J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

비정형 데이터셋 표준포맷 기반 국방 비정형 데이터셋 표준화 방안 제안☆

Proposal of Standardization Plan for Defense Unstructured Datasets based on Unstructured Dataset Standard Format

황 윤 영[1] 손 지 성[2*]
Yun-Young Hwang Jiseong Son

요 약

민간에서뿐 아니라 국방분야에서도 인공지능은 국방의 발전을 위해 꼭 도입되어야 하는 첨단기술로 받아들여지고 있으며, 특히 국방과학기술혁신의 핵심 과제로 인공지능이 선정되고, 데이터의 중요성이 확대되고 있다. 국방은 폐쇄적인 데이터 정책에서 데이터 공유·활성화로 방향을 전환하고 있으며, 국방의 발전을 위해 필요한 양질의 데이터를 확보하기 위한 노력을 기울이고 있다. 특히 AI·빅데이터의 고유한 특성이 반영될 수 있도록 관련 절차 개선 및 대량·양질의 데이터가 충분히 확보된 상태에서 연구개발이 시작될 수 있도록 데이터 확보를 위한 사업예산과 제도 검토를 추진하고 있다. 그러나 국방 차원의 정형데이터 및 비정형 데이터의 표준화·품질 기준 마련이 필요한 상황이나 지금까지 국방은 정형데이터의 표준화·품질 기준을 제안하고 있는 수준으로 이에 대한 보완이 필요하다. 본 논문에서는 국방 인공지능에서 가장 필요한 국방 비정형 데이터셋을 위한 비정형 데이터셋 표준포맷을 제안하고, 이를 바탕으로 국방 비정형 데이터셋 표준화 방안을 제안한다.

☞ 주제어 : 데이터셋, 비정형 데이터셋, 표준화, 표준포맷

ABSTRACT

AI is accepted not only in the private sector but also in the defense sector as a cutting-edge technology that must be introduced for the development of national defense. In particular, artificial intelligence has been selected as a key task in defense science and technology innovation, and the importance of data is increasing. As the national defense department shifts from a closed data policy to data sharing and activation, efforts are being made to secure high-quality data necessary for the development of national defense. In particular, we are promoting a review of the business budget system to secure data so that related procedures can be improved to reflect the unique characteristics of AI and big data, and research and development can begin with sufficient large quantities and high-quality data. However, there is a need to establish standardization and quality standards for structured data and unstructured data at the national defense level, but the defense department is still proposing standardization and quality standards for structured data, so this needs to be supplemented. In this paper, we propose an unstructured data set standard format for defense unstructured data sets, which are most needed in defense artificial intelligence, and based on this, we propose a standardization method for defense unstructured data sets.

☞ keyword : Dataset, Unstructured Dataset, Data Standardization, Data Standard Format

1. 서 론

『국방과학기술혁신기본계획(2023.5)』 [1]에서는 국방데이터 관련하여 ‘(과제1.1) 현존 위협에 대응하는 기술개발 역량 확보’와 ‘(과제2.4) AI·빅데이터 중심의 개발사업을 위한 제도 개선’ 등을 추진 과제로 선정하였다. 특히, 국방데이터 구축 로드맵에 따라 국방연구개발 추진에 필요한 AI 학습용 데이터를 선제적·체계적으로 구축하며 구축한 데이터는 ‘국방 지능형 플랫폼을 통해 통합적으로 관리·활용할 수 있도록 관리·협조 체계 구축 예정이다.

1 Defense CBM+ Agile Team, Korea Institute of S&T, Information, Daejeon, 34141, Korea.
2 Defense CBM+ Agile Team, Korea Institute of S&T, Information, Daejeon, 34141, Korea.
* Corresponding author (jsson@kisti.re.kr)
[Received 11 October 2023, Reviewed 24 October 2023 (R2 04 December 2023), Accepted 15 December 2023]
☆ 이 논문은 2022년 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(KRIT-CT-22-081, 무기체계 CBM+ 특화연구센터)

또한, AI·빅데이터의 고유한 특성이 반영될 수 있도록 관련 절차 개선을 검토 중이며, 대량·양질의 데이터가 충분히 확보된 상태에서 연구개발이 시작될 수 있도록 데이터 확보를 위한 사업예산과 제도 검토를 추진하고 있다.

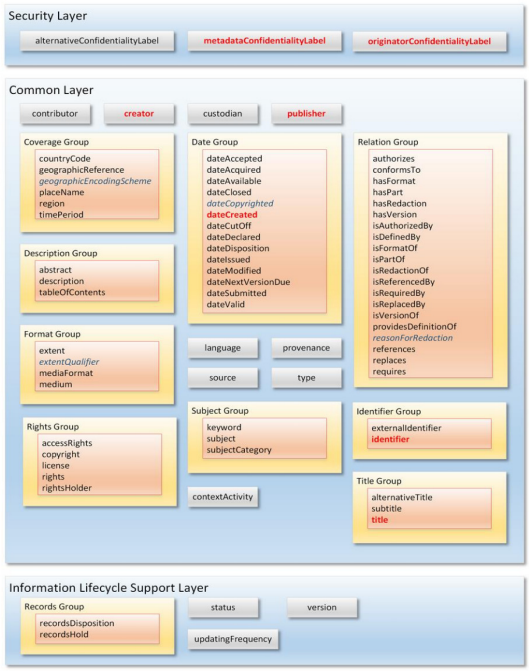
이러한 노력이 빛을 발하기 위해서는 양질의 데이터 확보를 위한 국방 차원의 정형데이터 및 비정형 데이터의 표준화·품질 기준 마련이 필요한 상황이나 지금까지 국방은 정형데이터의 표준화·품질 기준을 제안하고 있는 수준으로 이에 대한 보완이 필요하다. 현재, 국방 데이터 표준단어, 표준도메인, 표준코드, 데이터베이스 코드설계 표준을 분석해 보면 국방부 데이터양에 비례하여 표준단어, 표준도메인 관리 항목이 상대적으로 적은 실정이며, 이에 대한 보완책이 필요하다. 또한, 학습용 데이터의 경우 각 군과 기관의 필요에 따라 개별 구축되고 있어 전군 차원의 공동 활용 등 데이터 공유·융합에 많은 시간·비용이 발생한다.

고품질의 인공지능 결과를 도출하기 위해 비정형 데이터의 중요성이 높아지고 이에 비정형 데이터의 표준화를 위한 노력은 국방뿐 아니라 민간·공공영역에서 진행되고 있다. 국외에서 비정형 데이터 표준화는 JTC1 표준위원회 JTC I/SC 42(Artificial intelligence)에서 인공지능 전 분야 표준개발 및 가이드를 제정하여 제공한다[2]. 또한, 인공지능 및 머신러닝 데이터 품질, 인공지능 신뢰성 등 표준화를 진행하고 있다. 국내 인공지능 표준화는 TTA 표준위원회인 TTA 인공지능기반기술 PG (PG1005), TTA 스마트헬스 PG (PG419), TTA 메타데이터 PG (PG606)에서 주도하고 있다.

TTA 인공지능기반기술 PG(PG1005)는[3] 인공지능 기반 기술 및 인공지능 데이터 관련 국내 표준을 개발하는 위원회로서, 인공지능 기반 기술 및 데이터 분야 표준화와 데이터 품질 평가 지침 및 AI 신뢰성 확보 가이드라인 제정을 목표로 한다. 특히, 자율주행 AI 학습용 데이터 관리 분야 표준화에 주력하고 있으며, 'TTAK.KO-10.1339: 지도학습을 위한 데이터 품질 관리 요구사항 표준'을 제정하였다. TTA 스마트헬스 PG(PG419)는[4] 헬스케어 분야를 중점으로 비정형 데이터 표준화를 진행하고 있다. 의료 서비스 환경 구축을 위한 데이터 처리 및 헬스케어 시나리오별 프레임워크 표준개발을 목표로 하며, 인공지능을 위한 진단과 치료 관련 데이터 수집 및 응용서비스 모델 관련 표준을 개발한다. 현재, 표준화 산출물로는 '간호 업무 지원 음성 인식 모델 학습용 데이터 구축 표준'과 '진단 보존 인공지능 모델 개발을 위한 학습용 데이터 구축 방안 (병리조직 이미지)' 표준을 제정한 바 있다. TTA 메타데이터 PG(PG606)는[5] 메타데이터, 인공지능 데이터의 효율적 관리 목적으로 인공지능 데이터 거버넌스 기술을 개발한다. 인공지능 학습 데이터와 빅데이터 수집, 저장, 분석, 관리 목적으로 메타데이터 표준을 제정한다. 현재, TTAK.KO-10.1378: 마이데이터를 위한 데이터 품질점검 지침 표준을 제정하였다.

국내외 사례를 살펴보면, 헬스케어, 마이데이터, 자율주행 등 특정 도메인별 비정형 데이터셋의 표준화 방향을 정의하고 있으며, 아직까지 국내외 모두 비정형 데이터 표준을 제정하기 위한 초입 단계이다. 따라서, 본 연구팀은 NATO 표준 중 데이터 관련 표준인 STANAG 7023, STANAG 5636을 추가 분석하였다.

2022년 11월에 발행된 NATO Core Metadata Specification (NCMS)[6]는 나토 회원국 간의 기술적 데이터 상호운영성을 원활하게 하기 위한 토대를 형성하기 위한 표준이다. 이 표준은 공통 XML 기반 형식 및 구문을 제공하여 나토 회원국 간의 표준적인 기호를 이용할 수 있게 함으로써 호환성과 확장성을 보장하고 지상 분야에서는 C4I와 군사 개발, 작전, 훈련에서의 상호 이용성을 증가시키는 데 있다.



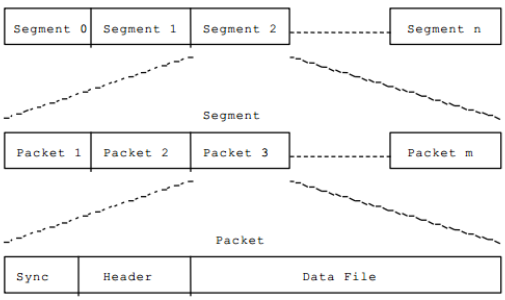
(그림 1) NCMS에서 제안하는 메타데이터 구조
(Figure 1) Metadata Structure of NCMS

표준에서는 그림 1과 같이 메타데이터를 레이어(Layers), 메타데이터 요소(Metadata Elements), 그룹(Group)으로 구성하고 있다. 레이어는 메타데이터의 최상위 요소를 의미하며, 메타데이터 요소는 리소스를 설명하는 가장 작은 기본단위이며, 그룹은 두 개 이상의 하위 메타데이터 요소를 포함하는 메타데이터를 의미한다.

이 표준에서는 레이어는 다음과 같이 3개로 구성된다.

- 보안레이어 : 접근제어, 공개 가능성 등의 정보 포함
- 일반레이어 : 데이터에 대한 설명정보, 관리정보 등을 포함하며, 주제, 설명, 범위와 같은 정보로 구성됨
- 정보 생애주기 제공 레이어 : 리소스의 전체 수명주기를 관리하는 데 필요한 현재 레코드 단계, 폐기정보, 보류정보, 업데이트 주기 및 간격 등을 포함

STANG 7023으로도 알려진 AEDP-7023[7]은 2022 년 7월에 발행된 표준으로 공중 정찰에 사용되는 기본 이미지 데이터에 대한 표준을 제공한다. 이 표준의 목적은 수집 시스템과 이용 시스템 간에 정찰 이미지와 보조 데이터를 전송하기 위한 표준 데이터 형식과 아키텍처를 제공하며, 통신 프로토콜을 의미하지는 않는다. 전송 데이터 포맷 단위는 그림 2와 같이 레코드이며, 레코드는 세그먼트로 구성되고 하나의 세그먼트는 패킷으로 구성된다. 패킷은 다시 헤더와 데이터 파일로 구성된다.



(그림 2) 레코드/세그먼트/패킷 구성
(Figure 2) Record/Segment/Packet Structure

패킷을 구성하는 헤더 정보는 데이터 파일에 대한 정보로 데이터 파일 사이즈, 순차번호, 타임 태그 등을 포함하고 있다. 데이터 파일은 센서 정보(Sensor Data)와 부가정보(Auxilary Data)로 구성되며, 센서 정보는 실제 센서 측정값이 저장되고, 부가정보는 센서 측정값에 대한 설명을 포함하는 메타데이터 정보이다. 부가정보를 구성하는 세부항목은 다음 그림과 같으며, 포맷 정보, 미션, 목표, 플랫폼, 센서 파라미터 정보 등을 포함하고 있다.

앞서 살펴본 NATO 표준은 우리나라 국방분야의 데이터 표준(안)을 수립하는데, 유용한 내용인 데이터 구조 및 메타데이터 구성항목을 제공하고 있다.

따라서 본 논문에서는 국방 비정형 데이터셋의 표준화를 위해 국방 비정형 데이터셋을 정의하고, NATO 표준에 기반하여 우리나라 국방 분야에 적합한 국방 비정형 데이터셋 표준포맷을 제안한다. 또한, 제안한 표준포맷에 기반하여 국방 비정형 데이터셋 표준화 방안을 제안한다.

2. 국방 비정형 데이터셋 표준포맷

2.1 국방 비정형 데이터셋 정의

2.1.1 비정형 데이터셋 정의 사례

국내외 사례를 분석하면 표 1과 같다. 대부분의 사례에서 비정형 데이터를 정의할 때 정형데이터와 비교하여 데이터의 구조와 저장 형태의 차이점 위주로 정의하고 있다. 정형데이터와 비정형 데이터의 가장 큰 차이점은 정형화된 구조의 여부에 따라 관계형 데이터베이스에 저장될 수 있는지에 대한 것으로서, 단일 구조로 정의되지 않고, 관계형 데이터베이스에 저장될 수 없는 데이터를 비정형 데이터로 정의하고 있다.

(표 1) 비정형 데이터 정의 사례
(Table 1) Example of Unstructured data Definition

국내외 사례	정의
TTA 정보통신 용어사전[8]	• 정의된 구조가 없이 정형화되지 않은 데이터
ISO/IEC 20546:2019, Information technology[9]	• 데이터란 비정형 데이터의 구조를 포함하지 않는 것 • 비정형 데이터란 데이터로 구성되지 않는 것
MongoDB[10]	• 미리 설정된 데이터 모델이나 스키마에 따라 구성되지 않아 기존의 관계형 데이터베이스(RDBMS)에 저장할 수 없는 정보
Data.gov[11]	• 범정부 영역에서 구조화된 데이터베이스에 저장되지 않는 모든 것

2.1.2 비정형 데이터셋 유형 정의 사례

표 2에서 볼 수 있듯이 대부분의 국내외 사례에서 동영상, 오디오, 텍스트 파일(메일본문, 보고서 등)을 비정형 데이터 사례로 제시하고 있다. MongoDB는 반정형 데

이터로 분류될 수 있는 웹페이지도 비정형 데이터로 분류하고 있으며, 센서 데이터를 포함한 신호 데이터는 제시하는 사례는 없었다.

(표 2) 비정형 데이터 유형 사례
(Table 2) Example of Unstructured data Type

국내외 사례	유형					
	동영상	오디오	이미지	보고서 (문서)	메일 본문	웹페이지
TTA 정보통신 용어사전	√	√	√	√	√	
ISO/IEC 20546:2019, Information technology					√	
MongoDB	√	√	√		√	√
Data.gov	√	√		√		

2.1.3 국방 비정형 데이터셋 및 유형 정의

본 논문에서는 앞서 살펴본 사례를 기반으로 국방 비정형 데이터와 국방 비정형 데이터 유형을 다음과 같이 정의한다.

- 국방 비정형 데이터 정의
  각 군, 국방부 산하기관 및 유관기관의 관계형 데이터베이스(RDBMS)에 저장·관리되지 않는 데이터
- 국방 비정형 데이터 유형 정의
  - 동영상
  - 오디오
  - 이미지
  - 텍스트 (메일본문, 보고서 등)
  - 센서데이터(BIT 데이터 등)

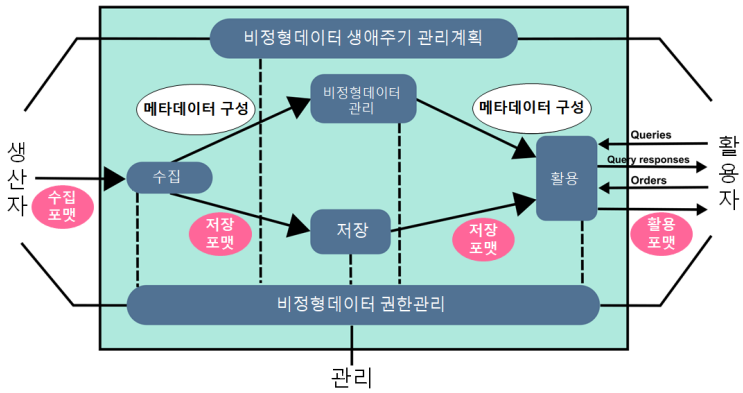
2.2 국방 비정형 데이터셋 표준포맷

2.2.1 국방 비정형 데이터셋 표준포맷

국제표준기구인 ISO 14721의 OAIS 참조모델[12]에서는 그림 3과 같이 데이터 생애주기 시점에 따라서 다음과 같이 데이터셋을 다르게 구성해야 한다고 명시하고 있다.

- 수집포맷(SIP: Submission Information Package): 생산자가 저장소 또는 플랫폼에 전송하는 정보 패키지
- 저장포맷(AIP: Archival Information Package): 저장소 또는 플랫폼에 저장되는 정보 패키지로 수집된 정보에 저장 및 관리에 필요한 정보가 추가되어 구성됨
- 배포포맷(DIP: Dissemination Information Package): 데이터를 활용하고자 하는 개인, 기관, 시스템에 제공되는 정보패키지로, 데이터 접근·열람·활용 권한에 따라 다른 정보로 구성될 수 있는 유연한 구조를 가짐

본 논문에서는 3가지 데이터 포맷 중 비정형 데이터 수집포맷에 대해 제안한다. 비정형 데이터 수집포맷은 각 군 및 관련기관에서 생성한 비정형 데이터를 국방 데이터플랫폼(가칭)에 전송하는데 필요한 국방 비정형 데이터 표준 수집포맷을 의미한다.



(그림 3) 비정형 데이터 표준포맷 유형
(Figure 3) Type of Unstructured Data Standard Format

(그림 4) 국방 비정형 데이터 구조(안)

(Figure 4) Draft Version of Defense Unstructured Data Structure

2.2.2 국방 비정형 데이터셋 구조

국방 비정형 데이터 표준화 및 품질관리 방안을 마련하기 위해서 국방 비정형 데이터 수집 포맷은 원시데이터 제공자와 수집자가 모두 이해할 수 있는 약속된 구조를 가져야 한다. 이를 위해 본 연구진은 국방 비정형 데이터의 수집 포맷에 대한 구조를 먼저 정의하였다. 메타데이터와 데이터값이 모두 정형화된 정형데이터와 달리 비정형 데이터는 정형화된 메타데이터와 비정형 데이터값으로 구성된다. 따라서 국방 비정형 데이터의 경우, 메타데이터와 비정형 데이터값의 개별적인 표준화·품질 기준이 마련되어야 한다. 본 논문에서는 국방 비정형 데이터 구조를 그림 4와 같이 제안한다. 국방 비정형 데이터는 메타데이터, 비정형 데이터셋, 전자서명으로 구성된다.

- 메타데이터: 정형데이터와 비정형 데이터 모두를 검색·활용·유통하는 플랫폼에서 데이터를 검색하는 데 필요한 정보로 구성됨
- 비정형 데이터셋: 비정형 데이터를 표현하는 비정형 데이터 메타데이터와 비정형 데이터값으로 구성되며, 비정형 메타데이터는 비정형 데이터값에 따라 다른 구성을 가지며, 비정형 데이터값은 데이터 특성에 따라 다양한 비정형 데이터형식으로 저장됨
- 전자서명 : 국방부에서 지침을 마련해야 할 사항으로 본 연구에서는 전자서명에 대해서는 다루지 않음

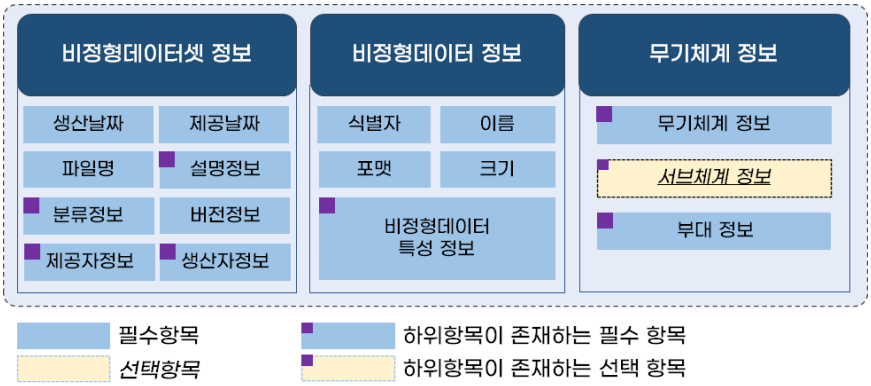
3. 국방 비정형 데이터셋 표준화 방안

국방 비정형 데이터 표준화는 그림 4의 구조에서 비정형 데이터셋의 비정형 데이터 메타데이터 표준화와 비정형 데이터값의 포맷 표준화로 나누어 정의될 수 있다.

비정형 데이터셋의 비정형 데이터 메타데이터 표준화는 비정형 데이터 유형에 따라 필수·선택 요소를 정의한다. 국방 비정형 데이터의 메타데이터 표준은 메타데이터 구조를 설계하고, 메타데이터 값의 표준화는 정형데이터 표준화와 같은 거버넌스를 따르도록 한다. 국방 비정형 데이터의 비정형 데이터값의 표준화는 비정형 데이터 유형별 표준 파일 포맷 표준화로 정의한다. 비정형 데이터값의 표준화는 데이터를 이용하는 사용자가 해결하려는 문제에 따라 필요한 표준화 수준이 다르고, 비정형 데이터의 유형별 표준화가 다르게 구성되어야 해서 비정형 데이터 포맷을 표준화하는 수준이 적정하다. 비정형 데이터 포맷을 표준화함으로써 특정 소프트웨어에 종속되지 않고, 필요한 데이터를 범용 소프트웨어로 활용함으로써 데이터의 활용성을 높일 수 있다.

3.1 비정형 데이터셋 메타데이터 표준화 방안

국방 비정형 데이터의 메타데이터는 NATO 표준을 참고하여 그림 5와 같이 ①비정형 데이터셋 정보, ②비정형



(그림 5) 국방 비정형 데이터 메타데이터 표준포맷

(Figure 5) Metadata Standard Format of Defense Unstructured Data

데이터 정보, ③무기체계 정보로 구성된다. 메타데이터는 필수항목과, 선택항목으로 구성되며, 각 항목은 하위항목 보유 여부에 따라 하위항목이 있는 경우 컨테이너 항목으로 구성한다.

- 비정형 데이터셋 정보

비정형 데이터셋 정보는 데이터셋의 설명 정보로 검색용 메타데이터로 확대하여 활용될 수 있다. 비정형 데이터셋 정보는 생산날짜, 제공날짜, 파일명, 설명 정보, 분류 정보, 버전 정보, 제공자 정보, 생산자 정보로 구성되며, 설명 정보, 분류 정보, 제공자 정보, 생산자 정보는 하위항목이 있는 컨테이너 항목으로 설계하였다.

- 비정형 데이터 정보

비정형 데이터 유형에 따른 특성정보를 설명하는 메타데이터 정보로, 식별자, 이름, 포맷, 크기 등의 필수항목과 비정형 데이터 특성 정보를 설명하는 컨테이너 항목으로 구성된다. 비정형 데이터 특성 정보는 비정형 데이터 유형에 따라 다른 정보로 구성될 수 있으며, 하위 구성요소 예제는 다음과 같다.

✓ 이미지: 이미지 가로, 이미지 세로, 이미지 비율 등
✓ 동영상: 해상도, 컬러, 초당 프레임, 영상 총프레임 등
✓ 오디오: 오디오 채널 수, 녹음 길이, 실내외, 스펙트럼 평균, RMS 게인 평균, 녹음장소 등
✓ 텍스트: 인코딩, 페이지수 등
✓ 센서: 센서종류, 센싱주기, 센싱종류(진동, 소리, 온도, 습도 등), 센싱단위 등
✓

- 무기체계 정보

무기체계 정보, 서브체계 정보, 부대 정보로 구성되며, 기존 국방 코드체계, KVMF(Korean Variable Message Format)[13] 과 같은 국방 무기체계 코드체계 등을 활용하여 구성한다. 서브체계 정보는 무기체계의 서브 체계 중 비정형 데이터와 관련된 서브 체계를 기술하며, 서브 체계 ID와 서브 체계 명을 하위 구성요소로 가진다. 부대 정보는 무기체계를 관리하는 부대 정보로 부대 ID, 부대 명 등이 기술된다.

3.2 비정형 데이터 파일 포맷 표준화 방안

3.2.1 비정형 데이터 파일 포맷 사례 조사·분석

본 논문에서는 표 3과 같이 미국, 호주, 영국, 핀란드 등의 사례와 우리나라 국립도서관, NIA, TTA 등에서 제안하는 표준 파일 포맷을 비교·분석을 통해 국방 비정형 데이터에 적합한 파일 포맷을 제안한다.

한국은 국립중앙도서관, 한국지능정보사회진흥원(NIA) AI Hub, 한국정보통신진흥협회(TTA)에서 제안하는 파일 포맷을 중심으로 살펴보았으며, 5개 국가에서 권고하는 동일한 파일 포맷을 분석하였다. 표 3에서 알 수 있듯이 비정형 데이터 유형별로 각 비정형 데이터에 적합한 표준 파일 포맷을 권고하고 있음을 확인할 수 있었다.

표 4는 표 3에서 조사된 파일 포맷의 빈도수를 정리한 것이다. 이미지는 TIFF를 표준포맷으로 가장 많은 국가가 권고하고 있으며, 2순위로 JPEG 2000과 PNG로 집계되었다. 동영상은 우리나라를 제외하고 4개국에서 Motion

(표 3) 이미지 데이터 메타데이터 항목구성(안)
(Table 3) Metadata Field for Image Data

번호	항목명	정의	컨테이너 여부	필수여부	반복 여부	데이터 타입
1	식별자	이미지 식별자	N	필수	N	String
2	파일포맷	이미지 포맷 ex) JPEG2000	N	필수	N	String
3	이미지 크기	이미지 크기 ex) 100MB	N	필수	N	Integer
4	이미지 가로	이미지 넓이	N	필수	N	Integer
5	이미지 세로	이미지 높이	N	필수	N	Integer
6	이미지 비율	이미지 비율	N	필수	N	Integer
7	이미지 해상도	이미지 해상도	N	필수	N	Array
8	촬영장소	촬영 장소	N	선택	N	Object
9	실내외 여부	실내외여부	N	필수	N	String
10	촬영환경	가상배경 활용 여부	N	필수	N	String
11	촬영장비 정보	촬영 장비 정보	N	필수	N	Object
12	측정 시작일시	이미지 측정 시작일시	N	필수	N	Date Time
13	측정 종료일시	이미지 측정 종료일시	N	필수	N	Date Time
14	생산자 정보	문서 생산자 정보	N	필수	N	Object

JPEG2000을 표준포맷으로 지정하고 있으며, 2순위에는 우리나라를 포함한 4개국에서 MPEG-4을 표준포맷으로 채택하고 있었다. 오디오는 Waveform Audio File Format과 MPEG-2 Audio Layer III을 모든 국가에서 표준포맷으로 채택하고 있으며, 우리나라를 제외하고 4개국에서는 Free Lossless Audio Codec을 채택하고 있었다. 텍스트는 5개국 모두 PDF/A를 채택하고 있으며, TIFF, XML, Plain text가 2순위로 집계된다.

센서데이터 표준화 연구는 국가별로 진행된 사례는 찾아보기 힘들었으며, 도메인(지리, 차량, 해양 등)별로 연구되고 있다. 한국은 서울시에서 미세먼지 농도를 IoT 센서를 이용해 실시간으로 수집하는 연구를 수행하며 S-DOT이라는 규격을 구축하였으며, OGC에서 제정한 SensorML은 지리적 특성에 맞도록 설계된 센서 데이터 포맷이다. ITU-T FG-AI4A에서는 차량에서 수집한 센서의 데이터 교환을 위해 JSON 파일 포맷을 제안하고 있으며, Leidos Company에서 관리하고 있는 GSF는 미국방부 해군에서도 이용하고 있는 해양 수심 측정 센서데이터 포맷으로 XML, CSV, JSON 등으로 변환이 가능하다.

(표 4) 국가별 이미지, 동영상, 오디오, 텍스트 파일 포맷 비교
(Table 4) Comparison of image, video, audio, and text file formats

구분	한국	미국	호주	영국	핀란드
출처	• 국립중앙도서관[14] • NIA AI Hub[15] • TTA[16]	• 미국의회 도서관[17]	• National Archives of Austrailia[18]	• United Kingdom Data Service[19] • UK.gov[20]	• Finnish Social Science Data Archive[21]
이미지	• TIFF • JPEG	• TIFF • JPEG2000(*.jp2) • PNG(*.png)	• PNG (*.png) • JPEG 2000(*.jp2) • JFIF, TIFF • GIF, DNG	• TIFF • DICOM (*.dcm,*.dcm30)	• DNG • JPEG • JPEG2000(*.jp2) • PNG, TIFF
동영상	• MXF • MPEG-4 (*.mp4)	• Motion JPEG2000(*.mj2) • AVI(*.avi) • QuickTime Movie (*.mov)	• MKV • Ogg(*.ogv) • Motion JPEG2000(*.mj2)	• MPEG-4 (*.mp4) • OGG vides (*.ogv, *ogg) • Motion JPEG 2000(*.mj2)	• DPX • Motion JPEG 2000 (*.mj2) • MPEG-4(*.mp4)
오디오	• Waveform Audio File Format (uncompressed) (.wav) • MPEG-2 Audio Layer III (.mp3)	• Waveform Audio File Format(uncompressed) (.wav) • Broadcast Waveform Audio File Format (uncompressed) (.bwav) • Audio Interchange File Format (.aiff) • Free Lossless Audio Codec (.flac) • MPEG-2 Audio Layer III (.mp3)	• Waveform Audio File Format (uncompressed) (.wav) • MPEG-2 Audio Layer III (.mp3)	• Free Lossless Audio Codec (.flac) • MPEG-2 Audio Layer III (.mp3) • Audio Interchange File Format (.aif) • Waveform Audio Format (.wav) • Audio Interchange File Format (.aiff) • Broadcast Waveform Audio File Format (uncompressed) (.bwav)	• Free Lossless Audio Codec (.flac) • Waveform Audio Format (.wav) • MPEG-2 Audio Layer III (.mp3)
텍스트	• TIFF • JPEG • PDF	• Plain text • XML • PDF/A-1 (*.pdf)	• PDF	• RTF(*.rtf), TIFF • PDF/A or PDF (*.pdf) • HTML (*.html) • ODT (*.odt)	• CSV, EPUB • XHTML, HTML • XML, ODF • PDF/A • Plain text

(표 5) 이미지, 동영상, 오디오, 텍스트 파일 포맷 순위
(Table 5) Ranking of image, video, audio, and text file format

구 분	이미지	동영상	오디오	텍스트
1순위	• TIFF (5개국)	• Motion JPEG2000 (4개국)	• Waveform Audio File Format (5개국) • MPEG-2 Audio Layer III (5개국)	• PDF/A or PDF (5개국)
2순위	• JPEG 2000 (3개국) • PNG(3개국)	• MPEG-4 (3개국)	• Free Lossless Audio Codec (4개국)	• TIFF (2개국) • XML (2개국) • Plain text (2개국)

3.2.2 국방 비정형 데이터 파일 포맷 표준화 방안

앞서 살펴본 사례와 우리나라 국방에서 사용하는 파일 포맷을 분석하여 다음과 같이 국방 비정형 데이터 표준 파일 포맷을 제안한다.

• 이미지 표준 파일 포맷(안)

이미지 표준 파일 포맷으로는 TIFF, JPEG2000, PNG, JPEG를 제안한다. TIFF, JPEG2000, JPEG는 우리나라에서 표준 파일 포맷으로 권고하고 많이 활용되고 있으며, 미국과 데이터 교환이 빈번한 상황을 고려하여 미국에서 권고하는 PNG도 우리나라의 표준 파일 포맷으로 고려할 필요가 있다.

• 동영상 표준 파일 포맷(안)

Motion JPEG2000, MPEG-4, AVI를 표준 파일 포맷으로 제안 한다. 우리나라를 제외하고 4개국은 Motion JPEG2000을 표준 파일 포맷으로 권고하고 있으며, MPEG-4와 AVI는 우리나라에서 권고하는 표준 파일 포맷으로 많은 동영상 파일이 MPEG-4와 AVI로 생성되고 있어 동영상 표준 파일 포맷으로 제안한다.

• 오디오 표준 파일 포맷(안)

오디오 표준 파일 포맷으로는 Waveform Audio File Format (uncompressed), MPEG-2 Audio Layer III를 제안한다. 이 2가지 파일 포맷은 우리나라를 포함하여 5개국 모두가 오디오 표준 파일 포맷으로 권고하고 있다.

• 텍스트 표준 파일 포맷(안)

PDF, XML, 텍스트파일을 표준 파일 포맷으로 제안한다. 우리나라는 종이 문서를 스캔하여 생산한 전자문서 파일의 표준 파일 포맷으로 TIFF, JPEG을 제안하고 있으나, 이는 이미지 표준 파일 포맷에 적합하다. 따라서 전자매체에서 생성된 전자문서 파일의 경우에는 PDF나 XML을 표준 파일 포맷을 사용하는 것을 제안한다.

• 센싱 데이터 파일 포맷(안)

센싱 데이터는 데이터 종류에 따라 파일포맷이 다르기 때문에 특정 파일 포맷을 제안하기는 어렵다. 그러나 센싱 데이터의 활용·유통의 목적으로 표준 파일 포맷의 필요성을 판단하여 범용적으로 사용되는 반정형 데이터 교환 포맷인 JSON, TXT 파일을 센서 데이터 표준 파일 포맷으로 제안한다. 또한, 현재 군에서는 CSV를 센서 데이터 포맷으로 활용하고 있어 CSV도 표준포맷으로 제안한다.

4. 결론 및 향후 연구

비정형 데이터는 정형데이터와 달리 표준화 및 품질관리가 어렵다. 그 이유는 비정형 데이터는 유형별로 데이터 구조가 달라서 일관된 형태로 표준화가 어렵고, 품질의 기준이 모호하기 때문이다. 비정형 데이터의 품질은 사용자가 어떤 문제를 해결하는 데 필요한 데이터가 충분한지에 따라 판단될 수 있는 부분이다. 데이터 공유 및 활용·확산을 위해서는 비정형 데이터 관리가 필수적이다. 특히 국방에서는 데이터 기반으로 무기의 첨단화, 후속군 수지원 등을 지원하기 위한 연구가 활발히 진행되고 있다.

본 논문에서는 국방 데이터의 활용·활성화를 위해 국방 비정형 데이터를 정의하고, 국방 비정형 데이터 유형을 정의하였다. 국방 비정형 데이터 유형은 국내외 사례 및 군에서 생산되는 데이터 유형을 조사하여 동영상, 이미지, 오디오, 텍스트, 센싱 데이터로 정의하였으며, 추후 군과 관련기관의 요구사항을 수용하여 추가될 수 있다.

또한, 국방 비정형 데이터셋 메타데이터와 비정형 데이터셋으로 구성된 표준 수집 포맷을 제안하고, 이 표준 포맷에 포함되는 메타데이터 표준(안)을 제안하였다. 메타데이터 표준(안)은 비정형 데이터를 설명하는 비정형 데이터 메타데이터로, 비정형 데이터 특성, 관련 무기체

계 정보 등을 포함하도록 제안하였다.

마지막으로 표 6과 같이 국방 비정형 데이터 파일 포맷 표준화 방안을 제안함으로써, 데이터 활용자가 특정 소프트웨어나 환경에 의존되지 않도록 범용적인 활용이 가능할 것으로 기대된다. 즉, 우리나라 무기체계에서 활용되는 이미지 파일의 경우에는 TIFF, JPEG 등으로 생산하도록 함으로써, 이미지 활용자는 파일 포맷 변환 시간을 줄일 수 있다. 또한, 국외에서도 많이 활용되는 파일 포맷을 사용함으로써 미국 등과의 합동 훈련 등에서 필요한 데이터 교환이 빠르게 이루어질 수 있다.

(표 6) 국방 비정형 데이터 파일 포맷(안)
(Table 6) Unstructured Data File Format

구 분	동영상
이미지	• TIFF • JPEG2000(*.jp2) • PNG • JPEG(*.jpg)
동영상	• Motion JPEG2000(*.mj2) • MPEG-4(*.mp4) • AVI(*.avi)
오디오	• Waveform Audio File Format (uncompressed) (.wav) • MPEG-2 Audio Layer III (.mp3)
텍스트	• TXT • PDF • XML
센싱	• JSON • CSV

향후, 본 논문에서 제안한 국방 비정형 데이터 표준포맷과 표준화 방안에 대해 각 군 및 관련자로부터 피드백을 받아 보완·수정할 계획이 있다. 무기체계에서 생산되는 데이터를 본 논문에서 제안한 파일 포맷, 표준 메타데이터 포맷 등에 맵핑하고, 생산 편의성, 활용성 등을 반영하여 보완할 예정이다. 2023년 하반기 까지 1차로 방산업체를 대상으로 검증을 완료하고, 2024년 상반기 국방부의 지원을 받아 각 군의 의견을 수렴할 예정이다.

참고문헌(Reference)

[1] Department of Defense, “2023~2027 National Defense Science and Technology Innovation Basic Plan” for Diabetes Management,” 2023.

[2] ISO/IEC JTC 1/SC 42 - Artificial intelligence, “ISO/IEC 8183:2023 Information technology - Artificial intelligence - Data life cycle framework”, International Standard, 2023.

[3] Telecommunications Technology Association (TTA), “Requirements for Data Quality Management of Supervised Learning”, TTAK.KO-10.1339, TTA Standard, 2021.

[4] Telecommunications Technology Association (TTA), “Image Exchange Platform for Smart Health Service - Part 4: Categorization of Data Label for Supervised Learning of Medical Image”, TTAK.KO-10.1231-Part4, TTA Standard, 2022.

[5] Telecommunications Technology Association (TTA), “Data Quality Measurement Guideline for MyData”, TTAK.KO-10.1378, TTA Standard, 2022.

[6] NATO, “NATO STANDARD ADatP-5636 NATO Core Metadata Specification (NCMS)”, 2022.11

[7] NATO. “NATO STANDARD AEDP-7023 AIR RECONNAISSANCE PRIMARY IMAGERY DATA STANDARD”, 2022.7

[8] Telecommunications Technology Association (TTA) Information and Communication Terminology Dictionary, http://terms.tta.or.kr/main.do

[9] ISO/IEC 20546, “Information technology - Big Data - Overview and vocabulary”, International Standard, 2019.

[10] MongoDB, https://www.mongodb.com/unstructured-data

[11] Data.gov, https://data.gov/

[12] ISO/IEC 14721:2003, “Open archival information system (OAIS) — Reference model”, International Standara, 2003.

[13] Choi. Il-Ho, Kim. Dae-Young, Kwon. Chul-Hee, and Lee. Sang-Myung, “An Implementation of KVMF(Korean Variable Message Format) in the Battlefield Management System of Ground Fighting Vehicles”, Journal of the Korea Institute of Military Science and Technology, Vol.17, Issue.5, pp.663-671, 2014, https://doi.org/10.9766/KIMST.2014.17.5.663

[14] National Library of Korea, “Guidelines for depositing and collecting online materials of the National Library of Korea”, 2002.

[15] AI Hub, https://www.aihub.or.kr/

[16] Telecommunications Technology Association (TTA), https://www.tta.or.kr/
[17] USA Library of Congress, https://www.loc.gov/
[18] National Archives of Australia, https://www.naa.gov.au/
[19] UK Data Service, https://ukdataservice.ac.uk/
[20] UK.gov, https://www.gov.uk/
[21] Finnish Social Science Data Archive, https://www.fsd.tuni.fi/en/

◐ 저 자 소 개 ◑

황 윤 영(Yun-Young Hwang)
2002년 충남대학교 정보통신 및 컴퓨터공학 (공학사)
2004년 충남대학교 대학원 컴퓨터공학과(공학석사)
2011년 충남대학교 대학원 컴퓨터공학과(공학박사)
2011년~2012년 충남대학교 BK21 박사후연구원
2012년~현재 한국과학기술정보연구원 책임연구원
관심분야 : 데이터베이스, 데이터표준화, CBM+(Condition Based Maintenance Plus), 국방데이터
E-mail : yyhwang@kisti.re.kr

손 지 성(Jiseong Son)
2007년 서울여자대학교 컴퓨터공학과(공학사)
2009년 고려대학교 대학원 컴퓨터전파통신공학과(공학석사)
2016년 고려대학교 대학원 컴퓨터전파통신공학과(공학박사)
2016년~현재 한국과학기술정보연구원 선임연구원
관심분야 : 지식그래프, 데이터표준화, CBM+, etc.
E-mail : jsson@kisti.re.kr