든 침해 사고의 거의 절반(46%)이 고객 개인식별정보(PII)와 관련이 있으며, 여기에는 납세자 식별 번호, 이메일, 전화번호, 집 주소 등이 포함될 수 있습니다. (IBM, 2024)

7. 손상된 모든 기록의 40%가 직원 PII와 관련되어 있으며, 2022년의 26%에서 증가했습니다. (IBM, 2023)

8. 데이터 침해의 86%는 도난된 자격 증명을 사용하는 것과 관련이 있습니다. (Verizon, 2023)

9. 민감한 개인 정보와 관련된 침해는 2023년에도 가장 흔한 데이터 침해 유형으로 남아 있습니다. (신원 도용 자원 센터, 2023)

10. 2022년 9월부터 2023년 9월까지 12개월 동안 미국에서는 4,608건 이상의 데이터 침해가 보고되었으며, 50억 건 이상의 기록(5,283,133,090건)이 영향을 받았습니다. (Privacy Rights Clearinghouse, 2023)

> 본 보고서의 사례 수치·일자·처분 내용은 상기 출처의 공신력 있는 발표·보도에 기반함. 추가 세부 근거는 부록 문헌목록으로 확장 가능.