オープンデータの安全な利活用のための典拠情報に基づくリスク管理手法

ラミレス カセレス ギジェルモ オラシオ† 是津 耕司†

†情報通信研究機構
619-0289 京都府相楽郡精華町光台 3 丁目 5 番地
E-mail: {ramirez.caceres, zettsu}@nict.go.jp

あらまし 今日、ネット上には多種多様なオープンデータが増え続けているが、これらのオープンデータを正しく利用するには、誰がどのようにしてデータを作成し公開したのかという典拠情報を把握することが重要である。我々は、オープンデータの典拠情報を取集・分析し、データの品質や信頼性、使用条件などデータ利活用の安全性を評価する方法について研究を行っている。本論文で提案するセキュリティリスク管理手法では、データの典拠情報をPROV形式で構造化し、提供者・利用者間での使用権限の違反、データの不整合、不完全な組合せなど7種類100項目以上に渡るセキュリティルールを知識ベース化しセキュリティリスクを自動検出するとともに、リスク分析の結果を分かりやすく可視化することを可能にしている。

Provenance-based Risk Management for Secure Use of Open Data

Guillermo Horacio Ramirez Caceres† Koji Zettsu†

†National Institute of Information and Communications Technology.
3-5, Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, JAPAN
E-mail: {ramirez.caceres, zettsu}@nict.go.jp

Abstract Nowadays, large-scale and heterogeneous open data is continuously increasing on the web. However, in order to use correctly this data, it is very important to possess information related to the provenance, such as how the data was created and published. In this research, to allow the assessment of the safety of use of the data, i.e., the data quality, trustworthiness, and the appropriate conditions for use, we are collecting and analyzing provenance of the open data. We propose a security risk management method: First, to manage the provenance information, we structure the data based on PROV. Then, in order to detect a violation of condition of use between providers and user or data inconsistency, we have developed an automatic risk detection system that uses a security knowledge base including more than 100 security rules grouped in 7 categories of risk. Finally, we provide a graphical representation of the risk result that allows users to graphically see where and what kind of data generates security conflicts.

1 Introduction

An open data, it would be a kind data that can be redistributed, reused and can be used freely [1]. Nowadays, large-scale and heterogeneous open data is continuously increasing on the web [2]. With the rapid growth of linked data on the web more and more application emerges that make use of this data. These organized information that is valuable and easily accessible to those who need it, are called information asset.

Daily many users rely on data from the Web, but often it is difficult or impossible to determine where it came from? How it was produced? Or the end users are allowed to use or re-use the new information assets?. Therefore, in order to use correctly this data, it is very important to possess information related to the provenance, such as how the data was created and published.

In this research, to allow the assessment of the safety of use of the data, i.e., the data quality, trustworthiness, and the appropriate conditions for use, we are collecting and analyzing provenance of the open data.

In this paper, we propose a security risk management method: First, to manage the provenance information, we structure the data based on W3C PROV [3]. Then, in order to detect a violation of condition of use between providers and user or data inconsistency, we have developed an automatic risk detection system that uses a security knowledge base including more than 100 security rules grouped in 7 categories of risk. Finally, we provide a graphical representation of the risk result that allows users to graphically see where and what kind of data generates security conflicts.

This paper is organized as follows. In Section 2, we briefly review some related work. In Section 3, we explain a use case scenario. In Section 4 and 5, we explained concept of provenance and how we implemented provenance in the risk management process. In section 6, describe a prototype developed and the result of the implementation. Finally, in Section 7 we conclude the paper and present some future works.

2 Background and Related Works

Security and data provenance are mutual related. Good security leads to detailed provenance, and good provenance lets users make good security decisions [4].

In the open data security scope we can find several risks like: license is not open enough, heterogeneous licenses across datasets, data quality, data available in heterogeneous formats, incomplete metadata, the language barrier, etc. [5].

ISO 31000 defines risk as the “effect of uncertainty on objectives” [6]. In most cases these effects are negative, but positive effects are possible.

Different methodologies exist for risk assessment, some of which are discussed in ISO/IEC 31010 and ISO/IEC 27005 [7] [8].

Tools to support provenance are continuously being developed. There appear to be two basic approaches to describing the objectives to which provenance records refer: coarse-grained and fine-grained. The Open Provenance Model (OPM) and most of the provenance vocabularies studied in Provenance adopt the coarse-grained model [9].

Attribution is important for making data

citable and to ensure that creators of data receive credit to offset the extra effort required to publish data [10]. There are some proposals for fine-grained citation of data in databases and for versioning that can begin to address this [11].

Much of the focus of work on provenance in workflows and distributed systems has been recording processing steps, the precise parameters used, and any other metadata considered important for ensuring repeatability of electronic experiments [12]. Some other works are focus on the provenance of scientific data processing [13]. The motivation of this paper is similar, but we are targeted on Open Data. Based on the semantic gap between data providers and user, we are working to detect security problems that affect the data utilization, like violation of condition of use between providers and user.

The Dublin Core Metadata Initiative (DCMI) is the current set of the Dublin Core vocabulary [14]. DCMI terms, tell us what was affected. In addition, according to the W3C PROV guideline, many DC terms can be used to describe provenance information about a resource: when it was affected in the past, who affected it and how it was affected. However, there is no direct information in DC describing where and why a resource was affected [15].

3 Motivating Scenario

Large-scale massive heterogeneous open data have been accumulated in various fields of scientific research and society. However, much open data emerges from activities that are different in purpose, contexts, and time from its eventual use.

Figure 1 Motivating scenario

The dataset are collected in different way by different organizations from different providers and location. Therefore, dataset may have different time frame, geographic units or other essential characteristics.

Many uses of the web involve the combination of data from diverse sources. User can access to a wide variety of open data like Open Science Data, Open Government Data, Open sensor data, and Open social data (Fig. 1).

As a result, the discovery of new knowledge by linking the big variety of data, sensing data and science data has been increasing [16]. However, data can only be meaningfully reused if the collection processes are exposed to users. This enables the assessment of the context in which the data was created, its quality and validity, and the appropriate conditions for use.

The information assets living on the Internet often change containers or move through a process that creates new information. Actual risk management approaches do not consider such dynamic scenarios.

The main objective of this research is to implement a provenance-base risk management to support secure leverage of distributed open data and to be able to analyze security of big data. Provenance information can help to provide a safe

Table 1 Provenance and Dublin Core mapping

Category	Name			
What	dct.Abstract	dct.AccrualMethod	dct.AccrualPeriodicity	dct.AccrualPolicy
	dct.Alternative	dct.Audience	dct.BibliographicCitation	dct.ConformsTo
	dct.Coverage	dct.description	dct.EducationLevel	dct.Extent
	dct.Format	dct.HasPart	dct.identifier	dct.InstructionalMethod
	dct.IsPartOf	dct.IsRequiredBy	dct.language	dct.Mediator
	dct.Medium	dct.Relation	dct.Requires	dct.subject
	dct.TableOfContents	dct.title	dct.type	(*)Type.NIST800-60
Who	dct.Contributor	dct.creator	dct.publisher	dct.RightsHolder
When	dct.Available	dct.created	dct.Date	dct.DateAccepted
	dct.DateCopyrighted	dct.DateSubmitted	dct.Issued	dct.Modified
	dct.temporal	dct.Valid		
Where	dct.spatial			
How	dct.accessRights	dct.HasFormat	dct.HasVersion	dct.IsFormatOf
	dct.IsReferencedBy	dct.IsReplacedBy	dct.IsVersionOf	dct.license
	dct.References	dct.Replaces	dct.rights	dct.source

service to use and re-use the information asset according to the creator requirements and to transfer the license statement to the final user. By using provenance representation, we can find the origin of the information asset, when was created and by who, then is possible to provide a clear and trust information at time.

4 Provenance-Based

4.1 Metadata Management

In our proposed model, we are working with a lot of information asset from different providers. Metadata is used to represent properties of information assets. Many of those properties have to do with provenance.

Table 2 Metadata mapping example

what	dcterms.title	carbohydrates in particulate matter of the northwest pacific
	dcterms.description	Patterns of distribution and variations of group and monosaccharide…
	dcterm.language	en
who	dcterm.creator	Vladlen E, Artem'ev
	dcterm.publisher	PANGAEA
when	dcterm.created	1973-11-22
where	dcterm.spatial	[[130.87, 33.778], [160.55, 33.778], [160.55, 18.86], [130.87, 18.86], [130.87, 33.778]]
how	dcterm.right	CC-BY

In order to manage the information assets we are implemented an assets classification by using Dublin Core (DC) for describing the metadata and a guidance to manage provenance (PROV) proposed by the provenance working group of W3C.

Table 1 describes a list of the DC term using to describe an information asset according to these five categories (what?, who?, when?, where?, and how?).

We employ a simple illustrative example from an existing Open science data in Table 2.

What contains all the terms describing a resource without referring to its provenance. *Who* contains agent related terms. *When* contains date and time related terms. *Where* define spatial information of the resource. This can be considered special regarding their relevance for search specific dataset in spatial-temporal environment. Finally, *How* contains derivation related terms. When a resource is derived from other resources, the original resource becomes part of the provenance chain of the derived resource. Finally, licensing, rights and their access are considered part of the

provenance of the resource as well, since they restrict and explain how the resource can be used for further derivation.

4.2 Provenance Representation

To provide a graphic representation of provenance information, we implemented the Open Provenance Model (OPM) [17], which defines three main entities in a provenance record: circles are artifact, rectangles are process, and octagons are agents. An agent is an entity capable of performing a process, an artifact is an immutable piece of a state, and a process is a series of actions that use artifacts to generate new artifacts.

As shown in figure 2, the entities are related though a number of properties like: *wasGeneratedBy*, *used*, *wasControledBy*, *wasTriggeredBy*, and *wasDerivedFrom*. Edge labels are in the past to express that these are used to describe past executions.

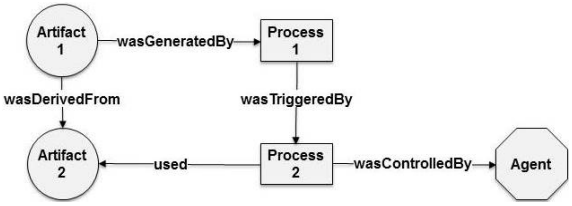

Figure 2 OPM Overview

Information asset used, processes performed, entities that perform these processes and any new information asset generated is captured and represented based on the OPM. Figure 3 shows how the metadata described in Table 2 is mapping to the OPM graph.

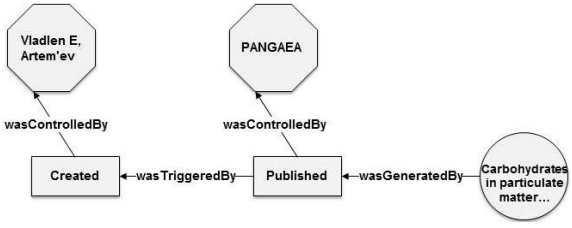

Figure 3 Mapping Dublin Core to OPM graph

5 Risk Management Model

In this research we propose a new approach to improve the risk management to be able to analyzing the security in the Big Data environment.

Our proposed risk assessment is based on the provenance information including in the metadata of the information asset. Then we are using the graphical representation of the provenance information to analyze possible security issues by accessing the security knowledge base.

5.1 Risk Identification

Based on the OPM graph data and the information asset attributes, we implement a risk assessment diagnostic that performs rule-based risk detection and returned the result of risk detected.

In order to analyze the security risk in large OPM graph, we define small subset of provenance, the patterns for verification fall into two groups (Fig 4)

- Group 1: is a pattern to carry out checks if it is constituted by one information asset (A), one agent (AG), and one process (P).
- Group 2: is a pattern to carry out checks if it is constituted by two information assets (A), one agent (AG), and one process (P).

For example, if we consider the language barrier issues. When a user tries to combine two types of data with different

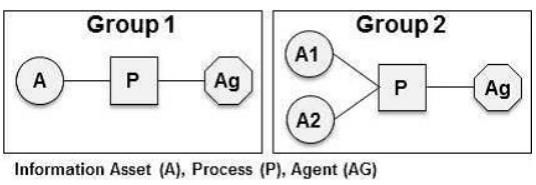

Figure 4 Security rules patterns

language can generate some conflicts that might be directly related to the confidentiality, integrity and availability of the data (Group 1). In addition, if there are language difference between dataset and end users, may cause availability issues (Group 2).

The security rules was created based on the attributes of each information asset (IA), the process that affect the information asset (P), and the user who want to use the asset (AG). The security knowledge base includes more than 100 security rules. This rule file consists of two parts: Rule condition that performs risk detection, and risk definition that return the result of risk detection. An example of the rule file is described in Fig. 5.

```
conds:
  group="1"
  and process["accion"]="any"
  and asset["dcterms.language"]!=agent["dcterms.language"]

risks:
  threat["riskId"]="rule base No. 9";
  threat["riskDescription"]=asset1["Name"]+
  "("+asset1["dcterms.language"]+")"+
  "and"+agent["Name"]+"("+agent["dcterms.language"]+")"+
  "Risk detected (If there are language difference between
  dataset and users, may cause availability issues.)
  threat["securityCategory"]="Availability";
  threat["riskScore"] = makeScore(0,0,1,0,0,0,0);
```

Figure 5 Rule file example

5.2 Risk analysis and Visualization

Risk analysis comprehends the nature of risk and determines its level. Risk analysis involves consideration of the causes and sources of risk, their consequences and the probability that those consequences can occur.

In addition to the conventional approach of security requirements that include confidentiality, integrity, and availability, based on the provenance information, we include other requirement to support trust, compliance, and completeness.

Based on the above security risk identification, we implemented function makeScore(co, in, av, au, po, us, pr). As shown in Fig. 5, each security rule file includes this function to describe the risks nature. It is a function of the score generation for each risk category and includes four values: 0 = not applicable, 1 = Low, 2 = Medium, 3 = High.

For each process that uses one or more information asset, we can identify several security rules. For example, as shows in left part of fig. 6, a user (AG) wants to join two assets (A1, A2) to create a new asset (A3). When checking the rules for the entire information asset related to the process, we can identify 6 security rules.

In risk analysis, the makescore function collects the value of each security rule to describe generically the risks nature. For example, as shown in Fig 6, makeScore (2,2,3,0,1,2,1) produces the following results: Confidentiality: 2, Integrity: 2, Availability: 3, Authenticity: 0, Possession: 1, Use: 2, and Provenance: 1.

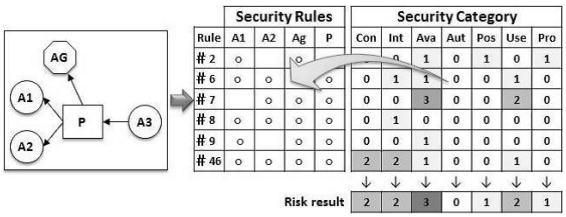

Figure 6 Mapping risk impact to artifact

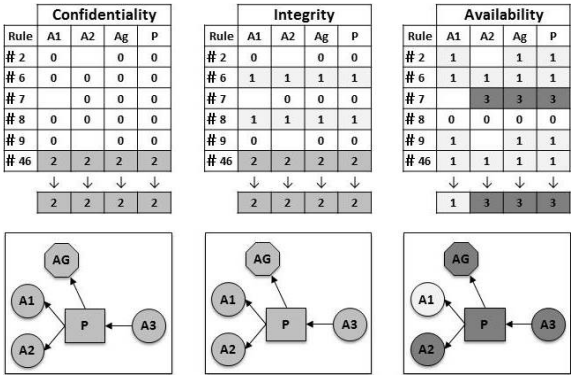

Figure 7 Transfer risk impact to OPM

Next, the value of each security category is transfer to the entities used in the rule (Artifact, Process, and Agent). Finally, as shown in fig. 7, the result of the risk is mapping to the OPM graph, then users are allowed to see graphically where and what kinds of data generated security conflicts.

6 Implementation and Evaluation

The National Institute of Information and Communication Technology, Japan (NICT) has a large-scale web archive that contains about four million documents.

Our proposed risk assessment framework works as a web application. The OPM generator provides a graphic representation of the provenance information by access user and accesses the profile information. Then the risk assessment diagnostic implements a risk assessment by accessing the OPM graph data and checking the security rule files. Finally, the renderer function provides the risk assessment results to users by accessing the security knowledge base.

To evaluate our proposed system, we verified the possible security risks based on the provenance information of about 349,298 information assets from the data publisher for environmental science (PANGAEA) [18]. Table 3 show the metadata collected and registered on our proposed system.

Based on the metadata attributes collected from the dataset, we implementing the risk diagnostic. We can identify security risks in some of this dataset as show in table 4. We found 55 dataset having some security issues related to confidentiality or integrity, and 23176 dataset with security issues related to availability.

Table 3 Metadata collected

Metadata	pangaea	%
dct.Publisher	349298	100%
dctIdentifier	349298	100%
dct.Language	349298	100%
dct.Title	349298	100%
dct.Source	349298	100%
dct.Description	9391	3%
dct.Creator	337361	97%
dct.Created	349298	100%
dct.Subject	349298	100%
dct.Spatial	346607	99%
dct.Rights	281548	81%
dct.Temporal	230643	66%

Table 4 Risk analysis result

Risk Category	pangaea	%
Confidentiality	55	0.02%
Integrity	55	0.02%
Availability	23176	6.46%

7 Conclusion

In this paper, we propose a new approach to implement risk assessment based on provenance information. In this research, we use provenance to support security risk assessment. Furthermore, use security risk assessment to improve provenance. Together with our security risk assessment system, we can provide a remarkably complete and rigorous record of everything and everyone who has interacted with your data in any way, be it access you have made or actions we have taken to ensure integrity and authenticity.

By using provenance, we can find the origin of the information asset, when was created and by who, then is possible to provide a clear and trust information at time. In addition, our proposed system implements a graphic representation of risk result using provenance graph, allowing users to see graphically where and what kinds of data generated security conflicts. The provenance-based risk assessment, allow users to analyze the security issues of information asset in a big data environment. As future work, after

identified the risk that affect the dataset, the next step is to implement a risk treatment to reduce the risk to an acceptable level. Risk reduction is the countering or elimination of security risks by the selection, application and assessment of security controls.

References

[1] Cuzzocrea A., et al.: Analytics over large-scale multidimensional data: the big data revolution! *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP, ACM*, pp. 101-104. (2011)

[2] The Fourth Paradigm: Data-Itensive Scientific Discovery. *Microsoft research*, ISBN 978-0-9825441-0-4 (2009)

[3] The Provenance Working Group of the World Wide Web Consortium. (W3C) (Online) available from < http://www.w3.org/2011/prov/wiki/MainPage (accessed 2014/08/15)

[4] Patrick McDaniel, Sean W. Smith: Data Provenance and Security. *IEEE Security and Privacy.* pp.83-85. (Mar. 2011)

[5] Sébastien Martin, et al. Risk Analysis to Overcome Barriers to Open Data. *The Electronic Journal of e-Government (EJEG)* (2013)

[6] ISO 31000:2009 - Risk management - Principles and guidelines. (2009)

[7] ISO/IEC 31010:2009, Risk management - Risk assessment techniques. (2009)

[8] ISO/IEC 27005:2011, Information technology Security technology Information security risk management. (2011)

[9] Moreau L, et al.: The open provenance model core specification (v1.1). *Future Generation Computer Systems.* Volume 27, Issue 6, pp.743-756. (2011)

[10] Buneman P.: How to cite curated databases and how to make them citable. *Proceedings of the 18th International Conference on Scientific and Statistical Database Management.* pp. 195-203. (2006)

[11] Oinn T, et al.: Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, pp. 3045-54. (2004)

[12] Simmhan Y, et al.: A survey of data provenance in e-science. *SIGMOD* Record. pp. 31-36. (2005)

[13] Bose R. et al.: Lineage Retrieval for Scientific Data Processing: A Survey. *ACM Comput. Surv.* Volume 37, Issue 1, pp.1-28. (2005)

[14] Dublin Core Metadata Initiative (DCMI) (Online) available from http://dublincore.org/ (accessed 2014/08/15)

[15] Buneman P., Khanna S. and Tan WC. Why and Where: A Characterization of Data Provenance. *Proceedings of the 8th International Conference on Database Theory.* pp.316-330. (2001)

[16] Groth P., et al.: Requirements for Provenance on the Web. *The International Journal of Digital Curation*, Volume 7, Issue 1, pp.39-56. (2012)

[17] The Open Provenance Model (Online) available from < http://openprovenance.org/> (accessed 2014/08/15)

[18] Data Publisher for Earth & Environmental Science (PANGAEA) (online) available from <http://www.pangaea.de/> (accessed 2014/08/15)