Data Provenance in the Internet of Things

This work was created in the context of the EU Project COMPOSE (FP7-317862).

Authors: Sabine Bauer
Supervisors: David Parra
           Daniel Schreckling
Submitted: 2013-09-16
Abstract. Data Provenance and the Internet of Things (IoT) are two key subjects which have to be brought together in the near future. Almost every object will be able to communicate with individuals and other objects using the abilities of embedded sensors and actuators. As a result the IoT implicates a huge amount of data whose processing necessitates a certain level of trust which can only be reached with the knowledge of Data Provenance. Security requirements like Integrity and Confidentiality will increase in value. An architecture model to overcome the arising challenges resulting from the implementation of Data Provenance in the IoT has to be acquired.

Keywords: Trust, Data Provenance, Internet of Things (IoT), Security Requirements, Architecture Model

1 Introduction

A wide range of individual objects identifiable via own IP addresses or similar identifiers and controllable over web portals are part of the future everyday life. The mechanisms of the Internet of Things will not even be introduced into the handling of typical smart objects like digital cameras or household appliances, they will also facilitate the handling of complex expiries in science and industry.

Almost every object will be able to deliver information to its users and due to it its importance and functionalities will increase. With the use of embedded sensors and actuators automatically collecting data, human interaction can be reduced. The innovation Google Glass symbolizes this technological progress - data is placed on a prism screen in the field of human vision and can complete what is seen with real-time information from the Web. But not only the IoT will gain in importance.

Fig. 1. The innovation Google Glass showing contextual information of what is seen.
Data Provenance identifying the origin of data and processed operations on it will help to guarantee security requirements like Integrity and Confidentiality of collected data in the IoT. Additionally Data Provenance can support the implementation of security applications and algorithms becoming more and more important when dealing with increasing data volumes. For this purpose Data Provenance mechanisms will be implemented in the Internet of Things and a common architecture model will be developed.

The paper is organized as follows. In Section 2 the concept of Data Provenance, its requirements, its characterizations as well as an architecture model and the data model Provenance Data Model are explained. Section 3 defines the Internet of Things and gives an overview of its infrastructure, range and architecture. Potential difficulties which can arise during the implementation of Data Provenance mechanisms in the IoT are discussed in Section 4. In Section 5 the presented aspects are combined in an architecture model for Data Provenance and the Internet of Things. Section 6 concludes the paper.

2 Data Provenance

Originally provenance - sometimes also called lineage or pedigree - played a vital role in the context of art, archives and archeology because people wanted to know the creator of an object. Recently this term became more important for data in various sectors. The spectrum ranges from finance and commerce to medicine and science as well as legality [9]. The central question has changed. People not only want to know the creator of data, they also care about the manipulators of data. Especially in the sense of security another important question arises additionally: How is the manipulation realized?

Data Provenance provides information about the ownership and both actions and modifications which have been performed on data - it is the history of an item’s ownership [15] and all processes like creation, transformation and copying. In the ages of Social Networks, Cloud Computing and the Internet of Things as well as distributed sometimes untrustworthy storage locations, the owner should have information about the accuracy, origin and timeliness of his data [4]. In the Sarbanes-Oxley Act, for instance, regularities and consequences for signing incorrect corporate financial statements became contractual. Therefore it is important to keep track of data which contributed to financial reports and to authenticate the people who worked on it [9].

2.1 Data Provenance Requirements

To get an accurate level of trust and to guarantee a safe data handling inside a system working with Data Provenance, information which have been collected by using provenance mechanisms have to meet several requirements [9]:

- Completeness: Every single action which has ever been performed is gathered.
– Integrity: Data has not been manipulated or modified by an adversary.
– Availability: The possibility to verify the collected information. In this context, availability is comparable to auditing.
– Confidentiality: The access to the information is only reserved for authorized individuals.
– Efficiency: Provenance mechanisms to collect relevant data should have a reasonable expenditure.

In addition to Hasan’s identified aspects, there are some other requirements which have to be met when creating a system with Data Provenance. The first one is Privacy - the protection of personal data. Unauthorized collection, storage and use of personal data have to be prevented in- and outside of a Data Provenance system.

The requirements Linkability and Unlinkability are placed in opposition. Unlinkability has to be guaranteed outside a Data Provenance system. All collected information, especially personal data, are not allowed to get out of the system and into the wrong hands. Security concerns, for instance arising when user profiles are created are thus able to be prevented.

Linkability, in contrast, is a requirement to be met inside a Data Provenance system and comes together with Transparency. Actions and modifications on data must be traceable. Behind the demand is a possible selection of necessary information when working with a huge amount of collected information. Maybe due to legal requirements, the selection has to be removed one-day and all the information are needed. Then it should be possible to trace the history of the used information and even to do profiling.

2.2 Data Provenance Characterization

To get accurate information which can be selected according to predefined criteria like a specific time period or a group of users, a characterization of provenance data is necessary. Buneman et al. developed an approach for Data Provenance in databases. They differentiate between why and where provenance. Why provenance describes the source of information which influenced the existence of data. Where provenance, in contrast, is the location in the source database of the extracted information. Thus it is possible to identify the source of errors.

For Data Provenance in the context of the Internet of Things the selection criteria have to be extended. In addition to why and where, there arise the questions about the creator and editor of information - who - and the timestamp and time periods of processes on data - when. Accurate mechanisms like the automatic collection of unique paths or timestamps must be implemented in a Data Provenance system.

2.3 Capturing Provenance Information

The current state of practice are humans directly entering information and manually copying data. Provenance responsible people have to decide how and
which information should be collected to get provenance data. A schema to cap-
ture the necessary meta data and a workflow system with all processing steps
have to be specified. Hereinafter a possible architecture model for collecting
provenance information out of office documents is presented. It serves as a basis
for an architecture bringing Data Provenance and the Internet of Things to-
gether following in Section 5.

format with a fine-grained information flow control architecture. Its purpose is
the deployment and enforcement of security policies for office documents. Initially
every component of the office document - paragraphs, shapes, images - is
labeled to get the meta information. Additionally, in a sequence of records the
performed changes on the document as well as author information are stored
in a provenance enabled format. Hereby Confidentiality can be supported. The
sticky provenance thus contains all labels and the document’s change history and
can be cryptographically attached to the document. Unauthorized users without
the cryptographic link are not able to edit or remove any provenance data. The
Provenance Manager, an extension module for the office application, contains
the architectural concept. It unifies the labeling process with the storage process
and supports an user-friendly representation and modification interface.
The description of the provenance information can be implemented by a generic
data model. W3C proposed the Provenance Data Model, a further development
of the Open Provenance Model OPM (openprovenance.org).

2.4 The Provenance Data Model

To provide insight into a description mechanism the Provenance Data Model
(PROV-DM) is considered in more detail. The basic model consists of three
classes: the entity, the activity and the agent. In the context of the data model
“provenance describes the use and production of entities by activities, which
may be influenced in various ways by agents” [14].

- Entities: Provenance is described of entities. They can be physical, digital,
conceptual or other type of things.
- Activities: Activities are described by duration and their correlation with
entities. Activities produce as well as utilize entities. The specified activities
are consuming, processing, transforming, modifying, relocating, using and
generating entities.
- Agents: Agents are specified by their responsibility. They can be responsible
for activities taken place, for the existence of entities and for the activities
of other agents.
- Extended structures: In addition to the three basic classes there are some
extended structures like subtypes, identifications and expanded relations.

The following relations describe the core model of the Provenance Data Model
[10][14], shown in Figure 2.
Fig. 2. The core model of the Provenance Data Model [10]

- *wasDerivedFrom*: Entities can be derivations of other entities.
- *wasAttributedTo*: Entities and agents are correlated.
- *actedOnBehalfOf*: Agents can act on behalf of themselves and other agents.
- *wasGeneratedBy*: Entities can generate activities.
- *used*: Activities in turn use entities.
- *wasInformedBy*: Activities can inform themselves and other activities, they are limited in time by fixed start and end dates.
- *wasAssociatedWith*: Activities are connected with agents.

There is no single correct provenance specification. “Multiple provenance descriptions about the same entity can co-exist” [14]. This phenomena is called provenance of provenance.

Fig. 3. A description example of PROV-DM
3 The Internet of Things

Already in the early 1990s Mark Weiser first hit upon the idea of creating an Internet of Things realized by the interconnection of physical items and the virtual world. Users can remotely control these physical items which take the role of access points to Internet services [12]. The items become smart and get communication and network interfaces to be accessible everywhere and everytime. They can be identified, located and addressed by unique characteristics like IP addresses. Embedded sensors and actuators take over the collection of information and the interaction with the environment. Humans do not have to be on-site to collect necessary information and to control and manage processes - remote control simplifies all operations. Smart objects additionally develop self-* capabilities which reduce human intervention and enable an independent configuration and maintenance.

3.1 The Range of the IoT

The Internet of Things can be part of every situation in everyday life. Hereinafter only a few applications will be displayed. Web-enabled and with cameras equipped mobile phones, the flagship of smart objects, are used by nearly every person. On the one hand they have an enormous entertaining factor for example when you want to share your current position with your friends. On the other hand they can improve everyday actions like shopping. A mobile phone can display additional product information such as calorie values or allergen information. Even the energy management at home during office times becomes possible. Energy providers sell smart metering systems providing consumption information and the individual management like the regulation of warm water [6]. Embedded sensors and actuators allow a rapid reaction to unpredicted situations guided by automated systems like automatic braking systems in cars and trucks - without any human intervention [5].

Figure 4 provides insight into a future interpretation of the Internet of Things. The user no longer depends on his physical location. Smart objects like mobile phones enable a remote control and access to items like the electric car or the water heater at home which are equipped with their own IP addresses. Out of this changed situation results a new meaning of the term Social Network where not only users are linked together but also objects among themselves and with
their users. Now objects have the ability to deliver information and thus they supply their owners with extended functionality and usability \[19\].

3.2 Resulting Challenges and the Connection to Data Provenance

In the Internet of Things sensors and actuators can take over the complete interaction with the environment and collect data automatically. But they not only collect information about the performed actions, the nodes also capture semantic information like the date and the author of an action taken with a specified object. For instance a sensor in a car can store the name and date of birth of his driver as well as his behavior in road traffic. By now, sensors are able to support a slot for exchangeable memory cards to bridge for longer periods without connection to the network.

Some challenges arise with the distribution of the IoT. It involves a huge amount of data which has to be stored and archived somewhere. Often the complete information are not necessary and a selection is possible. Here Data Provenance comes into play: the acquisition and the storage of information can be transferred into predefined data models enabling the reconstruction of actions. Additionally, a filtering of the collected information is possible and unusable contents can be deleted. Challenges like the treatment of personal data and privacy as well as the introduction of standards can be simplified. To bring IoT and Data Provenance together connection points between the two architecture models have to be identified.

Fig. 4. A future interpretation of the Internet of Things \[19\]
3.3 IoT Infrastructure Model

To identify connection points to Data Provenance a possible infrastructure model of the Internet of Things is considered [17].

The foundation of this conceptual reference model is the Domain Model. It defines the main IoT concepts and the relations between these concepts. To generate a common understanding, objects and attributes are specified. The Domain Model supplies the Information Model defining the structure like relations and attributes of all the information collected in the IoT. Its main parts are the modeling of the information flow and the storage. All the concepts of the Domain Model which are important for the representation and manipulation of data as well as their relations are specified. The Functional Model which is based on the Domain Model and the Information Model clarifies common semantics. For this purpose, it considers the main concepts of the Domain Model and builds groups of functionalities on it. Via Functional Groups the Functional Model can interact with the other models and their concepts. The Information Model is thereby used to structure the implemented information. The Communication Model manages the complexity of communication in the heterogeneous IoT environment. For this purpose, the Domain Model structures the connection of the communicating objects. The Security Model contains relevant security functionalities, their interdependencies and interactions. Deployment security, communication security and service security can be distinguished. The two finally presented models are in interaction with the Functional Model by their Functional Groups - instances of the respective model. Next to these two Functional Groups, there also exist seven other Functional Groups like IoT Service, IoT Business Process Management and Device.

![Diagram of IoT Infrastructure Model]

Fig. 5. An infrastructure model of the IoT

As possible connection point to a Data Provenance system the Information Model can be identified. The collected provenance information could be integrated in this model as data model of entities, agents and activities, comparable to PROV-DM. According to the provenance application there can also be interactions with the Communication and the Security Model.
3.4 IoT Architecture Model

The IoT can not only be considered as infrastructure concept. It can also be handled as architecture model. The base of the IoT architecture are smart devices like smart phones, cars and household appliances equipped with sensors and actuators. They are able to collect information about their environment and to interact with their environment. They form the Device Layer (green) on which the Sensing Layer (blue) is placed. The captured information can be retrieved with a web server which is the gateway to the users. The web server communicates with a database serving as storage platform for all the information. Their interaction is the basis of the Gateway Layer (yellow). On top of the layer stack is the User Interface Layer (red). It contains the web interface which can be represented by web browsers and offers the access for every user. The web interface can be equipped with an access control concept to authenticate the users. Figure 6 represents the stack of the four interacting layers. The colors in the brackets show the single layers.

![Fig. 6. An architecture model of the IoT - a stack of four interacting layers](image)

4 Joining Data Provenance and the IoT

Data Provenance and the Internet of Things are two terms which bring a lot of benefits and difficulties together. Subsequently some problems arising when linking these two concepts are shown. Hereinafter the Data Provenance requirements introduced in section 2.1 - Completeness, Integrity, Availability, Confidentiality, Efficiency, Privacy, Linkability and Unlinkability - will be examined in this context.
4.1 Amount of Data

When every device becomes a smart object which can communicate with other objects, Internet services and humans, a huge amount of data accumulates and overheads must be prevented. To meet the requirements Completeness and Efficiency, this huge amount of data has to be managed. Usually there are standalone database systems to store all collected provenance information together with referred data. Due to the consideration of a separation of data and related provenance - a kind of meta-data - an efficient architecture has to be implemented to avoid inconsistency. Especially actions like backup, restoration and copies have to be considered [15]. The enormous amount of data coming up with the IoT must result in considerations if this kind of storage management is still possible and with which changes it can be realized. Furthermore it should be clarified if the available storage capabilities suffice and which storage locations exist.

4.2 Interconnectedness

Expansive networks with an incalculable number of available devices like the IoT result in an enormous interconnectedness complicating traceability. Mechanisms to meet the requirements Completeness, Efficiency and Linkability have to be developed. For provenance in the context of Social Networks like Facebook or Twitter, Jennifer Goldbeck implemented an approach to categorize different levels of trust dependent on the relationship of two users [7]. Consequently personalized recommendations with trust values can be computed and be used for future analysis and tracing of user statements. The application of this attempt with regard to the categorization of smart objects in the IoT and the assignment of certain trust levels can give rise to some difficulties. There are no hard and fast rules about how to categorize because the abundance of available and communicating devices is incalculable. Rules for the assignment of trust levels have to be determined. Furthermore a possible misuse of traceability has to be considered.

4.3 Faulty Data Propagation

The IoT and its usage of a lot of different interconnected smart objects provides attackers with a target. It becomes more difficult to authenticate the real data origin and to preempt the modification and manipulation of information. The requirements Integrity and Availability nevertheless have to be met in a combined system. Mechanisms to check the correctness of data are needed. But false and modified data exchange can also result from unwanted false entries. Therefore algorithms helping to extract the necessary information of provenance data to identify and verify information’s history have to be implemented. Due to Linkability it is also possible to identify the propagation routes of faulty data throughout the network [16].
4.4 Profiling

The already mentioned difficulties lead to the main problem of profiling users with the collection of the available information over the expansive IoT network. The correct interception and combination of data enables for instance the construction of accurate motion and habit profiles. The basis for this is provided by the identification of user specific information possibly realized by the manipulation of selected sensors and actuators. These aspects complicate the fulfillment of the requirements Confidentiality, Privacy and Unlinkability. Access control models enabling the selective determination of pre-defined access rights produce relief. Different user groups with an explicit assignment of rights can be defined. For instance, the $UCON_{ABC}$ Usage Control Model \[18\] combines access control, trust management and Digital Rights Management. Sensitive information like provenance data can be protected with diverse usage decisions based on authorizations, obligations and conditions.

5 Common Architecture of Data Provenance and the Internet of Things

After the introduction of the two topics Data Provenance and the Internet of Things in combination with their challenges, a conceptual model for a common architecture will be specified. Therefore the architecture model of the IoT serves as basis. Its components Users, Browser, Access Control, Web Server, Database, Sensors and Devices remain and are extended by provenance specific components like the Provenance Event Handling.

Smart devices which are forming the foundation of the architecture are equipped with sensors and actuators for the communication and interaction with the environment. The actions of the sensors and actuators are controlled by the component Provenance Event Handling. It resorts to underlying algorithms managing the handling of the resulting provenance information. The Collection Algorithm specifies the collection of the necessary provenance information and determines the data format. The Verification Algorithm ensures the collection of the right information and recognizes manipulated data. With the help of the Categorization Algorithm the collected information can be classified. Different trust levels for the smart objects, identification characteristics as well as categorization rules have to be pre-defined. By the application of the Selection Algorithm provenance information can be limited, i.e. a special time period can be selected.

The captured information can be retrieved with a Web Server. It is connected with a Database, the storage system. The Database can be divided into two interconnected parts: the Device Information and the associated Provenance Information. Both parts must have unique identifiers to perform queries and references to identify their relations. The data stored in the Database is additionally saved at a backup storage system with a Management Interface. Furthermore a Cloud Backup can be part of the architecture model. The Web Server can be accessed via a Browser Interface - the interaction component. On the one hand the Browser Interface serves as communication platform
between the users and the devices. On the other hand, this component provides the visualization of the provenance information. At this point provenance profiles can be generated, the individual setting can be adapted and the manipulation of the provenance information is possible. To avoid unauthorized access, the Browser Interface is linked with an Access Control which is also coupled with the Database and the Management Interface of the backup component.

![Conceptual model for a common architecture of Data Provenance and the Internet of Things.](image)

**Fig. 7.** Conceptual model for a common architecture of Data Provenance and the Internet of Things. (A bigger representation of this model is part of the Appendix.)

The identified Data Provenance requirements are integrated in the presented conceptual model. With the Provenance Event Handling the requirements Linkability and Unlinkability can be controlled. The component determines the fundamental algorithms for the collection of the information and guarantees a secure handling. The Verification Algorithm proves possible manipulations and modifications and ensures Integrity. In addition, the Selection Algorithm makes it possible to verify the collected information and to select the required information. The requirements Efficiency and Completeness are implemented at the Backup. A storage component with a high data throughput and a sufficient data volume must be selected. The Database guarantees Availability of all the information by being accessible 24/7. Privacy and Confidentiality requirements are handled with the Access Control component. It can be supported with an usage control model like $UCON_{ABC}$. 
6 Conclusion

The Internet of Things is the new generation of data exchange supported by a huge amount of smart objects. The goal is the connection of the resulting volume of data with provenance data without producing a lot of overhead. Up to now Data Provenance is not provided in the existing Internet model [10].

To overcome this state, connection points between the Internet of Things and Data Provenance have been identified. A conceptual architecture model connecting these technologies has been developed. During the implementation some challenges arose. Provenance requirements imply the verification of collected information and call for an access rights management. Additionally, a characterization of provenance information is necessary to be able to select the desired information. The IoT provides different connection points to Data Provenance. The Information Model is the infrastructural interface enabling the integration of a provenance data model. Considering the IoT architecture model, the Gateway Layer as well as the User Interface Layer support the connection. A common architecture calls for provenance specific components. For this purpose, the Provenance Event Handling component has been implemented. Some challenges remain: The growing Internet of Things is in need of enough storage capabilities. Furthermore, secure and trustworthy authentication mechanisms for collected information and accessing users have to be integrated into a common architecture.
Fig. 8. Conceptual model for a common architecture of Data Provenance and the Internet of Things
References