Journal of Information Processing Vol.22 No.4 617–625 (Oct. 2014)

[DOI: 10.2197/ipsjjip.22.617]

Regular Paper

Provenance-Based Security Risk Assessment Framework

Guillermo Horacio Ramirez Caceres[1,a)] Koji Zettsu[1,b)]

Received: November 29, 2013, Accepted: June 17, 2014

Abstract: Large-scale massive heterogeneous data have been accumulated in various fields of scientific research and society. As a result, discovering new knowledge by linking sensing and science data, such as web archives, has attracted attention. We developed a Knowledge Language Grid (KLG) system that combines multiple asset data from different providers and allows users to use or re-use them. KLG structures a great quantity of information that can be confidential for individuals, companies, or institutions, but it can also be misused or disclosed to inappropriate people. In this paper, we propose a risk assessment framework based on provenance information. In addition, since KLG allows user to access security knowledge-bases, it is possible to provide actual and on time information about risk and security controls. Our proposed system implements a graphic representation of provenance using Open Provenance Model (OPM), and users are allowed to see graphically where and what kinds of data generate security conflicts.

Keywords: risk assessment, provenance, big data, OPM

1. Introduction

Owing to the advance of broadband mobile communications and the Internet, many users share resources on their personal networks and connect to them to enter a world of information. Large-scale massive heterogeneous data have been accumulated in various fields of scientific research and society [1]. By linking such huge amounts of data, including sensing and science data, the discovery of new knowledge is increasing [2].

Data curation enables data discovery and retrieval, maintains quality, adds value, and provides for re-use over time. This new field includes authentication, archiving, management, preservation, retrieval, and representation [3]. In the field of disaster prevention, the rapid analysis of disaster information is especially expected [4].

Such organized information, which is valuable and easily accessible to those who need it, is called information assets. Many approaches have been proposed to handle IT security issues, and many international standards have been developed in this field [5]. However, traditional security mechanisms, which are tailored to secure static data, are insufficient, and many security issues related to privacy and copyright are difficult to manage [6], [7]. Therefore, our proposed approach supports the secure leverage of information to cover such security issues as copyright and intellectual property.

In this paper, we propose a provenance-based security risk assessment framework that leverages such information assets as science and sensing data. Provenance refers to the chronology of the ownership, custody, or location of a historical object [8]. The provenance of information determines whether it can be trusted to identify a problem caused by invalid output and to credit originators when reusing it. Provenance in data curation refers to such information sources as data and programs involved in producing a new dataset [9].

We developed a Knowledge Language Grid (KLG) system, which often combines multiple asset data from different providers and allows users to use or re-use information assets. KLG structures a great quantity of information that can be confidential for individuals, companies, or institutions and can be misused or disclosed to inappropriate people. Awareness of the security risks that affect information assets is critical to protect intellectual property and privacy [10].

Our proposed prototype implements a graphic representation of provenance. Users are allowed to see graphically where and what kinds of data generate security conflicts. Using provenance representation, we can find the origin of information assets, when they were created and by who to provide clear and reliable information in time.

This paper is organized as follows. In Section 2, we briefly review related work to support risk assessment to clarify the different motivations between previous works and ours. In Section 3, we explain several use case scenarios of provenance-based risk assessment. In Section 4, we explain the concept of provenance and how we implemented provenance in the risk assessment process. Section 5 describes our prototype that was developed in an information services platform (ISP) laboratory and the implementation results of data curation. Finally, in Section 6 we conclude our paper and present future works.

2. Related Works

Security in computer science embraces two concepts: physical and logical [11]. Physical security refers to the protection of hardware and the support of data and the buildings and facilities

[1] Universal Communication Research Institute, National Institute of Information and Communications Technology (NICT), Soraku, Kyoto 619–0289, Japan

[a)] ramirez.caceres@nict.go.jp

[b)] zettsu@nict.go.jp

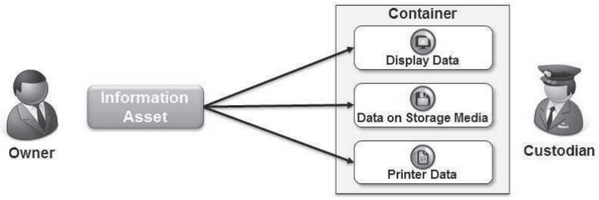


Fig. 1 Information assets.

that harbor them, including fire, sabotage, robberies, and natural disasters. Logical security refers to the secure use of software, the protection of data, processes, and programs and the orderly and authorized access of users to information.

Networks must satisfy the following requirements or characteristics to maintain physical security [12]:

- Confidentiality: its loss is the unauthorized disclosure of information.
- Integrity: its loss is the unauthorized modification or the destruction of information.
- Availability: its loss disrupts access to information, its use, or access to information systems.

For logical security, in our case, protecting information assets is more complicated. **Figure 1** shows the security issues of information assets.

The owners or creators of information assets have the primary responsibility for their viability and survivability. An information asset "lives or exists" in a container, where it is stored, transported, or processed. Custodians or administrators have the responsibility to protect an information asset as it is stored, transported, or processed.

Organizations operate in an uncertain world. Every project or activity has certain risks, but what is the risk? ISO 31000 defines risk as the "effect of uncertainty on objectives" [13]. In most cases these effects are negative, but positive effects are possible.

ISO/IEC 27001, also called the Information Security Management System (ISMS), is an international standard for initiating, implementing, maintaining, and improving information security management in organizations [14]. These standards are used by a broad range of organizations in most commercial and industrial market sectors: finance and insurance, telecommunications, utilities, retail and manufacturing sectors, various service industries, transportation sectors, and governments.

ISO/IEC 27002 provides guidance about the implementation of security control policies [15]. However, risk analysis and risk assessment, both of which are necessary for describing the environment of security control policies, are outside ISMS's scope.

Basically, to effectively manage risk, we must identify and assess the threats to assets to determine the vulnerability of critical assets to determine risks and identify methods to reduce and prioritize them [16].

Unfortunately, many organizations are unsure how or even where to deploy their scarce resources to protect their information assets. The steadily increasing technical and environmental complexity of global networks presents a significant obstacle. In addition, the list of information security vulnerabilities and threats continues to grow to which organizations are constantly subjected.

Vulnerability-centric approaches can also be implemented [17]. However, the existence of a significant vulnerability does not mean that an organization faces a significant risk. This distinction is important because assets and their value to an organization determine the context for risk rather than the vulnerability itself. To consider the importance of vulnerabilities in an organization, the National Infrastructure Advisory Council (NIAP) proposed a Common Vulnerability Scoring System (CVSS) [18].

A significant amount of guidance, including FIPS Publication 199 [19] and NIST Special Publication 800-60 [20], has been issued to help federal government agencies value their information assets.

Different methodologies exist for risk assessment, some of which are discussed in ISO/IEC 27005 [21], [22]. Therefore, the implementation of a secure system generally consumes a large amount of time and resources and requires much knowledge [23].

ISO/IEC 27001 specifies that the controls implemented within the scope, boundaries, and the context of ISMS need to be risk-based. The application of an information security risk management process can satisfy this requirement.

According to such risk assessment, to identify risks, we must know the asset, the threats to it, the vulnerabilities that might be exploited by those threats, and finally the impact of damage to its confidentiality, its integrity, and its availability.

Many available commercial and governmental risk assessment methodologies contain these basic activities, for example, the Operationally Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) information security risk assessment methodology [24].

However, as explained above, to protect information assets, we must consider confidentiality, integrity, and availability. Other important requirements include data authenticity, data possession, data accessibility, and data provenance.

3. Motivation Scenario

In this section, we illustrate how provenance-based security risk assessment can provide improved security requirements to protect information assets.

The information assets on the Internet often change containers or move through a process that creates new information. For example, the data from two different sources are sometimes combined to create a new information asset.

For security purposes, we ask the following questions: Who owns the new information asset? Were its security requirements transferred correctly? What are its security requirements? Are end users allowed to use or re-use it?

The following scenarios were based on three basic processes: *search*, *collect*, or *join*. These processes create a new information asset from different providers.

3.1 Use Case Scenario: "Collect"

In this scenario, we focus on KLG's "Collect" process (**Fig. 2**). For example, emergency responses involve the immediate actions taken to respond to a disaster. Government users access KLG and collect information from different providers.

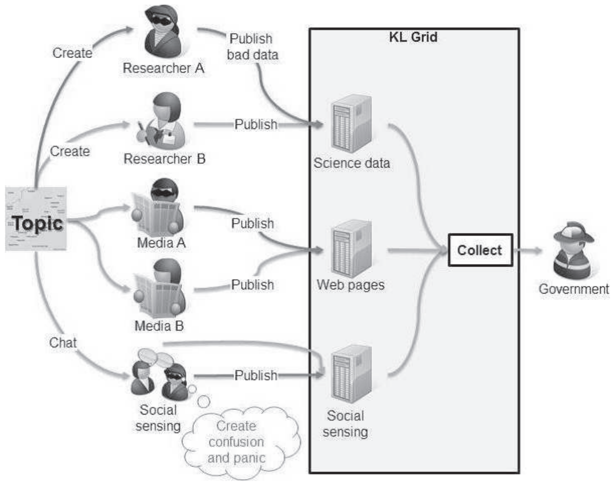


Fig. 2 Use case scenario: "Collect".

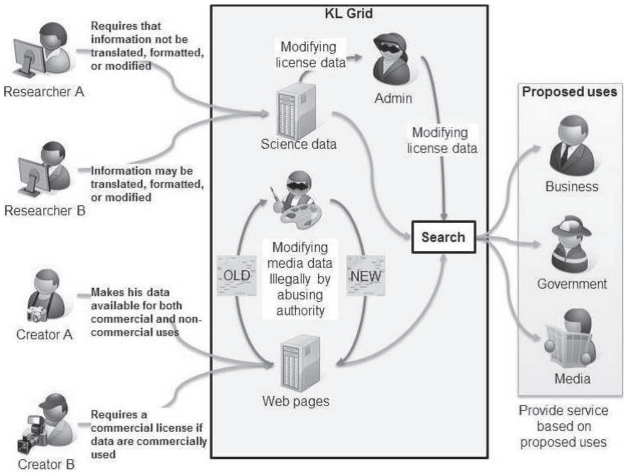

Requires that information not be translated, formatted, or modified
Researcher A
Information may be translated, formatted, or modified
Researcher B
Makes his data available for both commercial and non-commercial uses
Creator A
Requires a commercial license if data are commercially used
Creator B
KL Grid
Modifying license data
Admin
Science data
Modifying license data
Modifying media data Illegally by abusing authority
OLD
NEW
Web pages
Search
Proposed uses
Business
Government
Media
Provide service based on proposed uses

Fig. 3 Use case scenario: "Search".

An information asset from *Researcher B* might be objective and useful, but not the information from *Researcher A*. Another case is the information generated for media companies. This information might be objective and centered on the *topic*. However, some media companies pretend to describe the topic from a commercial point of view.

Information collected from social media, like Facebook or Twitter, is sometimes anonymous, and malicious users might abuse this application to create confusion or cause a panic.

The consequences of unauthorized modification or the destruction of emergency response information usually depends on whether the information is time-critical. Data supporting emergency responses may be available, and delays are not tolerated.

Using provenance representation, we can find the origin of an information asset, when it was created and by whom, and then we can timely provide clear and reliable information.

3.2 Use Case Scenario: "Search"

The next scenario shows how provenance information helps protect the intellectual property of information assets. As shown in **Fig. 3**, we are working with research information and such creators of media contents as photographer.

Research and development involve the gathering and analysis of data, the dissemination of results, and the development of new products, methodologies, and ideas.

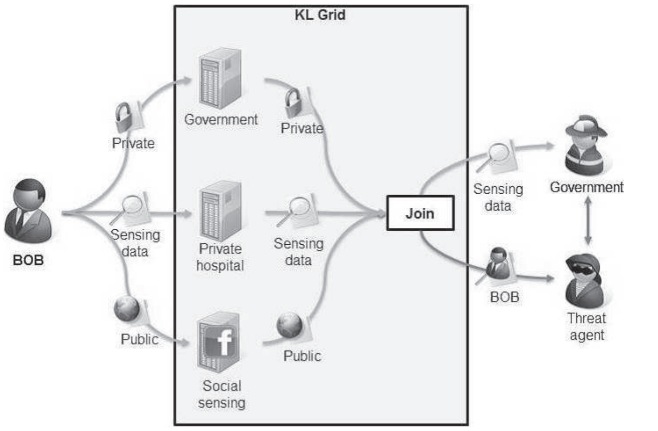


Fig. 4 Use case scenario: "Join".

Most research and development information is proprietary. Unauthorized disclosure of such information violates regulations. Pre-publication or other unauthorized disclosure of research findings can have seriously adverse effects.

For example, *Researcher A*, who owns the information asset, forbids it from being translated, formatted or modified. But the information created by *Researcher B* is allowed to be translated, formatted, or modified.

In the case of a media creator, *Creator A* makes his data available for both commercial and non-commercial uses. However, *Creator B* requires a special commercial license if his data are commercially used.

Some information assets might be modified or the original license requirement could be changed, accidentally or not.

Provenance information can help provide a safe service to use and re-use information assets based on creator requirements and to transfer the license statement to the final user.

3.3 Use Case Scenario: "Join"

The last scenario describes how to meet privacy issues by implementing a provenance-based risk assessment.

As explained in Section 1, to provide a service, multiple asset data are often combined from different providers, allowing users to use or re-use the information.

In our case study (**Fig. 4**), suppose that some privacy information for user *BOB* is stored on government servers, including personal information about city hall matters. Bob also stores some sensing information, like medical records, on a public hospital server. He is also very active on social networks like Facebook where he often shares pictures and information.

Imagine that a government user wants to implement a new service for a community to know more about its needs. He accesses KLG and adds information from a specific area, including government records, hospital records, and information from social networks.

The information generated by KLG might include some sensing data, and a user's private information could be disclosed after combining various records that a threatening agent can use to identify such specify users as *BOB*.

With a provenance-based risk assessment, we can identify possible violations of privacy disclosures at the moment of implementation in the join process on KLG. Then we can reclassify the information asset and inform the final user about the risk to use or re-use it.

4. Provenance-based Risk Assessment Framework

The main objective of this research is to implement a provenance-based security risk assessment to support the secure leverage of information assets.

Our proposed risk assessment combines multiple asset data from different providers and allows users to use or re-use the information. After implementing a graphical representation of provenance, users can see graphically where and what kinds of data generate security conflicts. Finally, allowing users to access security knowledge-bases, we can provide actual and timely information about risks.

Figure 5 shows the security risk assessment concept of our research. The owners of assets analyze the possible threats to determine which apply to their environment. The results are called risks. This analysis supports the selection of countermeasures or security controls to counter risks and reduce them to an acceptable level.

Safeguarding the information assets of interest is the responsibility of owners (providers) who value them. Owners set the security requirements for information assets and communicate them to asset custodians, who must meet the owner's security requirements by implementing appropriate security controls on the container where the asset is stored. Security controls are imposed to reduce vulnerability. Residual vulnerability might remain after the imposition of security controls. Such vulnerability may be exploited by threatening agents who represent the residual level of risk to the assets. Owners will seek to minimize the risk given to other constraints.

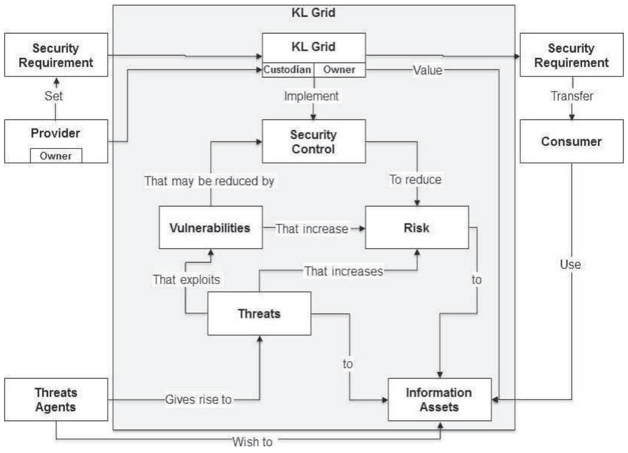


Fig. 5 Provenance-based security risk assessment.

Actual or presumed threatening agents may also value the assets and seek to abuse them in a manner contrary to the owner's interests. Owners will probably realize that such threats might damage their assets and reduce their value. Specific security impairment commonly includes (without being limited to) damaging disclosure of the asset to unauthorized recipients (loss of confidentiality), damage to the asset through unauthorized modification (loss of integrity), or unauthorized deprivation of access to the asset (loss of availability).

Finally, the custodian communicates the necessary security requirements to the system user (consumer) who wants to access a specific information asset.

4.1 Provenance-based

One central aspect on which this research is based is the provenance management of each information asset. In our proposed model, we deal with many information assets from different providers and manage them by asset classification using Dublin Core (DC) for describing metadata [25] and a guidance to manage provenance (PROV) proposed by the provenance working group of W3C [8]. **Table 1** lists the DC terms that describe an information asset.

4.1.1 Open Provenance Model

To provide a graphic representation of provenance information, we implemented the Open Provenance Model (OPM) [26], which defines three main entities in a provenance record: *Agent*, *Artifact*, and *Process*. 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 **Fig. 6**, the entities are related by a number of properties: *used*, *wasGeneratedBy*, *wasControledBy*, *wasTrig-*

Table 1 Provenance and Dublin Core mapping.

Category	Name	Description
What	dcterms.identifier	unambiguous reference to the resource within a given context.
	dcterms.description	account of the resource.
	dcterms.title	name given to the resource.
	dcterms.language	language of the resource.
	dcterms.subject	resource's topic.
	dcterms.type	resource's nature or genre.
Who	dcterms.creator	entity primarily responsible for making the resource.
	dcterms.publisher	entity responsible for making the resource available.
When	dcterms.created	date of resource's creation.
	dcterms.temporal	resource's temporal characteristics.
Where	dcterms.spatial	resource's spatial characteristics.
How	dcterms.accessRights	information about who can access the resource or an indication of its security status.
	dcterms.license	legal document giving official permission to do something with the resource.
	dcterms.license.cc	license based on Creative Commons (CC)
	dcterms.license.odrl	license based on Open Digital Rights Language (ODRL)
	dcterms.rights	information about rights held in and over the resource.
	dcterms.source	related resource from which the described resource is derived.

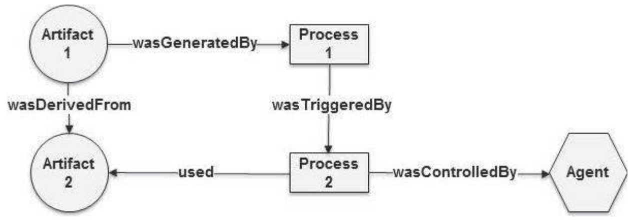


Fig. 6 OPM overview.

geredBy, and *wasDerivedFrom*. The edge labels are in the past tense because they describe past executions.

Process 2 is controlled by an Agent, which is represented using the *wasControlledBy* property. A process can be controlled by multiple agents.

wasTriggeredBy defines a relationship among processes, when a process is made operational by another process.

The relations between process and artifact are represented by the *used* and *wasGeneratedBy* properties. For example, Artifact 2 was *used* in Process 2, and Artifact 1 was created by Process 1 (Fig. 6).

Finally, the *wasDerivedFrom* property specifies that one artifact was derived from another. Artifact 2 was derived from Artifact 1.

4.2 Risk Assessment

As explained in the previous section, to find, identify, and describe the risk, we must identify the asset, the threats that affect it, and the vulnerabilities that might be exploited by the threats and their impact. In other words, we must obtain deep knowledge about attacks and how they are implemented.

We are working on risk assessment based on the risk management described in ISO 31000. As explained in Section 2, risk assessment includes the following three activities:

- risk identification
- risk analysis
- risk evaluation

4.2.1 Risk Identification

Next we look at the process of finding, recognizing, and describing risks. Based on the OPM graph data and the information asset attributes, we implemented a risk assessment diagnostic service that performs risk detection and returns its result.

The risk assessment purpose in this research provides a framework to support rule-based, risk detection engine implementation.

We created security rules based on the attributes of each information asset (IA), the process that affects it (P), and the user who uses it (A). Our proposed model currently includes 106 security rules. Since they were created based on the metadata attributes of information assets, security rules can be applied to any new information asset added to the system. In addition, our proposed prototype includes a web-based interface and allows system administrators to edit or register a new security rule to cover future security issues.

We implemented a risk diagnostic algorithm based on OPM graph data to carry out verifications on the basis of Process, which is an algorithm comprised of the relationship among Information Asset (IA), Agent (A), and Process (P). The patterns for verification fall into four groups (**Fig. 7**):

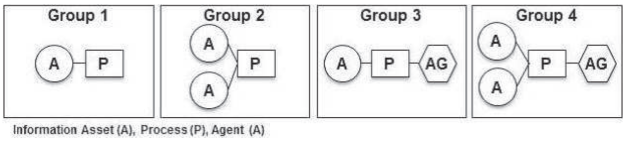


Fig. 7 Security rule patterns.

```
conds:
 (group == "1") && isNull(asset1["dcterms.identifier"])

risks:
 threat["riskId"] = "rule based check result No.1";
 threat["riskDescription"] = asset1["Name"] + "(" + asset1
 ["Id"] + ")" + "," + "risk detected";
 threat["securityCategory"] = "Possession, Provenance,";
 threat["riskScore"] = makeScore(0,0,0,0,1,0,1);
 threat["riskComment"] = "if the ID is null is impossible to
 provide asset Possession & Provenance) Id:1 ";
```

Fig. 8 Rule-based example.

- Group 1: carries out verifications if it is composed of one information asset (IA) and one process (P).
- Group 2: carries out verifications if it is composed of two information assets (IA) and one process (P).
- Group 3: carries out verifications if it is constituted by one agent (A), one information asset (IA), and one process (P).
- Group 4: carries out verifications if it is constituted by one agent (A), two information assets (IA), and one process (P).

Rule-based, risk verification is implemented by the risk inspector function. Verifying risks is based on the rules described in the rules file. An example of a rule file is described in **Fig. 8**.

As shown in Fig. 8, this rule file consists of two parts: rule condition and risk definition. The description of the conditions resembles the description of JavaScript, whose regular expressions are also available. The risk definitions are shown for the X conditions. Description complies with JavaScript's syntax. By setting information that corresponds to the threat, risks are reflected as detection risks.

Figure 9 shows an OPM graph example of the risks of diagnostic transactions. The rule check was implemented in two steps based on the process. **Table 2** summarizes the security rules found in the example.

In the first step, the OPM graph verifies the rule by focusing on P1 (PROCESS). For P1 we can find two combinations for group 1 and one for group 2. One agent is related with the process, and we can also find two combinations for group 3 and one for group 4.

Next we implemented verification by focusing on P2 (PROCESS) and checked the rules for all the information assets related to the process. However, since *A3* is composed of *A1* and *A2* and inheriting attributes *A3(A1)* and *A3(A2)*, we must conduct additional combinations for P2. Since there is no *Agent* for P2, we did not do any rule verifications for groups 3 and 4.

4.2.2 Risk Analysis

Risk analysis comprehends the nature of risk and determines its level. As explained in Section 2, the Confidentiality, Integrity, and Availability triad (CIA) is the core principle of information security.

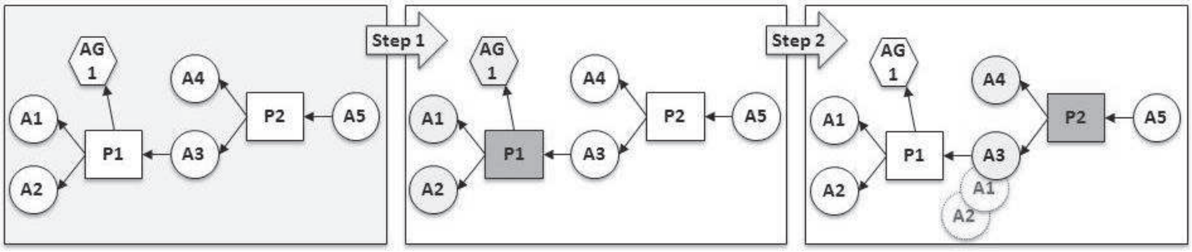


Fig. 9 Rule verification.

Table 2 Security check.

Step	Group 1	Group 2	Group 3	Group 4
1	A1-P1 A2-P1	A1-A2-P1	A1-P1-AG1 A2-P1-AG1	A1-A2-P1-AG1
2	A3-P2 A4-P2 A3(A1)-P2 A3(A2)-P2	A3-A4-P2 A3(A1)-A4-P2 A3(A2)-A4-P2		

In this research, we chose a comprehensive approach to record and confirm data authenticity, recording and providing a complete record of how we handled all of the data, tracking each and every user who accesses them, and logging any unauthorized access attempts.

The provenance of information is used to determine whether information can be trusted to identify a problem that causes invalid output and credits the originators when it is reused. We must provide detailed information that defines the information, its source, and everyone and everything that's happened to it since its creation. In addition to the conventional approach of security requirements that include confidentiality, integrity, and availability, based on the provenance information, we include other requirements to support trust, compliance, and completeness.

We define seven categories of risk based on provenance information by the following questions:

Confidentiality: Are your data always confidential?

Integrity: Can you guarantee their integrity?

Availability: Are your data available when you need them?

Authenticity: Do you know your data's authenticity?

Possession: Do you know who possesses your data at all times?

Use: Are you confident that you can always use them?

Provenance: Can you always assert the provenance of your data?

Based on the above security risk identification, we implemented function *makeScore(co, in, av, au, po, us, pr)*. As shown in Fig. 8, each security rule file includes this function to describe the risk's nature. It is a function of the score generation for each risk category and includes two values: 0 = not applicable, 1 = applicable.

Based on the risk identification explained in Section 4.2.1, for each process that uses one or more information asset, we can identify several security rules. In risk analysis, the makescore function sums the value of each security rule to describe generically the risk's nature.

For example, *makeScore (2,1,2,1,1,3,1)* produces the following results: Confidentiality: 2, Integrity: 1, Availability: 2, Authenticity: 1, Possession: 1, Use: 3, and Provenance: 1. They are shown in radar charts (**Fig. 10**).

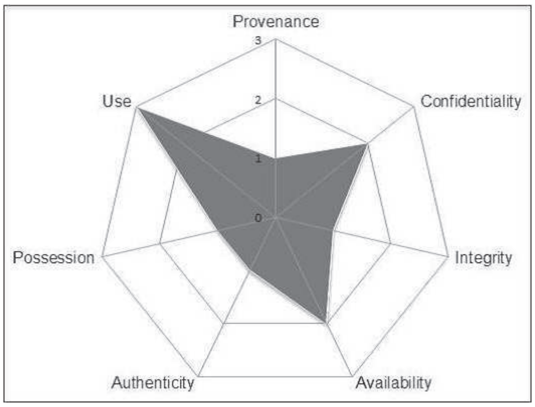


Fig. 10 Risk analysis: radar chart.

As an axis, this chart displays the security risk results for each category. In addition, we can display the overall trends of all risks.

4.2.3 Risk Evaluation

In this section we compare the process of the risk analysis results with the risk criteria to determine whether the risk or its magnitude is acceptable or tolerable. Placing a monetary value on information assets has proven to be very hard for many organizations. Information assets are often not carried on the books as capital investments, so determining their monetary equivalent can be convoluted. Often an information asset's value is found in the process it supports and not in the information itself.

Our evaluation of risk analysis results is based on the possible impact of the risks described in FIPS 199 and the security guide for mapping types of information in the security categories described in a special publication: NIST SP 800-60.

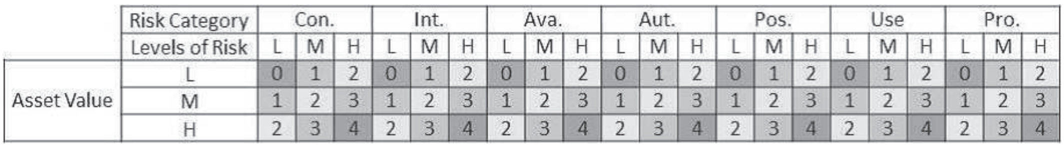
FIPS 199 defines the tree level of the risks for each security category. This means the possible impact of the risk on organizations.

Low: The potential impact is low if the loss of confidentiality, integrity, or availability might have a limited adverse effect on the organization's operations, its assets, or its individuals.

Moderate: The potential impact is moderate if the loss of confidentiality, integrity, or availability might have a seriously adverse effect on the organization's operations, its assets, or its individuals.

High: The potential impact is high if the loss of confidentiality, integrity, or availability might have a catastrophic effect on the organization's operations, its assets, or its individuals.

An asset's value is determined by looking at the potential impact on an organization if the asset's security is compromised.



	Risk Category	Con.			Int.			Ava.			Aut.			Pos.			Use			Pro.		
	Levels of Risk	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
Asset Value	L	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
	M	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
	H	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4

Fig. 11 Risk evaluation.

Table 3 Security rules.

Group	ID	AttrName1	AttrValue1	AttrName2	AttrValue2	Process	AgentAttr	AgentValue
3	9	Language	JA	N/A	N/A	N/A	Language	EN
3	16	AccessRights	OPEN	N/A	N/A	N/A	Class	COM
2	47	License.cc	CC-BY-SA	License.cc	CC-BY-NC-SA	Join	N/A	N/A
4	80	License.cc	CC-BY	License.cc	CC-BY-NC-SA	Join	Class	COM
4	104	License.cc	CC-BY-SA	License.cc	CC-BY-NC-SA	Join	Class	COM

First, we classified the information asset by type (research and development, disaster monitoring, or prediction, for example). Then for each type of information asset, the potential impact is rated on a simple scale of high (H), medium (M), or low (L).

The asset values and the security risk category levels, which are relevant to each type of consequence, are matched in a matrix (**Fig. 11**) to identify each combination of the relevant measures of risk on a scale of 0 to 4. The values are placed in the matrix in a structured manner. The appropriate row is identified by the asset value, and the appropriate column is identified by the possible impact for each risk category. For example, if the asset has a (M) value, the possible impact of the confidentially (Con) is (H), and the measure of the risk is 3.

5. Implementation and Evaluation

Our proposed risk assessment framework works as a web application. **Figure 12** shows the system architecture. This system uses a membership function to provide risk assessment based on the current user. After logging onto the system, users can access information assets. The OPM generator provides a graphic representation of the provenance information (Section 4.1) by access user and accesses the profile information. Then the risk assessment diagnostic implements a risk assessment (Section 4.2) 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, including SP 800-60.

The National Institute of Information and Communication Technology, Japan (NICT) has a large-scale web archive that contains about four million documents. In this research we are working on the discovery of new knowledge by a large variety of data, including such sensing and science data as web archives. Especially in the field of disaster prevention, these data can be used to rapidly analyze disaster information. For each information asset, we include the metadata attributes to enable the provenance representation of information assets (Table 1). **Figure 13** shows the information asset management screen of the Knowledge Language Grid.

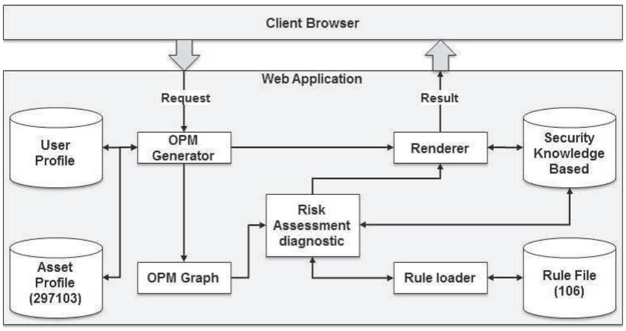
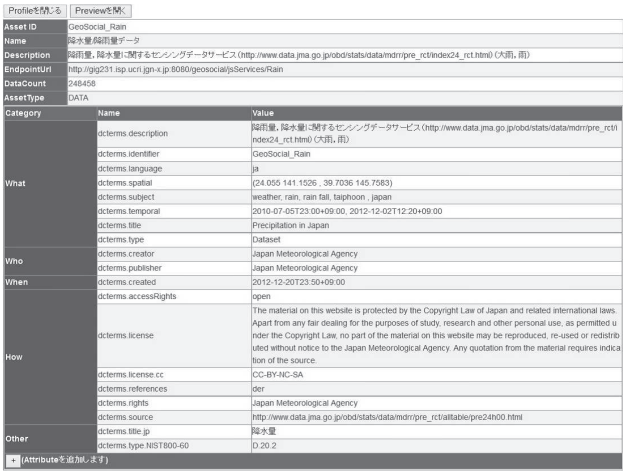


Fig. 12 System architecture.



Asset ID	GeoSocial_Rain
Name	降水量/降雨量データ
Description	降雨量，降水量に関するセンシングデータサービス（http://www.data.jma.go.jp/obd/stats/data/mdrr/pre_rct/index24_rct.html）（大雨，雨）
EndpointUrl	http://gig231.isp.ucri.jgn-x.jp:8080/geosocial/jsServices/Rain
DataCount	248458
AssetType	DATA

Category	Name	Value
What	dcterms.description	降雨量，降水量に関するセンシングデータサービス（http://www.data.jma.go.jp/obd/stats/data/mdrr/pre_rct/index24_rct.html）（大雨，雨）
	dcterms.identifier	GeoSocial_Rain
	dcterms.language	ja
	dcterms.spatial	(24.055 141.1526 , 39.7036 145.7583)
	dcterms.subject	weather, rain, rain fall, taiphoon , japan
	dcterms.temporal	2010-07-05T23:00+09:00, 2012-12-02T12:20+09:00
	dcterms.title	Precipitation in Japan
	dcterms.type	Dataset
Who	dcterms.creator	Japan Meteorological Agency
	dcterms.publisher	Japan Meteorological Agency
When	dcterms.created	2012-12-20T23:50+09:00
How	dcterms.accessRights	open
	dcterms.license	The material on this website is protected by the Copyright Law of Japan and related international laws. Apart from any fair dealing for the purposes of study, research and other personal use, as permitted under the Copyright Law, no part of the material on this website may be reproduced, re-used or redistributed without notice to the Japan Meteorological Agency. Any quotation from the material requires indication of the source.
	dcterms.license.cc	CC-BY-NC-SA
	dcterms.references	der
	dcterms.rights	Japan Meteorological Agency
	dcterms.source	http://www.data.jma.go.jp/obd/stats/data/mdrr/pre_rct/alltable/pre24h00.html
Other	dcterms.title.jp	降水量
	dcterms.type.NIST800-60	D.20.2

Fig. 13 Information asset repository.

To evaluate our proposed system, we verified the possible security risks based on the provenance information of about 297,103 information assets. For example, an international company wants to collect information about hay fever in Japan to produce a brochure for the Japanese market. First it collects pollen, rainfall, and wind speed data and combines this information.

To enable meaningful risk assessment, the generated risk must be identified when two or more difference sources are combined. We developed a graph representation interface based on OPM (Section 4). As shown in the top of **Fig. 14**, the information asset used, the processes performed, the entities that perform these processes, and any new information asset generated are captured and represented based on OPM.

Based on the proposed risk assessment in Section 4, by analyzing the security attributes of each information asset and implementing the risk diagnostic, we can identify security risks (**Table 3**).

For example, the information asset language is Japanese (JA) and the language of the user who wants to access this information is set to English (EN). The user encountered problems un-

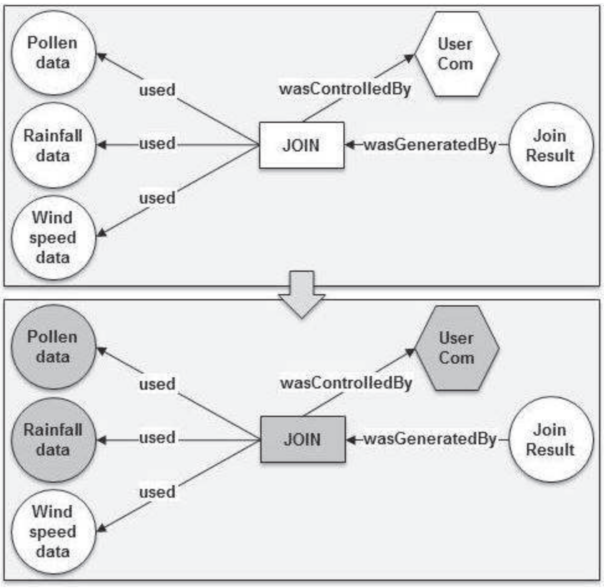


Fig. 14 Provenance-based risk identification.

derstanding the information asset. In addition, information assets have creative commons [27] license as CC-BY-NC-SA (Attribution + Noncommercial + ShareAlike), and the user is a commercial entity. The information asset cannot be used.

The security risks found in the object are represented by gray, as shown at the bottom of Fig. 14. After identifying risks, users can find the necessary security control to counter them by searching on the security control knowledge-base.

6. Conclusion

The information assets living on the Internet often change containers or move through a process that creates new information. An information asset is protected or secured through controls implemented at the asset container level. The degree to which an information asset is protected or secured is based on how well the implemented controls (at the container) align with and consider its security requirements. Any risks to the containers in which information assets live are inherited by the information assets themselves.

Actual risk management approaches do not consider such scenarios. Our proposed system implements a graphic representation of provenance using OPM, and users are allowed to see graphically where and what kinds of data generated security conflicts. In addition, KLG allows users to access security knowledge-bases and provides actual and timely information about security risks and countermeasures.

In this paper, we propose a new approach to implement risk assessment based on provenance information. Even though much further research is needed in this area, the major challenges that clearly capture provenance are in engineering and changing the mindset of data providers about recognizing its importance. Another challenge is mapping information related to a particular hardware to an information asset to provide wide security.

References

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

[2] Groth, P., Gil, Y., Cheney, J. and Miles, S.: Requirements for Provenance on the Web, *International Journal of Digital Curation*, Vol.7, No.1, pp.39–56 (2012).

[3] Bertot, J.C. and Choi, H.: Big data and e-government: Issues, policies, and recommendations, *Proc. 14th Annual International Conference on Digital Government Research*, pp.1–10 (2013).

[4] Nagata, M., Kinpara, I., Takemura, M. and Mineno, H.: The Development of the Scalable Large-scale Safety Information System Using AWS, *2013-CDS-8*, pp.1–8 (2013).

[5] Cloud Security Alliance (CSA) (online), available from ⟨https://cloudsecurityalliance.org/⟩ (accessed 2013-11-27).

[6] Takabi, H., Joshi, J.B.D. and Ahn, G.: Security and Privacy Challenges in Cloud Computing Environments, *IEEE Security and Privacy*, pp.24–31 (Nov. 2010).

[7] Loukides, G. and Gkoulalas-Divanis, A.: Privacy challenges and solutions in the social web it Magazine: Crossroads – The Social Web, Vol.16, No.2, pp.14–18 (Dec. 2009).

[8] The Provenance Working Group of the World Wide Web Consortium (W3C) (online), available from ⟨http://www.w3.org/2011/prov/wiki/Main_Page⟩ (accessed 2013-11-27).

[9] Buneman, P., Chapman, A., Cheney, J. and Vansummeren, S.: A provenance model for manually curated data, *IPAW'06 Proc. 2006 International Conference on Provenance and Annotation of Data*, pp.162–170 (2006).

[10] Microsoft Corp., A Guide to Data Governance for Privacy, Confidentiality, and Compliance, Part 3: Managing Technological Risk (2010).

[11] Cloud Security Alliance (CSA): Security guidance for critical areas of focus in cloud computing V3.0 (online), available from ⟨https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf⟩ (accessed 2013-11-27).

[12] National Institute of Standards and Technology (NIST): Special Publication 800-33, Underlying technical models for information technology security (Dec. 2001).

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

[14] ISO/IEC 27001:2005, Information technology - Security techniques - Information security management systems - Requirements (2005).

[15] ISO/IEC 27002:2005, Information technology - Security techniques - Code of practice for information security management (2005).

[16] National Institute of Standards and Technology (NIST), Special Publication 800-30, Risk management guide for information technology systems (July 2002).

[17] Elahi, G., Yu, E. and Zannone, N.: Security Risk Management by Qualitative Vulnerability Analysis, *Proc. 2011 3rd International Workshop on Security Measurements and Metrics*, pp.1–10 (2011).

[18] National Institute of Standards and Technology: National Vulnerability Database, Common Vulnerability Scoring System Version 2 Calculator (CVSS) (online), available from ⟨http://nvd.nist.gov/cvss.cfm⟩ (accessed 2013-11-27).

[19] Federal Information Processing standards publication: FIPS PUB 199, Standards for Security Categorization of Federal Information and Information System (Feb. 2004).

[20] National Institute of Standards and Technology (NIST): Special Publication 800-60, Guide for mapping types of information and information systems to security categories (Aug. 2008).

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

[22] Institute of Management Accountants: Enterprise Risk Management: Tools and Techniques for Effective Implementation (online), available from ⟨http://poole.ncsu.edu/erm/documents/IMAToolsTechniquesMay07.pdf⟩ (accessed 2013-09-21).

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

[24] CERT, Sofware Engineering Institute: Operational Critical Threat, Asset, and Vulnerability Evaluation (OCTAVE) (online), available from ⟨http://www.cert.org/octve/⟩ (accessed 2013-11-27).

[25] Dublin Core Metadata Initiative (DCMI) (online), available from ⟨http://dublincore.org/⟩ (accessed 2013-11-27).

[26] The Open Provenance Model (online), available from ⟨http://openprovenance.org/⟩ (accessed 2013-09-09).

[27] Creative Commons (online), available from ⟨https://creativecommons.org/⟩ (accessed 2014-03-17).

Guillermo Horacio Ramirez Caceres was born in Corrientes, Argentina. In 1991, he started supporting and creating computer systems for small and middle-sized companies. In 1996, he began teaching programming, mathematics, and logic classes at senior high schools and in 1998 joined Argentina's Ministry of International Trade and Industry. He received a B.E. degree in international trade at Nordeste University in 1999. In 2000, he enrolled at Soka University in Japan and joined its Network Laboratory a year later. He received an M.E. degree in engineering in September 2003 and a Ph.D. in March 2010 from Soka University and is currently a researcher at the National Institute of Information and Communications Technology (NICT) in Japan. His research interests include the fields of Information Network Security, Network Management, and Human Networks. He is a member of IEICE, IPSJ, IEEE, and IIIS.

Koji Zettsu is a Director of Information Services Platform Laboratory at Universal Communication Research Institute of National Institute of Information and Communications Technology (NICT), Japan. He was a visiting associate professor of Kyoto University, Osaka University and Nara Institute of Science and Technology from 2008 to 2012. He was also a visiting researcher of Christian-Albrechts-University Kiel, Germany in 2009. He received his Ph.D. in Informatics from Kyoto University, Japan in 2005. He was in IBM Yamato Software Laboratory from 1992 to 2003. His research interests are information retrieval, databases, data mining, and software engineering. He was a vice chair of Technical Committee on Data Engineering (DE) of IEICE from 2011 to 2012. He was the technical editor of Value-creating Network sub-working group of New Generation Network Forum, Japan from 2009 to 2010. He is a member of IEICE, IPSJ, and ACM.