논문 2017-1-2

소프트웨어 컴플라이언스를 위한 SPDX Parser 및 Validator

윤호영*, 조용준*, 정병옥*, 신동명*+

SPDX Parser and Validator for Software Compliance

Ho-Yeong Yun*, Yong-Joon Joe*, Byung-Ok Jung*, Dong-Myung Shin*+

요 약

수 많은 파일로 이루어진 소프트웨어 패키지를 일일이 분석하는 것은 많은 시간과 비용을 요구하는 작업이다. 이에 리눅스 재단의 워킹그룹인 SPDX에서는 소프트웨어의 명세정보(메타데이터) 규약을 발표하였다. SPDX 문서는 2017년 상반기 기준 2.1버전이 발표되었으며, 총 7개의 콘텐츠에 66개 항목이 존재한다. 또한 Tag/Value 형식과 RDF형식을 권장하며, 스프레드시트 형식을 지원한다. 본 연구에서는 SPDX 문서를 각 항목별로 분류하고, 유효성 검사를 해주는 SPDX Parse & Validator 툴을 개발하였다. 추후 SPDX 문서를 생성(Generator)하는 툴을 개발하여 보다 효율적으로 소프트웨어 패키지를 관리하고자 한다.

Abstract

Analyzing a software package which is consisted of big numbers of files takes enormous costs and time. Therefore, SPDX (Software Package Data Exchange) working group collaborate with Linux Foundation published a software information(metadata) specification: SPDX. On the first half of 2017, the specification contains seven chapters and 66 items, according to Ver 2.1 of SPDX spec. It prefers Tag/Value or RDF forms but also supports spreadsheet form. In this paper, we introduce SPDX parsing & validation tools to check the validity of SPDX document. We'll develop SPDX document generator to manage software package more efficiently for our next target.

한글키워드 : 소프트웨어 패키지, 소프트웨어 정보교환명세, 구문 분석, 유효성 검사

I. 서 론

소프트웨어 개발에는 많은 인력이 투입된다. 규모가 크고, 전문성을 띄는 소프트웨어일수록 다양한 업체(혹은 부서)가 참여하기도 한다. 그림 1을 보면, 수주업체(원도급자)는 발주기관에서 소프트웨어 개발을 의뢰받아 하도급자 A, B에게 개발을 도급한다. 하도급자 A와 B는 또 다른 전

* 엘에스웨어(주)
+ 신동명 (교신저자)
(email: roland@lsware.com)
접수일자: 2017.05.22 수정완료: 2017.06.13

문 인력에게 개발을 의뢰하기도 하는데, 주로 프로젝트 매니저(PM)가 각 단계별 개발을 관리하게 된다. 프로젝트 매니저는 개발 모듈을 취합하는 단계에서 파일들을 분석하는 컴플라이언스 과정을 거쳐야 한다. 하지만 오픈소스 소프트웨어 라이선스 충돌 검사, 특허권 위반, 저작권 위반 등을 검사하기 위해 다양한 형태의 수많은 파일을 분석하는 작업은 시간과 비용을 요구한다.

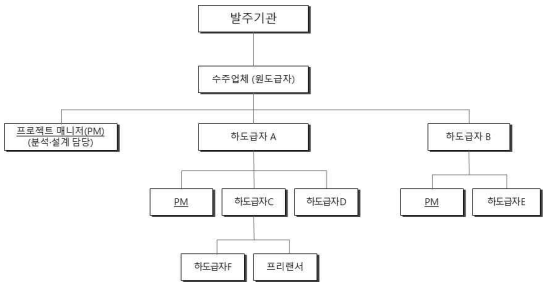

그림 1. 소프트웨어 업계 하도급 구조 예시

리눅스 재단의 워킹그룹에서는 위의 예시를 비롯하여, 소프트웨어 정보 공유를 원활하게 하는 것을 목적으로 데이터 교환 형식을 제안하고 있다. 이를 SPDX(Software Package Data Exchange)라고 하며, 2010년 1.0버전을 시작으로 1.1, 1.2, 2.0버전에 이어 2016년에 2.1버전을 발표하였다. SPDX 문서는 SPDX 문서 생성자 정보, 패키지 정보, 패키지 구성파일 정보, 파일 일부(Snippet) 정보, 라이선스 정보, SPDX 요소 관계 정보, 주석 정보를 담고 있으며, Tag/Value, RDF, HTML, 스프레드시트 형식을 지원한다 [1-2].

본 연구에서 개발한 SPDX Parser 및 Validator는 SPDX 문서 정보를 항목에 맞게 추출해주며, 해당 항목의 유효성 검사를 진행해주는 기능을 가지고 있다. 또한 지원 형식은 Tag/Value와 RDF이며, 2.0버전과 2.1버전을 지원한다. SPDX 문서의 지원 형식과 문서의 구성요소들에 대한 설명을 2장에서 서술하였다. 3장에서는 SPDX Parser와 Validator의 기능에 대해 서술하였고, 4장에서는 결론 및 향후 연구에 대해 서술하였다.

2. SPDX 문서 분석[3]

2.1 지원 형식

SPDX 문서는 공식적으로 Tag/Value 형식, RDF 형식을 권장하고 있지만, HTML형식이나 스프레드시트 형식 또한 활용되기도 한다.

SPDX 문서에서 Tag/Value 형식은 HTML에서 사용되는 괄호 형태(예:<text>apple</text>)의 Tag/Value 방식이 아닌 서술형에 가깝다. SPDX 문서의 구성요소 제목 앞에는 '##'를 명시하고, 해당 구성요소의 컨텐츠들을 데이터 형식에 맞게 기술한다. Tag/Value 형식의 예는 표 1과 같다.

표 1. SPDX Tag/Value 형식의 예

```
## Creation Information
Creator: Tool: LicenseFind-1.0
Creator: Organization: Example()
Creator: Person: Jane Doe()
Created: 2010-01-29T18:30:22Z
```

RDF(Resource Description Framework) 형식은 웹상의 자원의 정보를 표현하기 위한 규격으로 상이한 메타데이터 간의 어의, 구문 및 구조에 대한 공통적인 규칙을 지원하는 것을 목적으로 한다 [4]. RDF는 정보 자원(Resource), 속성 유형(Property Type), 속성값(Value)으로 구성된다. SPDX 문서에서 RDF 형식의 예시는 표 2와 같다.

표 2. SPDX RDF 형식의 예

```
<spdx:creationInfo>
  <spdx:CreationInfo>
    <spdx:creator>Person: Jane Doe ()</spdx:creator>
    <spdx:creator>Organization: Example()</spdx:creator>
    <spdx:creator>Tool: LicenseFind-1.0</spdx:creator>
    <spdx:created>2010-01-29T18:30:22Z</spdx:created>
</spdx:CreationInfo>
```

2.2. Document Creation Information

문서 생성 정보(Document Creation Information)는 SPDX 문서의 기본적인 정보를 의미하며, 반드시 기술되어야 하는 정보이다. 항목으로는 SPDX 문서의 버전(SPDX Version), 데이터 라이선스(Data License), SPDX 고유 식별번호(SPDX Identifier), 문서의 이름(Document Name), SPDX 문서 네임스페이스(SPDX Document Namespace), 외부 문서 참조(External Document References), 라이선스 목록 버전(License List Version), 생성자(Creator), 생성날짜(Created), 생성자 의견(Creator Comment), 문서 의견(Document Comment)이 있다.

문서 주석의 경우 2.0버전 대비 2.1버전에서 새롭게 추가된 항목이다. 표 3은 문서 생성 정보 항목의 필수 여부를 기술하였다.

2.3. Package Information

패키지 정보는 SPDX 문서를 만드는데 관여한 개인, 조직 및 도구들의 정보를 의미하며, 반드시 기재할 필요는 없다. 항목으로는 패키지 이름(Package Name), 패키지 SPDX 고유 번호(Package SPDX Identifier), 패키지 버전(Package Version), 패키지 파일 이름(Package File Name), 패키지 공급자(Package Supplier), 패키지 원작자(Package Originator), 패키지 다운로드 위치(Package Download Location), 파일 분석여부(File Analyzed), 패키지 검증 코드(Package Verification Code), 패키지 체크섬(Package Checksum), 패키지 홈페이지(Package Homepage), 출처 정보(Source Information), 확정된 라이선스(Concluded License), 파일에 대한 모든 라이선스 정보(All Licenses Information from Files), 선언된 라이선스(Declared License), 라이선스 설명(Comments on License), 저작권 문구(Copyright Text), 패키지 요약 설명(Package Summary Description), 패키지 상세 설명(Package Detailed Description), 패키지 설명(Package Comment), 외부 참조(External Reference), 외부 참조 설명(External Reference Comment)이 있다.

2.0버전과 비교했을 때, 파일 분석 여부와 외부 참조, 외부 참조 설명 항목이 2.1버전에서 새롭게 추가됐으며, 표 4는 패키지 정보의 항목별 필수 여부를 기술하였다.

표 3. 문서 생성 정보의 항목별 필수 여부

항목	필수 여부
SPDX Version	필수
Data License	필수
SPDX Identifier	필수
Document Name	필수
SPDX Document Namespace	필수
External Document References	선택
License List Version	선택
Creator	필수
Created	필수
Creator Comment	선택
Document Comment+	선택

+ : 2.1 버전에서 추가된 항목

표 4. 패키지 정보의 항목별 필수 여부

항목	필수 여부
Package Name	필수
Package SPDX Identifier	필수
Package Version	선택
Package File Name	선택
Package Supplier	선택
Package Originator	선택
Package Download Location	필수
Files Analyzed+	선택
Package Verification Code	필수
Package Checksum	선택
Package Home Page	선택
Source Information	선택
Concluded License	필수
All Licenses Information from Files	필수
Declared License	필수
Comments on License	선택
Copyright Text	필수
Package Summary Description	선택
Package Detailed Description	선택
Package Comment	선택
External Reference+	선택
External Reference Comment+	선택

+ : 2.1 버전에서 추가된 항목

2.4. File Information

파일 정보는 소프트웨어 패키지에 들어있는 각각의 파일에 대한 라이선스 및 저작권을 포함한 정보를 의미한다. 항목은 파일 이름(File Name), 파일 SPDX 고유 식별번호(File SPDX Identifier), 파일 유형(File Type), 파일 체크섬(File Checksum), 확정 라이선스(Concluded License), 파일내의 라이선스 정보(License Information in File), 라이선스에 관한 설명(Comments on License), 저작권 문구(Copyright Text), 파일 설명(File Comment), 파일 고지(File Notice), 파일 기여자(File Contributor)가 있다.

2.0버전에 존재했던 파일의 출처가 되는 프로젝트 명칭을 의미하는 파일 출처 프로젝트명(Artifact of Project Name)과 파일의 출처가 되는 프로젝트 홈페이지를 의미하는 파일 출처 프로젝트 홈페이지(Artifact of Project Homepage), 파일출처 프로젝트의 링크 정보를 의미하는 파일 출처 프로젝트 URI 항목은 2.1버전에서 채택되지 않았다. 표 5는 파일 정보의 항목별 필수 여부를 기술하였다.

표 5. 파일 정보의 항목별 필수 여부

항목	필수 여부
File Name	필수
File SPDX Identifier	필수
File Type	선택
File Checksum	필수
Concluded License	필수
License Information in File	필수
Comments on License	선택
Copyright Text	필수
Artifact of Project Name-	선택
Artifact of Project Homepage-	선택
Artifact of Project Uniform Resource Identifier-	선택
File Comment	선택
File Notice	선택
File Contributor	선택

- : 2.1 버전에서 제외된 항목

2.5. Snippet Information

Snippet 정보는 SPDX 2.1버전에서 처음 채택된 내용이다. Snippet은 작은 정보, 코드조각 따위의 사전적 의미를 갖는데, SPDX 문서에서는 웹이나 다른 소프트웨어 제품에서 복사한 저작권 및 라이선스 조건이 첨부된 코드로 정의한다. 본 논문에서는 영문으로 기재하였다.

Snippet 정보의 항목으로는 Snippet SPDX 고유 식별번호(Snippet SPDX Identifier), Snippet 파일의 SPDX 고유 식별번호(Snippet from File SPDX Identifier), Snippet 바이트 범위(Snippet

Byte Range), Snippet 라인 범위(Snippet Line Range), Snippet 확정 라이선스(Snippet Concluded License), Snippet의 라이선스 정보(License Information in Snippet), Snippet의 라이선스 정보에 대한 설명(Snippet Comments on License), Snippet 저작권 문구(Snippet Copyright Text), Snippet 설명(Snippet Comment), Snippet 이름(Snippet Name)으로 구성되어있다. 표 6은 Snippet 정보의 항목별 필수 여부를 기술하였다.

표 6. Snippet 정보의 항목별 필수 여부

항목	필수 여부
Snippet SPDX Identifier	필수
Snippet from File SPDX Identifier	필수
Snippet Byte Range	필수
Snippet Line Range	선택
Snippet Concluded License	필수
License Information in Snippet	선택
Snippet Comments on License	선택
Snippet Copyright Text	필수
Snippet Comment	선택
Snippet Name	선택

2.6. Other Licensing Information Detected

감지된 기타 라이선스 정보는 패키지에 포함된 라이선스 중 표준에 정의되지 않은 기타 라이선스들에 대한 정보이다. SPDX는 널리 사용되고 있는 표준 라이선스들의 전문을 공지하고 있으며, 이는 공식 홈페이지에서 확인가능하다. 항목은 라이선스 식별자(License Identifier), 추출된 문장(Extracted Text), 라이선스 이름(License Name), 라이선스 상호 참조(License Cross Reference), 라이선스 설명(License Comment) 으로 이루어져 있다. 표 7은 감지된 기타 라이선스 정보의 항목별 필수 여부를 기술하였다.

표 7. 감지된 기타 라이선스 정보의 항목별 필수 여부

항목	필수 여부
License Identifier	필수
Extracted Text	필수
License Name	필수
License Cross Reference	선택
License Comment	선택

2.7. Relationships between SPDX Elements

SPDX 구성요소 사이 관계성을 나타내는 항목으로는 2개의 구성요소 사이의 관계에 대한 정보를 기재하는 관계성(Relationship) 항목과 관계성에 관련된 일반적인 설명을 기재한 관계성 설명(Relationship Comment) 항목으로 이루어져 있다. ‘foo.c’ 파일이 ‘bar.tgz’ 압축파일에 존재한다는 사실이나 ‘Apache Xerces’의 메타파일은 ‘pom.xml’이다 등이 관계성의 예로 들 수 있다. 표 8은 SPDX 구성요소 사이 관계성 정보의 항목별 필수 여부를 기술하였다.

표 8. SPDX 구성요소 사이 관계성 정보의 항목별 필수 여부

항목	필수 여부
Relationship	선택
Relationship Comment	선택

2.8 Annotations

주석은 파일, 패키지 또는 전체 문서에 대한 설명을 기술한다. 주석자(Annotator), 주석일자(Annotation Date), 주석 유형(Annotation Type), 정보교환명세 식별자 참조(SPDX Identifier Reference), 주석 설명(Annotation Comment) 항목으로 이루어져 있으며, 주석이 존재하는 경우 위의 항목들은 필수로 기재되어야 한다. 표 9는 주석 정보의 항목별 필수 여부를 기재하였다.

표 9. 주석 정보의 항목별 필수 여부

항목	필수 여부
Annotator	선택
Annotator Date	선택
Annotator Type	선택
Annotator Reference	선택
Annotator Comment	선택

3. SPDX Parser & Validator

SPDX Parser와 Validator는 Python 3.5로 개발되었으며, Python 3.0 이상의 버전이 설치된 환경에서 실행가능하다. SPDX Parser는 SPDX 파일을 분석하여, 항목별로 내용을 분류하는 기능이다. Tag/Value 형식과 RDF 형식을 지원하며, SPDX 버전 2.0과 2.1을 지원한다. Parser 기능은 주로 ElementTree 라이브러리를 활용하였다. ElemenTree는 Fredrik Lundh가 개발한 XML Generator & Parser 라이브러리로 외부 라이브러리로 존재하다가 Python 2.5 버전부터 통합되었다.

Validator는 SPDX 항목이 데이터 포맷에 맞게 입력되었는지를 검증하는 기능이다. 대부분의 항목은 문자열 형식이기 때문에 유효성 검사를 필요로 하지 않으며, 문자열 형식이 아닌 항목과 기타 검증이 필요한 항목에 대해 표 10에 기술하였다.

4. 결론 및 향후 연구

소프트웨어 패키지에 존재하는 파일들을 일일이 분석하여 정보를 추출하는 과정은 시간과 비용을 요구한다. 하지만 소프트웨어 패키지의 정보 명세 규격인 SPDX 문서가 존재한다면 해당 작업을 간소화할 수 있다. 본 연구는 SPDX 문서의 정보를 추출하고, 검증하는 도구인 SPDX

표 10. SPDX 항목의 검증방식

항목	검증방식
SPDX Version	SPDX-M.N의 숫자 체크
Data License	License 리스트에 있는지 체크
SPDX Document Namespace	“#”문자가 포함되어 있는지 체크
License List Version	M.N 숫자 체크
Creator	괄호안의 E-mail 체크
Created	UTC 시간형식 검사
Package Version	M.N 숫자 체크
Package Supplier	괄호안의 E-mail 체크
Package Originator	괄호안의 E-mail 체크
Package Download Location	URL 체크
File Anaylzed	True, False 여부 검사
Package Verification Code	160bit binary 형식 검사
Package Checksum	해쉬값(SHA1) 유효성 검사
Package Homepage	URL 유효성 검사
Source Information	<text> 태그 여부
Concluded License	License 리스트에 있는지 체크
Declared License	
Package Summary Description	<text> 태그 여부
Package Detailed Description	<text> 태그 여부
Package Comment	<text> 태그 여부
File Format	확장자 리스트에 포함되었는지 체크
File Checksum	해쉬값(SHA1) 유효성 검사
Copyright Text	<text> 태그 여부
File Comment	<text> 태그 여부
File Notice	<text> 태그 여부
Snippet Byte Range	num 1:num 2 분류 체크
Snippet Line Range	num 1:num 2 분류 체크
Snippet Comments on License	<text> 태그 여부
Snippet Copyright Text	<text> 태그 여부
Snippet Comment	<text> 태그 여부
Extracted Text	<text> 태그 여부
License Name	License 리스트에 있는지 체크
License Cross Feference	License 리스트에 있는지 체크
License Comment	<text> 태그 여부
Relationship Comment	<text> 태그 여부
Annotator	괄호안의 E-mail 체크
Annotation Date	UTC 시간형식 검사
Annotation Type	REVIEW|OTHER 분류
,Annotation Comment	<text> 태그 여부

Parser & Validator를 개발하기 위해 SPDX 규격 문서를 분석하였다. 총 7개의 콘텐츠에는 66개 항목이 있으며, 해당 항목의 데이터 형식에 맞게 도구를 구현하였다.

SPDX 문서는 일반 개발자들이 입력하기에는 항목이 많은 편이다. 본 연구의 향후 연구로 SPDX 문서를 생성해주는 도구를 개발하고자 한다. 널리 사용되고 있는 마이크로소프트사의 Visual Studio의 플러그인 형태로 개발하여, 사용자가 프로젝트를 컴파일할 때, 자동으로 SPDX 문서가 생성될 수 있도록 하는 것을 목적으로 한다. SPDX 문서가 활성화되면, 소프트웨어 컴플라이언스와 같은 분석 단계에서 적극 활용될 것으로 기대된다.

.

Acknowledgment

본 연구는 문화체육관광부 및 한국저작권위원회의 2017년도 저작권기술개발사업의 연구결과로 수행되었음

참 고 문 헌

[1] SPDX, “Software Package Data Exchange; Specification Version: 2.1”, 2016

[2] CIO Korea, “리눅스 재단, 오픈소스 라이선스 고민 해결책 SPDX 발표“, 2011.08.

[3] TTA, “공개소프트웨어 정보교환명세”, TTAK.KO-11.0182/R1, 2015

[4] RDF, http://www.w3.org/RDF

저 자 소 개

윤호영
2012년 한성대학교 산업경영공학과 학사 졸업
2016년 연세대학교 정보산업공학과 박사과정 수료
2016년-현재 엘에스웨어(주)

<주관심분야 : 최적화 이론, 알고리즘>

조용준
2011년 큐슈대학교 전기정보공학과 학사 졸업
2016년 큐슈대학교 정보학과 박사과정 수료
2016년-현재 엘에스웨어(주)

<주관심분야: 게임이론, 분산 최적화 이론, 인공지능 >

정병옥
2007년 대전대학교 컴퓨터공학 석사
2006년-2016년 ㈜디지캡
2016년-현재 ㈜엘에스웨어

<주관심분야: 클라우드 보안 서비스, 빅데이터, 응용 보안>

신동명
2003년 대전대학교 컴퓨터공학과 박사
2001년 - 2006년 한국정보보호진흥원 응용기술팀 선임연구원
2006년 - 2014년 한국저작권위원회 저작권기술팀 팀장
2014년 - 2016년 한국스마트그리드사업단 보안인증팀 팀장
2016년 - 현재 엘에스웨어(주) 연구소장/이사
<주관심분야 : 오픈소스 라이선스, 시스템/네트워크보안, SG인증/보안, SW취약점 분석・감정>