ETSI GR CIM 007 V1.1.1 (2022-03)

Context Information Management (CIM); Security and Privacy

Disclaimer

The present document has been produced and approved by the cross-cutting Context Information Management (CIM) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/CIM-007-SEC

Keywords
API, architecture, GAP, information model, privacy, security, smart city

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) cross-cutting Context Information Management (CIM).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

The present document identifies the security and privacy aspects that are relevant when building systems based on the NGSI-LD API and the NGSI-LD information model. It defines high-level objectives that have to be taken into account when specifying the mechanisms that enable addressing the security and privacy aspects.

Contributions to the present document have been supported by the following European Union Horizon 2020 research projects: Fed4IoT (Grant number 814918) and IoTCrawler (Grant number 779852).

1 Scope

The present document provides a security and privacy review of the ISG CIM specifications, in particular the NGSI-LD API [i.1] and the Data Model [i.2]. The review identifies the risks from attack and means to mitigate the risk in the form of core security objectives and privacy protection objectives to be met by NGSI-LD Systems.

NOTE: The scope of the security and privacy protection objectives include those related to data provenance, and the role of data aggregation as impacting the attack surface of NGSI-LD System deployments.

2 References

2.1 Normative references

Not applicable to the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long-term validity.

The following referenced documents are not necessary for the application of the present document, but they assist the user with regard to a particular subject area.

[i.1] ETSI GS CIM 009: "Context Information Management (CIM); NGSI-LD API".

[i.2] ETSI GS CIM 006: "Context Information Management (CIM); Information Model (MOD0)".

NOTE: Available at https://www.etsi.org/deliver/etsi_gs/CIM/001_099/006/01.01.01_60/gs_cim006v010101p.pdf.

[i.3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (GDPR).

[i.4] Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union (NIS Directive).

[i.5] Directive 2014/53/EU of the European Parliament and of the Council of 16 April 2014 on the harmonisation of the laws of the Member States relating to the making available on the market of radio equipment and repealing Directive 1999/5/EC (Radio Equipment Directive (RED)).

[i.6] European Treaty Series No. 185: "Convention on Cybercrime".

NOTE: Available at https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=185.

[i.7] ETSI EN 303 645: "CYBER; Cyber Security for Consumer Internet of Things: Baseline Requirements".

NOTE: Available at https://www.etsi.org/deliver/etsi_en/303600_303699/303645/02.01.01_60/en_303645v020101p.pdf.

[i.8] ETSI TS 102 165-1: "CYBER; Methods and protocols; Part 1: Method and pro forma for Threat, Vulnerability, Risk Analysis (TVRA)".

NOTE: Available at https://www.etsi.org/deliver/etsi_ts/102100_102199/10216501/05.02.03_60/ts_10216501v050203p.pdf.

[i.9] ETSI TS 102 165-2: "Telecommunications and Internet converged Services and Protocols for Advanced Networking (TISPAN); Methods and protocols; Part 2: Protocol Framework Definition; Security Counter Measures".

NOTE: Available at https://www.etsi.org/deliver/etsi_ts/102100_102199/10216502/04.02.01_60/ts_10216502v040201p.pdf.

[i.10] IETF RFC 3986: "Uniform Resource Identifier (URI): Generic Syntax".

NOTE: Available at https://tools.ietf.org/html/rfc3986.

[i.11] Jean Louis Raisaro, Juan Ramon Troncoso-Pastoriza, Mickael Misbach, Joao Sa Sousa, Sylvain Pradervand, Edoardo Missiaglia, Olivier Michielin, Bryan Ford, and Jean-Pierre Hubaux. 2019. MedCo: Enabling Secure and Privacy-Preserving Exploration of Distributed Clinical and Genomic Data. IEEE/ACM Trans. Comput. Biol. Bioinformatics 16, 4 (July 2019), 1328-1341.

NOTE: DOI: https://doi.org/10.1109/TCBB.2018.2854776.

[i.12] i2b2, Informatics for Integrating Biology & the Bedside, National Center for Biomedical Computing, USA.

NOTE: Available: https://www.i2b2.org/index.html.

[i.13] L. Cinquini et al., "The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data," 2012 IEEE 8th International Conference on E-Science, 2012, pp. 1-10.

NOTE: DOI: https://doi.org/10.1016/j.future.2013.07.002.

[i.14] ETSI TS 103 485: "CYBER; Mechanisms for privacy assurance and verification".

[i.15] ETSI TS 103 486: "CYBER; Identity Management and Discovery for IoT".

[i.16] ETSI TS 187 020: "Radio Frequency Identification (RFID); Coordinated ESO response to Phase 1 of EU Mandate M436".

[i.17] ETSI TS 102 894-2: "Intelligent Transport Systems (ITS); Users and applications requirements; Part 2: Applications and facilities layer common data dictionary".

[i.18] OASIS: "eXtensible Access Control Markup Language (XACML) Version 3.0".

[i.19] ETSI TR 103 719: "CYBER; Guide to Identity Based Cryptography".

[i.20] ETSI TS 103 352: "CYBER; Attribute Based Encryption for Attribute Based Access Control".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

Access Control: means of prevention of unauthorized use of an Object or Service by a User or the public

Action: operation involving modifying or reading an Object or its Attribute. e.g. create, read, update, delete

Actor: individual person, group of persons, organization, or company

Administrator Policy: Policy defined by an administrator and applicable to any of the current data and services in the NGSI-LD System

Agent: software program that represents Actors to produce, consume or manipulate data

Attribute: characteristic of an Object or User

NOTE: An Attribute is the minimal piece of data that the system grants access to or bases access control decisions upon.

Consumer: User consuming data

Context: measured and inferred knowledge that describes the environment of an Entity

Context-based Access Control: Access Control decision process based on Context

Contract: formal agreement governing part of the collective behaviour of the involved Actors

Credentials: data that is transferred to establish or confirm the claimed Identity of a User

Data Controller: natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of Personal Data

Data Processor: natural or legal person, public authority, agency or other body which processes Personal Data on behalf of the Data Controller

Data Provenance: metadata that is associated with data that details the origin, changes to, and details supporting the confidence or validity of data

Data Subject: identified or identifiable natural person, who can be identified directly or indirectly in particular by reference to an identification number or to one or a combination of factors specific to physical, physiological, mental, economic, cultural or social identity

Entity: Object that is an informational representation of something that is considered to exist in the real world, physically or conceptually

Group: named set of Users, Objects or Attributes

Identity: set of Attributes by which an Entity or User is uniquely described, recognized or known

Integrity: surety that the data or service has not been altered or destroyed in an unauthorized manner

NGSI-LD Actor: human or legal entity, operating the NGSI-LD Agents and legally responsible for their actions.

NGSI-LD Agent: software components that interact with each other using NGSI-LD

NGSI-LD Attribute: NGSI-LD Property or an NGSI-LD Relationship

NGSI-LD Context Information: measured and inferred knowledge that describe the environment represented by means of NGSI-LD Entities and/or NGSI-LD Attributes

NGSI-LD Consuming Actor: NGSI-LD Actor consuming data from an NGSI-LD Providing Actor including the NGSI-LD Broker

NGSI-LD Entity: Entity in an NGSI-LD System

NGSI-LD Property: description which associates a main characteristic, i.e. an NGSI-LD Value, to either an NGSI-LD Entity, an NGSI-LD Relationship or another NGSI-LD Property

NGSI-LD Providing Actor: NGSI-LD Actor providing data to an NGSI-LD System

NGSI-LD Relationship: description of a directed link between a subject which is either an NGSI-LD Entity, an NGSI-LD Property or another NGSI-LD Relationship on one hand, and an object, which is an NGSI-LD Entity, on the other hand

NGSI-LD System: set of all interconnected software components that use the NGSI-LD API for communicating among each other

NGSI-LD User: User that is registered in an NGSI-LD System

NGSI-LD Value: JSON value (i.e. a string, a number, true or false, an object, an array), or a JSON-LD typed value (i.e. a string as the lexical form of the value together with a type, defined by an XSD base type or more generally an IRI), or a JSON-LD structured value (i.e. a set, a list, a language-tagged string)

Object: data unit created or requested by an application

Personal Data: any information relating to an identified or identifiable natural person (Data Subject)

Policy: set of Access Control rules defining allowed Users for certain operations within specified contexts that each User has to comply with to be granted access to an Object

Provider: User providing data

Service: software functionality that different Users can reuse, together with the Policies that should control its usage

Subject: User acting as Consumer or Provider of a Service

Trust: level of confidence in the capabilities and Integrity of a User or Service

User: Virtual representation of an Actor or Agent

NOTE: If users have to be registered in a registration process then this requires issuing Credentials

User Policy: Policy defined by a User and restricted to the set of data inserted by that User

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

API	Application Programming Interface
CIA	Confidentiality, Integrity and Availability
CoE	Council of Europe
DTE	Deterministic Encryption
EAV	Entity Attribute Value
EN	European Norm
ESGF	Earth System Grid Federation
EU	European Union
GDPR	General Data Protection Regulation
HIV	Human Immunodeficiency Virus
HTTP	Hypertext Transfer Protocol
HVAC	Heating Ventilation and Air Conditioning
IBC	Identity Based Cryptography
ICT	Information and Communication Technology
IdM	Identity Management
IP	Internet Protocol
IRI	Internationalized Resource Identifier
ISG	Industry Specification Group
JSON	JavaScript Object Notation
JSON-LD	JSON Linked Data
LD	Linked Data
MQTT	Message Queuing Telemetry Transport
NGSI	Next Generation Service Interfaces
PDP	Policy Decision Point
PEP	Policy Enforcement Point
SA	Security Association
SPU	Storage and Processing Unit

SSL	Secure Sockets Layer
TCP	Transport Control Protocol
UDP	User Datagram Protocol
UML	Unified Modelling Language
URI	Uniform Resource Identifier
VPN	Virtual Private Network
XACML	eXtensible Access Control Markup Language
XML	eXtensible Markup Language
XSD	XML Schema Definition

4 Security and Privacy in the context of NGSI-LD Systems

ETSI ISG Context Information Management (CIM) has defined NGSI-LD as a means of managing and exchanging context information (in a broad sense) between a variety of systems. The present document addresses security and privacy with respect to NGSI-LD and the NGSI-LD API.

In examining the role to be played by Security and Privacy provisions for NGSI-LD, several characteristics regarding the NGSI-LD architecture are considered as below:

- NGSI-LD systems use the NGSI-LD API through which they can query other systems, provide notifications, and receive responses to queries.

EXAMPLE: A Parking Management System could query a Traffic Management System about road occupancy at a particular egress gate.

- Means are available for NGSI-LD systems to discover, register, and report existence of entities and relationships within and across several instances of platforms.
- Distributed NGSI-LD system instances need to be able to reconcile the identity of entities referenced in different systems.

5 System Architecture

5.1 The security model

In most ICT security systems, the provisions for security are made with respect to Confidentiality, Integrity and Availability (CIA) characteristics of data, processes, protocols or systems. In addition, many security analysis approaches consider Security Associations (SA) between peers, conventionally referred to as Alice and Bob, with Eve playing the role of universal adversary. Thus, the CIA model is presented as follows:

- Confidentiality, wherein data shared by Alice with Bob cannot be accessed by Eve.
- Integrity, offering assurance that data shared by Alice with Bob cannot have been manipulated by Eve.
- Availability, offers assurance that data is available to Alice when it is required and not available to any Eve masquerading as Alice.

NOTE 1: The CIA triad, or CIA paradigm, has no straightforward defining reference as its source but, rather, has evolved in a number of environments over time, with references found in increasing numbers from the early 1970s and leading to almost universal acceptance over the intervening period. Offering a single authoritative source as a reference to the term "CIA" is likely to be misleading, or to be contested.

NOTE 2: The convention of naming actors Alice, Bob and Eve has been widely adopted in the security domain as a more reader friendly approach than referring to Peer-A, Peer-B, Peer-C and so on. A Wikipedia article (https://en.wikipedia.org/wiki/Alice_and_Bob) offers some additional insight although there is some dispute regarding the origin of the Alice and Bob names and the origins may be related to any story like adaption used to describe a complex process.

In other words, the wider impact of applying the CIA paradigm is that data originating from Alice is assured to have come from Alice (Availability), that is has not been modified since Alice released the data for transmission (Integrity), and that no unauthorised party has been able to access the content of the data (Confidentiality). A number of consequences follow from this including the need to be able to reliably identify Alice and Bob, to have assurance that Eve cannot masquerade as Alice, thus promoting the requirement for identity management, and authentication. If the CIA paradigm is underpinned by cryptographic mechanisms the nature of the key management has an impact on the core architecture of the system as well as on the problem of key binding to any established SA.

5.2 CIA and Trust frameworks

In consideration of NGSI-LD API trust is established between the calling entity and the called entity such that end-to-end trust is established. The simplified architecture shown in figure 5.2-1 places a broker entity between the context consumer and each of the context source and context producer although the description in ETSI GS CIM 009 [i.1] allows for a direct relationship between the consumer and source/producer (thus bypassing the role of the broker as trusted intermediary).

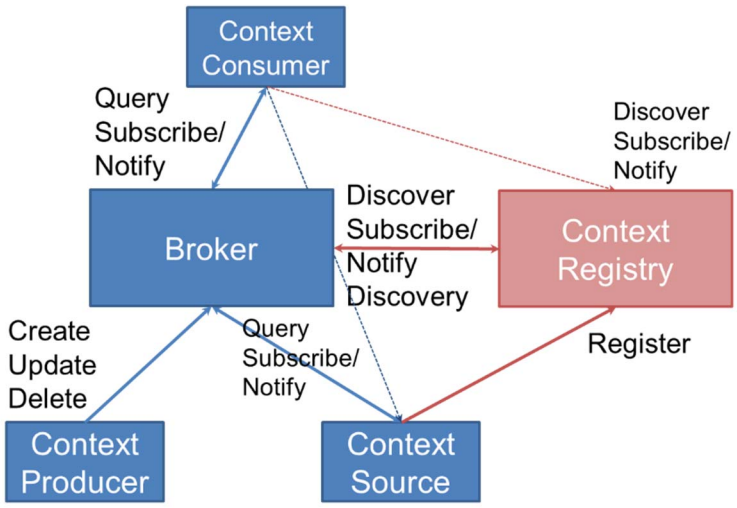

Figure 5.2-1 NGSI-LD architectural roles

In the context of figure 5.2-1 the Context Producer and Context Source have to be trusted by the broker but have no direct trust relationship to the Context Consumer. The NGSI-LD API specification does allow for direct connection between the Context Consumer and each of the Context Producer and Context Source. If the same data is retrieved with both the broker involved, and not involved, in the transaction the trust calculation may be different for each path.

5.3 Security and privacy constraints for NGSI-LD Systems

It is important to recognize that the constraints listed below apply to the user or provider of data, such as used in an NGSI-LD system. The consequences from a technology perspective are that provisions to allow an implementation to comply to regulation (e.g. data protection) are expected to be made available to users and providers. One purpose of the present document is therefore to assist in the identification of those technical provisions. Of itself a secure variant of NGSI-LD would be insufficient to confer regulatory compliance to mechanisms such as GDPR but may, when deployed, give greater likelihood of an operator being able to claim compliance.

It is also noted that regulation applies to legal entities and not to the technical entities (i.e. if a regulatory breach is discovered it is the legal entity that is held liable and not the technology).

There are a number of constraints placed on the use of data, including those implied by a number of regulatory frameworks, that are likely to apply to any deployment of NGSI-LD into network based systems and this includes the following (this list is indicative and no claim is made for its completeness):

- General Data Protection Regulation (GDPR) defined in [i.3] and equivalent regulations in non-EU markets.
- Network Information Systems directive (NIS) defined in [i.4] and equivalent regulations in non-EU markets.

NOTE: There is, at the time of writing, a development to update and strengthen the NIS Directive [i.14] in order to further improve the resilience and incident response capacities of public and private entities, competent authorities and the EU as a whole in both the field of cybersecurity and critical infrastructure protection.

- The Radio Equipment Directive (RED) defined in [i.5] and equivalent regulations in non-EU markets where radio equipment is used.
- Right to repair legislation may apply to ensure that when equipment is repaired and maintained independently of the original manufacturer and supply chain that data in the equipment maintains protection (this may add new entities into the trust model for NGSI-LD).
- Regional and national regulation concerning the safety of equipment and any consequences relating to data safety.
- Regional and national regulation concerning the disposal of equipment at end of life (see also GDPR) wherein data has to be disposed of.

In addition, in many markets there is a broad requirement to enable lawful access to data and content of networks and specific obligations fall onto operators to ensure that their networks and services are appropriately enabled.

EXAMPLE: The European Treaty 185, "Convention on Cybercrime" [i.6] applies for members of Council of Europe and places obligations on CoE members that are in turn placed on data and service providers to ensure reasonable access to data and other digital domain services to prevent crime conducted in the digital domain.

Where NGSI-LD is implemented in devices the security considerations given in ETSI EN 303 645 [i.7] should be taken into consideration.

In summary NGSI-LD deployments are impacted, but NGSI-LD itself is not requested to ensure and enforce compliance, but just to support it.

5.4 Contextualising Security in the scope of an NGSI-LD system

NGSI-LD operates within a wider protocol stack, e.g. in order to allow connectivity of consumer to broker to source NGSI-LD is bound to protocols such as HTTP and MQTT (see ETSI GS CIM 009 [i.1], clauses 6 and 7). Each of HTTP and MQTT have their own security properties. In addition HTTP/MQTT are bound to lower layer connectivity protocols including TCP and UDP over IP, which again have their own security properties. There is a contribution to system security and to system trust, of the entire protocol stack across the networks that connect consumer to broker to source, that should be taken into account in the overall risk management of systems that deploy NGSI-LD.

A conceptual overview of an NGSI-LD system is shown in figure 5.4-1, illustrating the different sources of information that are managed, as well as the security and privacy functionality, which protects it.

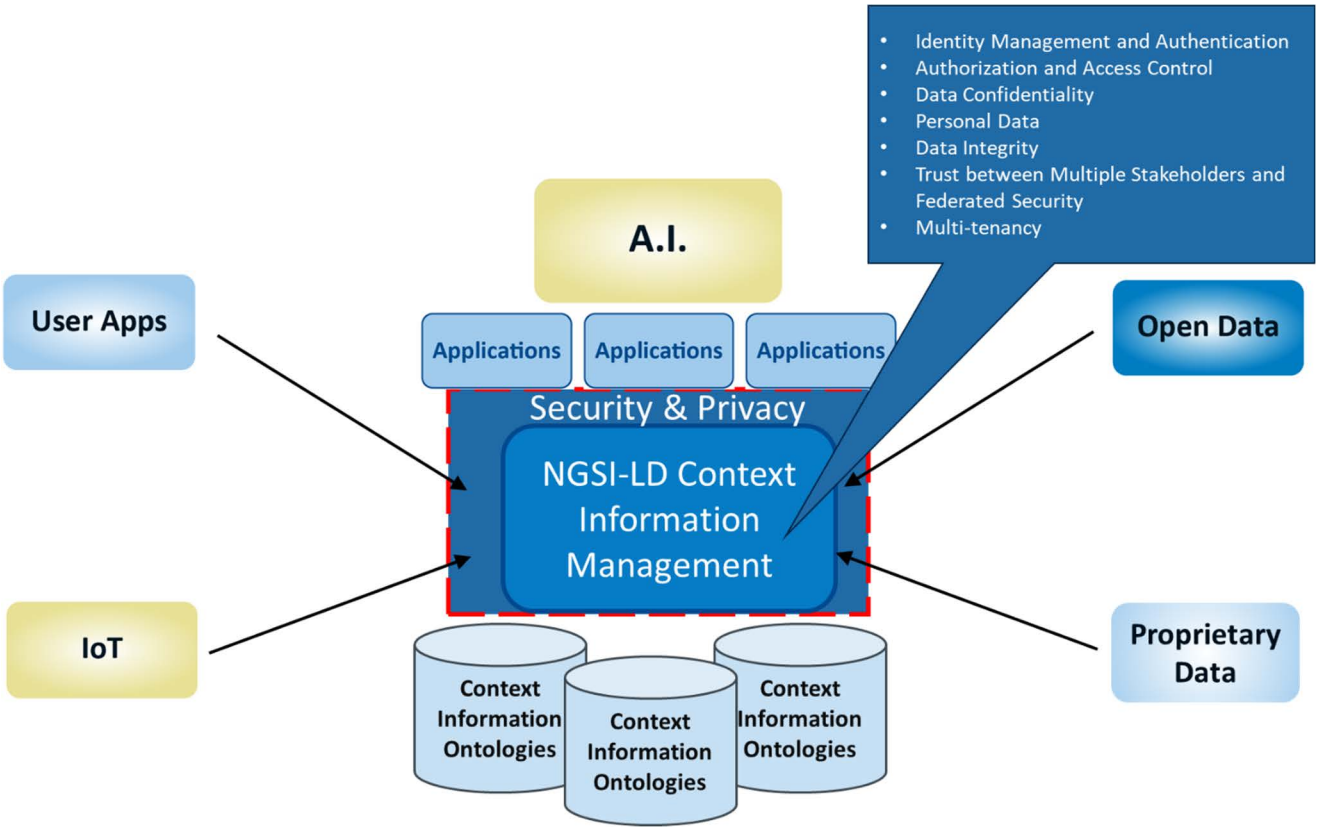

Figure 5.4-1: Conceptual view of an NGSI-LD System where Security & Privacy functionality is represented

In the following, security and privacy features, listed below, are investigated with respect to objectives for NGSI-LD Systems:

- Identity Management and Authentication
- Authorization and Access Control
- Data Confidentiality
- Personal Data
- Data Integrity
- Trust between Multiple Stakeholders and Federated Security
- Multi-tenancy

5.5 Possible system configurations

The NGSI-LD API [i.1] builds on the NGSI-LD Information Model [i.2]. It does not prescribe all possible architectural configurations that can be built on top of it, but instead introduces architectural roles and three prototypical architectures. The NGSI-LD API is designed in such a way that these prototypical architectures can be supported efficiently, but additional architectures can be envisioned as well.

For the purposes of the present document, the default design decision for a deployment of NGSI-LD is that all the components are fully decentralised and may also be federated. From a security analysis perspective, designing security measures for the most complex of scenarios is appropriate, even if deployment decisions are later made for simpler approaches. The underlying protocol stacks that support connectivity of NGSI-LD, in particular HTTP and MQTT, assume a network interconnection, and the identity structure of NGSI-LD using URIs for identification makes this analysis decision appropriate (i.e. global access and naming is assumed).

Figure 5.5-1 shows the NGSI-LD architectural roles as they are defined in the NGSI-LD specification [i.1]. Context Consumers request context information using the NGSI-LD API, typically from a Broker, but possibly also directly from a Context Source. For requesting information, they can use synchronous queries, getting the currently available information in response, or they can use the asynchronous subscription / notification interaction pattern for being notified about any future changes fitting the subscription.

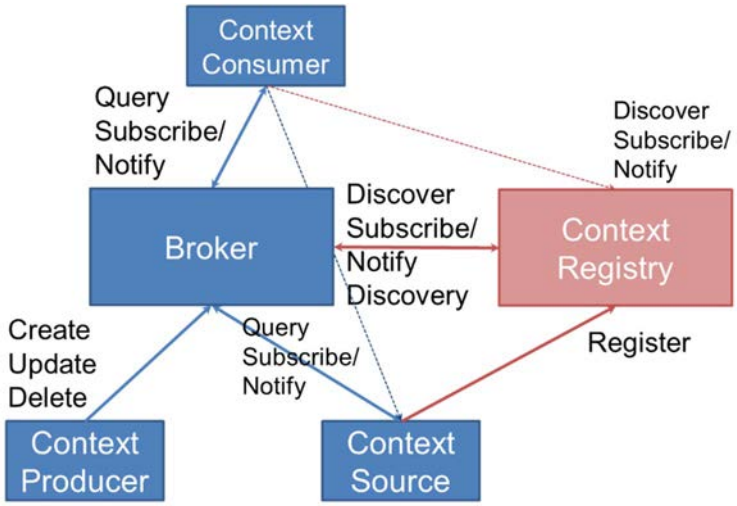

Figure 5.5-1 NGSI-LD architectural roles

Brokers provide a single point of access for Context Consumers. They provide access to the information they store themselves and/or the information they can retrieve from Context Sources registered in the Registry Server, aggregating the information for Context Consumers.

Context Producers create, update and delete information in the Broker. They cannot be directly accessed. Context Sources implement the query and subscribe operations of the NGSI-LD API and thus give access to the information they can provide themselves. To be found by Brokers, they register themselves with their access point and what kind of information they can provide. As Brokers implement query and subscription functionality required for Context Sources, a Broker can also act as a Context Source.

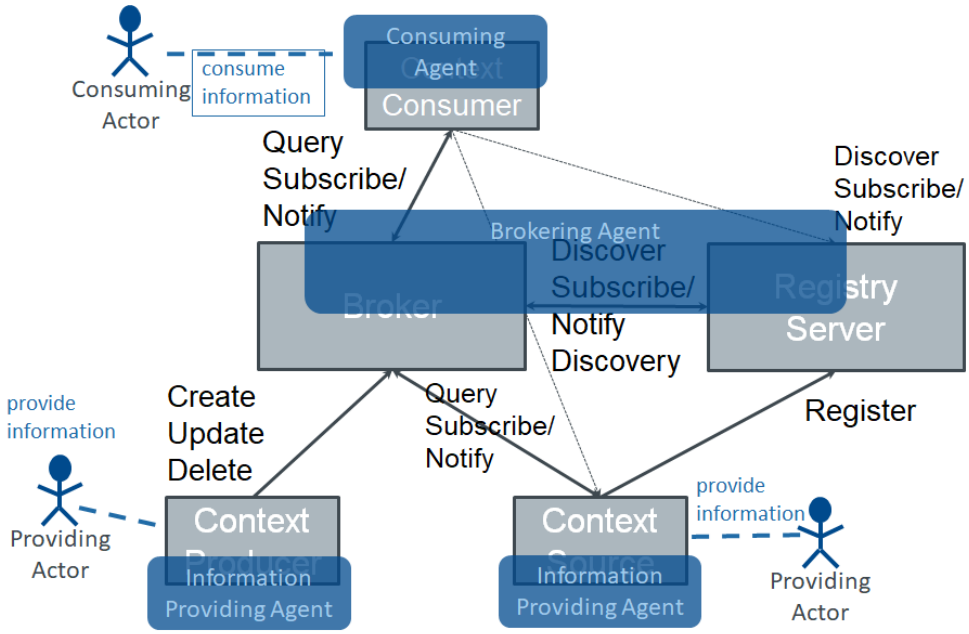

Figure 5.5-2 Actors and Agents

Figure 5.5-2 shows that agents act on behalf of actors. Actors are humans or legal entities, who operate the agents and are legally responsible for their actions.

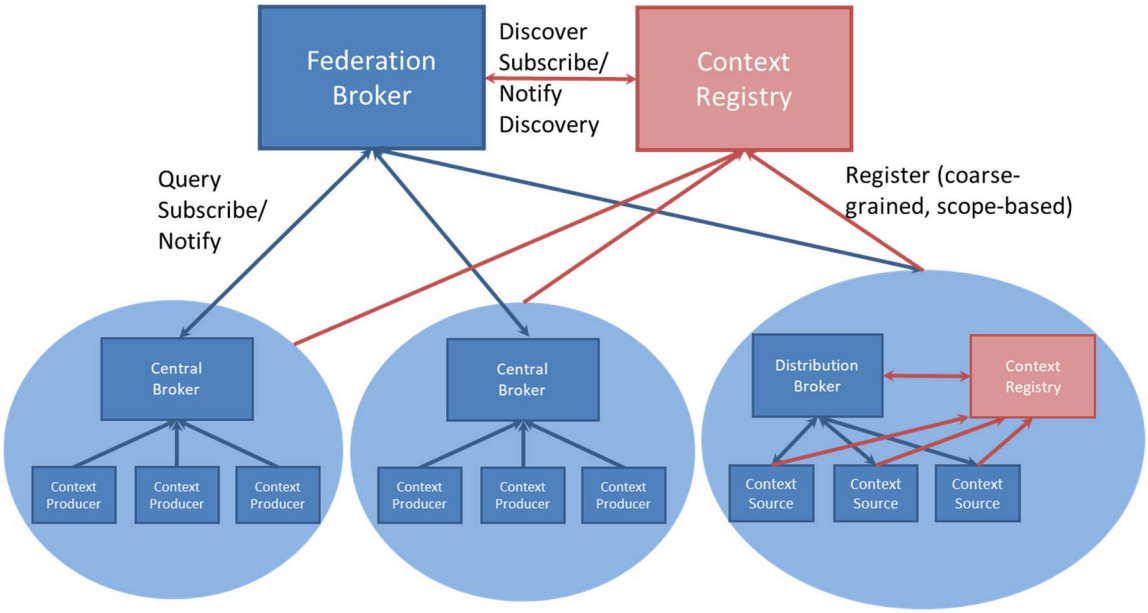

Figure 5.5-3 Federation Example

Figure 5.5-3 shows an example of the federation of NGSI-LD Systems. As can be seen, the Context Sources registered in the Context Registry on the Federation Level are complete Brokers that are themselves responsible for a domain, which could also be a whole federation again, i.e. it is possible to build hierarchical federation structures. From a security perspective, the whole NGSI-LD federation could be a single security domain, or it could also be a federation of security domains.

5.6 Open vs. Closed Deployments

An open NGSI-LD deployment is distinguished from a closed NGSI-LD deployment in the following characteristics:

- Where data is stored (e.g. single data store, multiple data stores, on private cloud storage, on shared cloud storage)
- Resources used while in-transit (e.g. all transit connectivity owned and maintained by the deploying entity, public shared resources)
- Being worked on by remote applications, prior to consumption

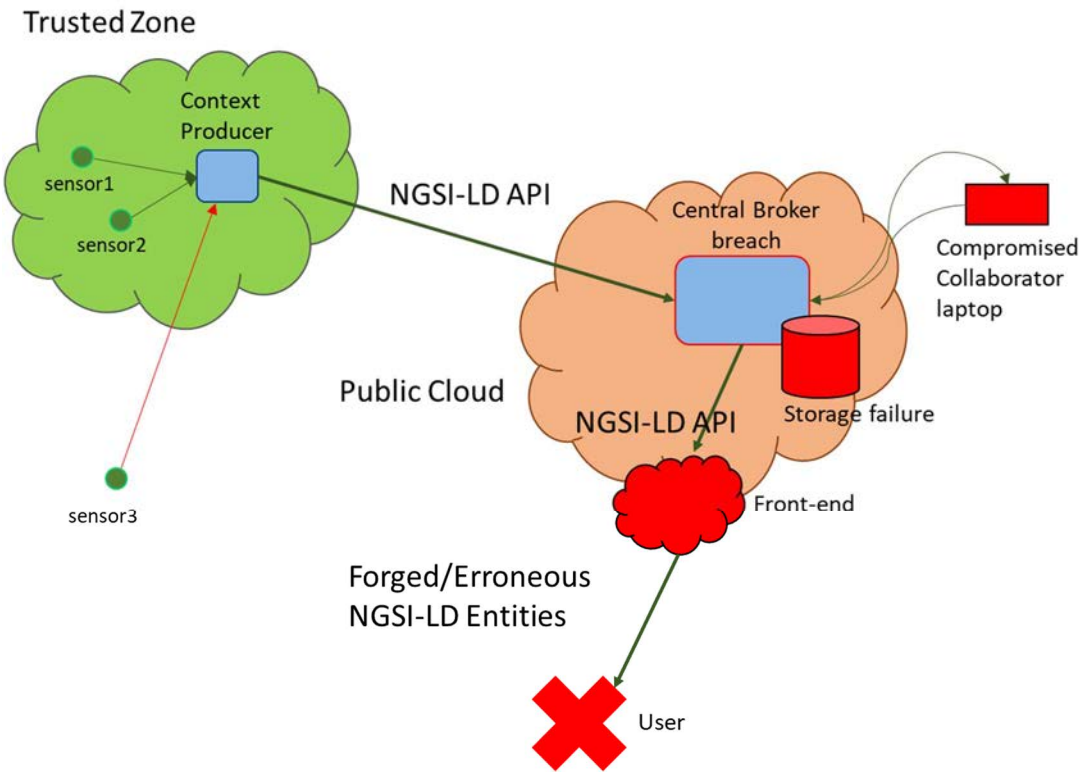

Figure 5.3-1: Examples of security threats impacting an open NGSI-LD deployment

Figure 5.3-1 is a synthesis of the typical open deployment where a User consumes NGSI-LD Entities, which can possibly undergo forging or errors due to threats manifesting themselves within the boundaries of the untrusted zone.

In deployment, the intent is to always operate within a bounded trust zone. Where more than one organization is required to cooperate in management of the trust zone, as for an open deployment, the trust zone policies will become more complex by default. However, the mechanisms used to verify the CIA attribute set are to a large extent identical although the scope and extent of where threats are introduced to the deployed system are more numerous (commonly referred to as an increase in threat surface).

EXAMPLE 1: When considering confidentiality the suite of mechanisms available serve to provide segregation of Alice's data from Bob's data. If provided by encryption there are many ways to achieve this in terms of key agreement strategy, algorithms, algorithm modes, but ultimately the data, when encrypted by a key known only to Alice can only be seen by Alice and with whomsoever she has shared the decryption key with.

EXAMPLE 2: The means by which integrity is verified is in part driven by path conditions and the data. Schemes such as cryptographic hash functions do prove that data has not been modified in such a way that the hash of the received data does not match the hash of the transmitted data, but require that the hash is generated and preserved (e.g. signed) independently of the data it seeks to verify.

EXAMPLE 3: Authentication can be achieved in a number of ways such as by proving knowledge of a shared secret, or by having attestations supported by a 3rd party. The result is that Alice is assured that Bob is really Bob irrespective of the path taken to get to that result.

To summarize: The choice of an open or closed deployment of NGSI-LD System is a deployment choice only and the same standards apply in each instance. The risks are however different in each case and can possibly be mitigated by using different techniques.

6 Security topics

6.1 Introduction

ETSI TS 102 165-2 [i.9] provides a general overview of mechanisms and protocols in support of the CIA paradigm. In an NGSI-LD API system, represented in UML in figure 6.1-1 (taken from ETSI GS CIM 009 [i.1]) only properties contain values. All identifiers in NGSI-LD are URIs, as mandated by IETF RFC 3986 [i.10].

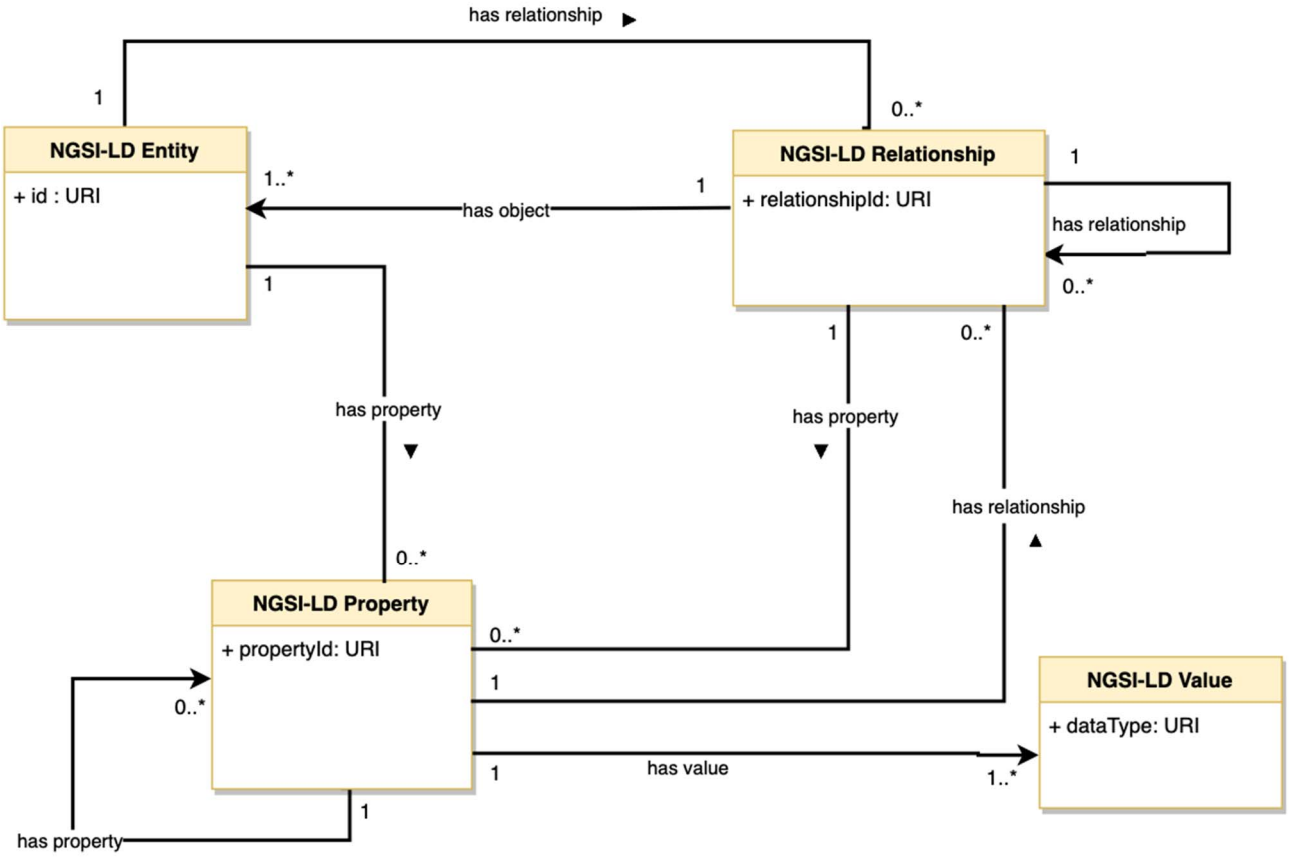

Figure 6.1-1: NGSI-LD information model as UML (from ETSI GS CIM 009 [i.1])

6.2 Identity Management and Authentication

6.2.1 Identity Management

In ETSI TS 187 020 [i.16] and in ETSI TS 103 486 [i.15] identity in ICT systems, of which NGSI-LD is a component, is described as a collection of related identifiers. The identity model of [i.16] and [i.15] maps to the illustration in figure 6.2.1-1, in which a user or a smart object, is represented by a set of identifier types and their assigned value. In ETSI TS 103 486 [i.15] in particular it is stated that not all attributes that contribute to identity need to be exposed. The consequence is that knowledge of any one attribute should not allow an observer to infer knowledge of any other attribute.

EXAMPLE 1: Knowing the location of an object should not allow inference of the object's shape or colour or any other identifying attribute.

In NGSI-LD, the security consequence of identity and identity management is that the principle of data minimization is followed (this is consistent with the obligations on users and providers under GDPR [i.3]). The role of identity management in NGSI-LD is to give assurance of the association of an identifier to an identity in such a way that the binding (association) can be independently verified by means of an authentication mechanism.

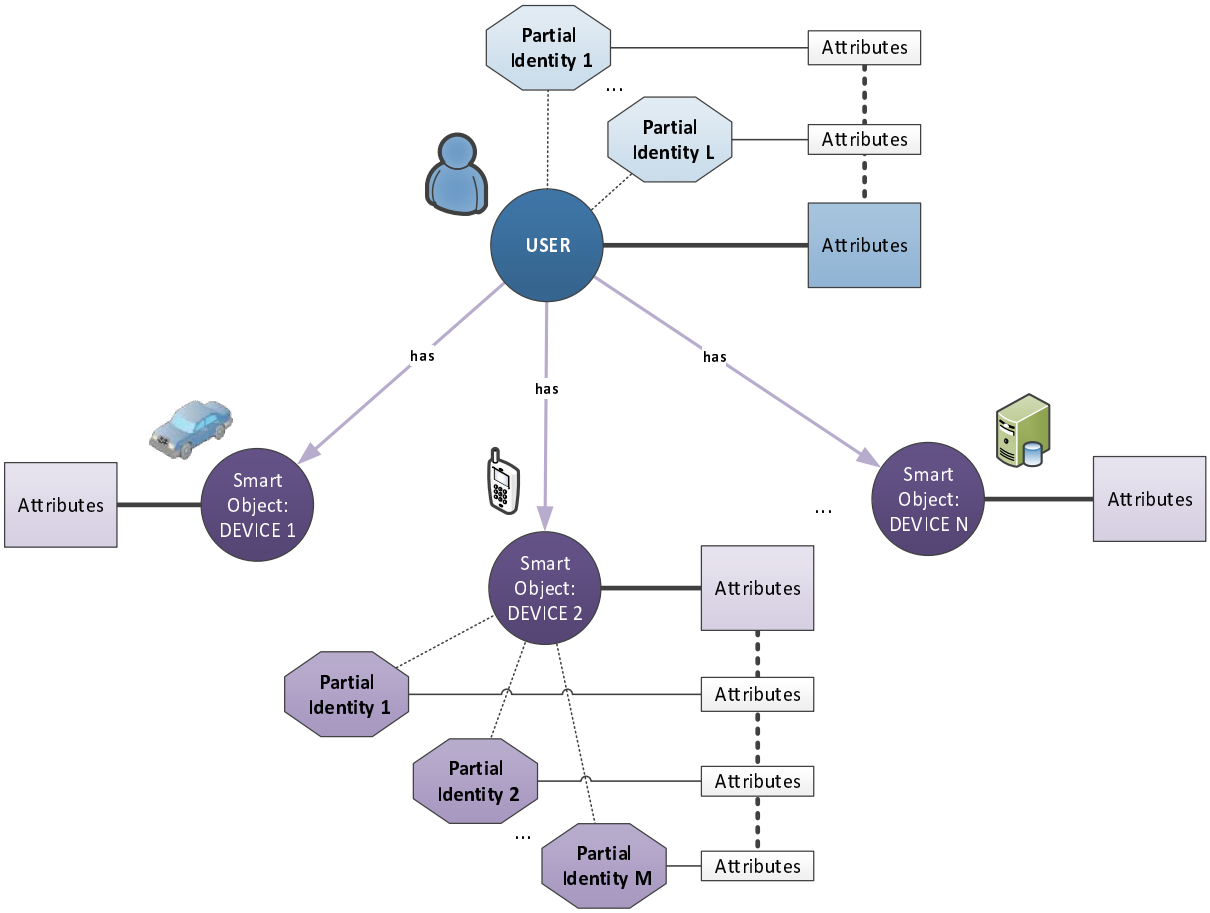

Figure 6.2.1-1: Identity as a sum of identifiers

In NGSI-LD all identifiers are of the type URI and specified in Annex A of ETSI GS CIM 009 [i.1] to exist within the NGSI-LD namespace.

EXAMPLE 2: In Annex A of ETSI GS CIM 009 [i.1] it is stated that an NGSI-LD Entity of Type *Vehicle* may have a Property named *licencePlateNumber*. In any practical model of a *Vehicle* it will also have static attributes including *marque*, *model*, *modelVariant*, *vinNumber*, *colourExternal*, and a number of dynamic attributes including *currentSpeed*, *currentHeading*, and so forth.

EXAMPLE 3: The ITS Common Data Dictionary specified in ETSI TS 102 894-2 [i.17] may be incorporated and made available as JSON data structures.

Some attributes of an NGSI-LD Entity may only be assigned by specific authorities.

EXAMPLE 4: An NGSI-LD Entity of Type Vehicle with a property of *licencePlateNumber* should have the value assigned to that property attested to by a relevant vehicle licencing authority. If the value of licence plate is only assigned by the vehicle licensing authority after validation of all other static attributes, then that single authority may be able to attest to all properties of that NGSI-LD Entity.

NOTE: In some jurisdictions some identifiers are considered as personal data, for example a vehicle licence plate identifier is a pseudonym for the registered owner, a mobile phone number is a pseudonym for the user. This is a result of the normal (societal, cultural) association of object to person.

6.2.2 Authentication

The role of authentication is such that Bob can be assured that Alice is Alice and that Eve cannot masquerade as Alice. Mechanisms for giving proof of authentication are described in ETSI TS 102 165-2 [i.9], and in the context of attribute based cryptography in ETSI TR 103 719 [i.19], with a number of variations described in the publications of multiple standard bodies (see bibliography), and in product offerings.

In the NGSI-LD API context authentication is used in part to verify the identity of each actor in the NGSI-LD API system.

NOTE: The underlying mechanisms of NGSI-LD that extend from JSON-LD and thence from JSON may inherit some of the functionality of the JSON Web Token and in turn on JSON Web Signature and JSON Web Encryption.

6.3 Authorization and Access Control

The role of authorization in the context of access control is to ensure that Bob can verify that Alice is allowed to perform a requested action. In this case Bob is the custodian and Alice is the requestor. Whilst an academic review of access control identifies that there are 2 primary approaches to access control, Discretionary and Mandatory, the more conventional view is that access control is driven by a set of policies that specify who or what is allowed to access data and under what conditions. Thus access may be restricted by requestor identity, by requestor location, by requestor role, by date and time and so forth. As part of the access control policy the custodian (Bob) may require that the requestor (Alice) proves she has authorization. Authorization may be granted directly based on attributes of Alice or may be granted only after 3rd party verification of authority.

NOTE: Access Control and Authorization are somewhat synonymous, access control verifies that Alice has authority to perform the action requested. If authorization is verified, access is granted, else access is denied.

Access Control and Authorization are often dependent on other processes including Identification and Authentication, i.e. access control requires that Bob (as custodian) is able to verify the identity and other credentials or attributes of Alice (as requestor) prior to applying the details of the access control policy.

The access control policy should identify in precise detail the conditions under which a resource can be made available. An exemplar for access control is the XACML model [i.18], which defines architectural elements to enforce the policy (e.g. Policy Enforcement Points (PEPs), Policy Decision Points (PDPs)). In addition the techniques of Attribute Based Cryptography (ABC), often in collaboration with Identity Based Cryptography (IBC), can be integrated to an access control policy, see ETSI TS 103 352 [i.20] and ETSI TS 103 485 [i.14].

6.4 Data Confidentiality

Confidentiality is a tool in which the information content of data sent from Alice to Bob is not visible to Eve.

NOTE 1: The term "sent from Alice to Bob" is often used in the security world to refer to data stored by Alice at time t to be retrieved at a time t+d, i.e. for later recovery/retrieval, where Bob refers to Alice in the future.

NOTE 2: Whilst confidentiality protection is a tool in privacy protection they are very different things. If Bob on receiving data legitimately from Alice chooses to make it public against the implicit or explicit agreement of Alice that may be a violation of Alice's privacy. The role of confidentiality protection for the data in transit is therefore independent of the wider obligation on Bob of maintaining confidentiality after receipt.

In many instances data confidentiality is afforded by the application of encryption to content, although techniques such as path segregation (e.g. VPNs) and the use of secure enclaves in data processing are other measures that can be applied to give assurance of data confidentiality.

6.5 Personal Data

Personal data, or personal identifying data, is any data that can be used to uniquely identify a person. Where personal data is identified, regulatory frameworks such as GDPR apply. As described in article 35.4 of GDPR [i.3] the data controller and relevant organization is expected to perform a Data Protection Impact Assessment as part of the GDPR compliance that is intended to identify if data held and processed is classified as personal.

Where personal data is identified there should be technical and organization means to ensure it is processed in accordance with GDPR (e.g. anonymized, encrypted, restricted) wherein the mechanisms for access control also apply.

However, full identification of the actors in GDPR (e.g. data controller, data processor) is not a technical consideration but an organizational one. Thus, whilst it may be reasonable to identify the broker in the NGSI-LD API architecture of figure 5.2-1 as the Agent of a data processor as defined in the GDPR, this role cannot be granted without reference to the data controller and to the explicit policies of the organization.

6.6 Data Integrity

In the context of the CIA paradigm the role of integrity is to give assurance to Bob that on receipt of data from Alice that what he receives is exactly as sent by Alice and that the data has not been manipulated by Eve. In addition, there is a close relationship between integrity and availability with the defining characteristic that the system delivering data is complete and has itself not been corrupted/manipulated by Eve. Integrity is therefore a key characteristic of non-repudiation services.

6.7 Trust between Multiple Federated Stakeholders

In a distributed network with relatively open interfaces offering access to resources at each node, the purpose of access control is, in part, to ensure that resources are only used by known and authorized parties. A failure of access control has the potential to deny access to resources to legitimate parties and may be used as part of an attack vector to enable masquerade of entities, or to allow manipulation of the content of operation of resources.

The access control policy has to also address data architectures including federated data access, de-centralized access control, and others as identified in ETSI GS CIM 009 [i.1]. In such cases, policies from more than one stakeholder and at more than one level may need to be addressed, so that the policy model includes, for example, global policies and domain-specific policies.

In figure 6.7-1, a policy management encompassing both local and global policies is shown.

Global data sharing policies: In a federation model, that includes many largely independent domains, it is necessary to maintain a common practice of rules for the federation of data.

Domain-specific data access policies: Each domain is a data owner and it has full rights to set its own policies. These policies are typically about who (external to the domain) can access which data under which circumstances. Importantly, domain-specific policies are not allowed to be conflicting with global policies to ensure consistency for federation in managing and sharing data securely.

Above the global policies, a middle layer (for instance a search functionality) is often times implemented, leveraging them.

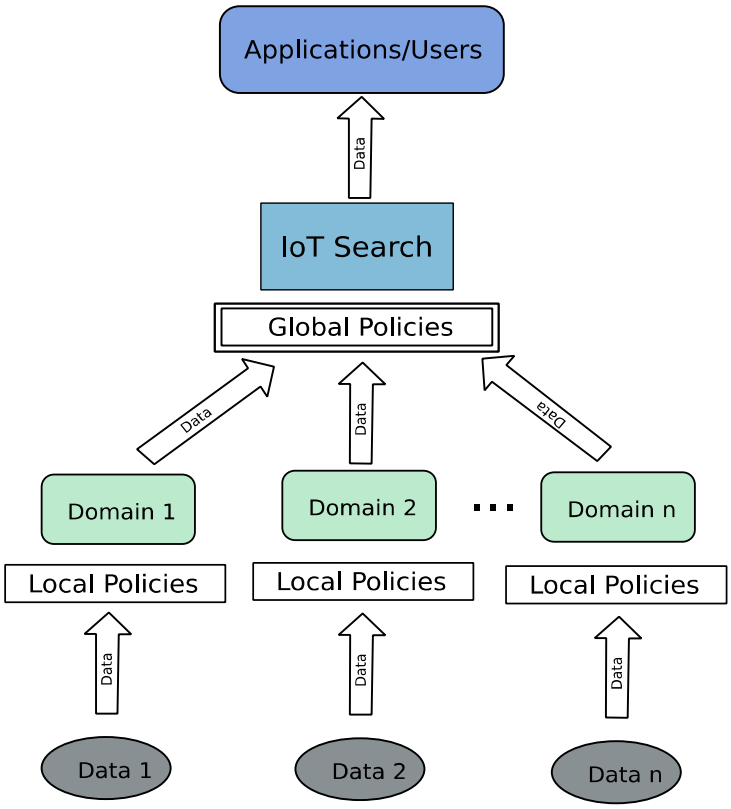

Figure 6.7-1: Federated policy management example

In achieving access control, it is important to consider the role of an access control policy for each federated resource on the network. For example, where multiple organizations work together to offer a service, such as in the example given in clause 4 in which a Parking Management System could query a Traffic Management System about road occupancy at a particular egress gate each of the systems would normally be expected to have an independent data controller and a data policy that defines under which conditions data can be made available between the collaborating systems, see ETSI TS 103 485 [i.14].

6.8 Multi-tenancy

The NGSI-LD API enables NGSI-LD Systems to support the concept of multi-tenancy. The idea is that a single software instance can serve multiple groups of Users, each belonging to a tenant, as if they each had their own software instance. For a User of one tenant, the Users of other tenants, their requests and their information are completely invisible. There is complete isolation between the tenants, which can, for example, be realized by storing their information in different databases. NGSI-LD Systems have the concept of a default tenant, i.e. if no tenant is specified in the request, the request is targeted to the default tenant.

To protect access to a tenant, access control also has to take into account tenant information. A User can have access to multiple tenants, but the access rights can be different for each tenant.

7 Desired Security Features

7.1 Introduction

This clause identifies future standardization requirements by identifying specific objectives to be met by future standards that, when implemented, will give an assurance of required control of data operations within the NGSI-LD system and the resilience of the NGSI-LD system to attack.

In the following, the term "system" refers to "NGSI-LD System".

Creating a security perimeter around data relies on understanding which is the sensitive data: a data classification is needed to help identify and mark which data is sensitive. In turn, identification of sensitive data requires knowledge of the context in which to:

- mask parts of data, ensuring that access to sensitive data does not potentially expand the attack surface
- comply with regulations
- apply Rights Management to different portions of data

Context Information Management and data-centric protection go hand-in-hand and cover a broader variety of security use cases, including completing security in open deployment scenarios.

7.2 Identity Management (IdM) and Authentication

The full model of identity management for NGSI-LD should follow the general model of ETSI TS 103 486 [i.15] wherein each attested User attribute can be independently verified.

The following security objectives, which are considered to be core objectives of an NGSI-LD system, are related to Identity Management and Authentication as described in clause 6.2:

- **IDM-1:** The IdM subsystem should be able to manage the Identity of a User as the collation of a set of independent attributes

EXAMPLE: Identifying attributes include the User's name, role, company, address.

- **IDM-2:** The IdM subsystem should be able to define groups of attributes in order to enable attribute based access controls
- **IDM-3:** Identities when represented as a collation/collection of attribute values should be designed in such a way that knowledge of any individual attribute does not infer the value of any other attribute associated to the User's full Identity.
- **IDM-4:** The IdM architecture should be designed in a manner that any architecture for the NGSI-LD API defined in clause 4.3 of ETSI GS CIM 009 [i.1] is supported.

7.3 Authorization and Access Control

The following security objectives are related to Authorization and Access Control as described in clause 6.3. It is strongly recommended to maintain discrete rules and policies for each of the Administrator and User policies. An access control policy should be able to support one or many rules:

- **AC-1:** the functionality exposed through the NGSI-LD API should be managed in the form of Access Control Policies.
- **AC-2:** Access Control Policies should be structured according to the 4-tuple: <subject, action, resource, condition>.
- **AC-3:** An access control policy should be able to support one or many rules.

The following examples are illustrative of the forms of access control granularity to be addressed.

EXAMPLE 1: An access control policy rule is specified to limit access to a specific user.

EXAMPLE 2: An access control policy rule is specified to allow public access.

EXAMPLE 3: An access control policy rule is specified to limit access to a specific user roles (e.g. building manager).

EXAMPLE 4: An access control policy rule is specified to limit access where a to a specific user.

Table 7.3-1 provides a number of detailed security objectives, consistent with the above examples, which are expected to be met within NGSI-LD systems.

Table 7.3-1: Subjects, Actions, Resources and Conditions to be supported in Policies

Subject	Specific Users	**AC-3.1:** Policies are able to allow access to a set of Users, who are registered (authenticated) users of the system. E.g. access is opened to user1, user2 and user3.
	Public access	**AC-3.2:** Policies are able to define public access to any Resource.
	Registered Users access	**AC-3.3:** Policies are able to define public access to any Object.
	User Groups	**AC-3.4:** Policies are able to allow access based on group of Users (e.g. Staff of Company A, friends).
	User Role	**AC-3.5:** Policies are able to allow access based on a specific role of Users (e.g. Manager of a building).
	User Attributes	**AC-3.6:** Policies are able to allow access based on multiple attributes of Users (e.g. e-mail, gender, location, company).
Action	Create	**AC-3.8:** Policies are able to allow creating a Resource.
	Read	**AC-3.9:** Policies enable specifying that reading a Resource is allowed.
	Update	**AC-3.10:** Policies enable specifying that updating a Resource is allowed.
	Delete	**AC-3.11:** Policies enable specifying that deleting a Resource is allowed.
	Subscribe	**AC-3.12:** Policies enable specifying that subscribing to a Resource is allowed.
	Register	**AC-3.13:** Policies enable specifying that registering for a Resource is allowed.
Resource	NGSI-LD Element	**AC-3.14:** Policies are able to define access to the whole data of NGSI-LD Elements (Entity, Registration, Subscription), identified by their NGSI-LD identifier or identifier pattern. E.g. access to all NGSI-LD Entities of id = sensor_room.
	NGSI-LD Type	**AC-3.15:** Policies are able to define access to the whole data of NGSI-LD Entities, identified by their NGSI-LD Types. E.g. access to all NGSI-LD Entities of type = water pipe.
	NGSI-LD Attribute	**AC-3.16:** Policies are able to restrict access to a single attribute of an NGSI-LD Entity. E.g. access sensor reading, but not the location.
	NGSI-LD Attributes Group	**AC-3.17:** There is an option to group NGSI-LD Attributes of an Entity, and attach an access. Policy to the group. E.g. access to group NotPrivate (sensor reading and sensor location). Note that enabling the creation of such groups is planned, but not yet implemented in the NGSI-LD API, as of V1.5.1 [i.1].
Condition	NGSI-LD Queries	**AC-3.18:** Policies are able to define conditions via NGSI-LD Queries, so that, if conditions are met, the Policy grants access. E.g. grant access if temperature > 30 ºC, createdAt between 12 p.m. and 1 p.m., location within 1 km from point X, data is not tagged "PERSONAL" hence not subject to GDPR restrictions, and scope is "/Madrid/Gardens ").
	Policy time window	**AC-3.19:** Policies enable limiting the access to Objects to a specific time of the day. Policies are able to restrict the access to Objects to a specific schedule. (e.g User can access ObjectX from 9.00 am to 10.00 am).
	Object time window	**AC-3.20:** Policies enable limiting the access to temporal information to a specific time window. Policies are able to restrict the access to temporal evolution of Objects. (e.g. only last hour data or everything prior to last hour).

- **AC-4:** The reading and managing of Policies is governed by (meta-level) Policies.

Rationale: It needs to be defined who can read and manage policies. For this reason, there have to be meta level policies for reading and managing policies. For these types of policies, there is a clear objective to enable some kind of bootstrapping.

- **AC-5:** To provide the required flexibility and granularity, the Subjects, Actions, Resources and Conditions specified in Table 7.3-2, there is an objective that there be Policies about reading and managing Policies.

Table 7.3-2: Subjects, Actions, Resources and Conditions to be supported in Policies for Policies

Subject	Specific Users	**AC-5.1:** Policies are able to allow access to policies to a set of Users, who are registered (authenticated) users of the system. E.g. access is opened to user1, user2 and user3.
	User Groups	**AC-5.2:** Policies are able to allow access to policies based on group of Users (e.g. Staff of Company A, friends).
	User Role	**AC-5.3:** Policies are able to allow access to policies based on a specific role of Users (e.g. Manager of a building).
Action	Create	**AC-5.4:** Policies enable specifying that creating a Policy is allowed.
	Read	**AC-5.5:** Policies enable specifying that reading a Policy is allowed.
	Update	**AC-5.6:** Policies enable specifying that updating a Policy is allowed.
	Delete	**AC-5.7:** Policies enable specifying that deleting a Policy is allowed.
Resource	Policy	**AC-5.8:** Policies are able to define access to Policies.
Condition	Administrator Policy	**AC-5.9:** An Administrator Policy is considered to apply to all associated Resources, unless there is a fitting User Policy.
	User Policy	**AC-5.10:** User Policies only apply to Resources created by the given User and in such case take precedence over Administrator Policies.

Explanation for AC-5.9 and AC-5.10: Two types of access control policies are defined: Administrator and User policies:

- A User Policy is a Policy, where the pool of applicable Resources (what can be queried) is automatically restricted to the set of data inserted by the User who is formulating the Policy, regardless of how the Policy is formulated.
- An Administrator Policy is a Policy, where the pool of applicable Resources is the whole set of current data in the Broker.

7.4 Data Confidentiality

The following security objectives are related to Data Confidentiality as described in clause 6.4 and complement the data confidentiality provisions made by access control and authorization:

- **DC-1:** It should be possible to encrypt the value of NGSI-LD properties when returned to either a synchronous or asynchronous NGSI-LD API query.

NOTE 1: If the value of any property is required to be known to an intermediate node (e.g. to a broker) the value cannot be encrypted using a mechanism where only the end point can decrypt it.

NOTE 2: Mechanisms that allow a party to decrypt only part of the value of an attribute and that allow authorized middlebox like entities (e.g. a Broker) to see relevant values of any property can be specified.

- **DC-2:** It should be possible for authorized parties to determine if a value of an NGSI-LD Property is encrypted.

In NGSI-LD Systems, data confidentiality is also given through the explicit access control restrictions also described in clause 7.3.

7.5 Personal Data

The following desired security features are related to Personal Data as described in clause 6.5:

- **PD-1:** NGSI-LD should provide in its cross-domain core ontology the means to express metadata about personal data and sensitive personal data.

Rationale: To control access to personal data, it is necessary to know whether certain data has been identified as personal data or even sensitive data:

- **PD-2:** Policies should be able to encode consent, so as to take it into account when controlling access to personal and sensitive information.

Rationale: When controlling access, it is important to know, whether a user has given consent to access personal information, to whom and under what conditions:

- **PD-3:** Policies should be able to access both consent information and type of personal data, in order to make access control decisions.

7.6 Data Integrity

The following desired security features are related to Data Integrity as described in clause 6.6:

- **INT-1:** The data consumer should be able to determine that data integrity has been preserved.
- **INT-2:** Verification of integrity should be independent of syntactical re-ordering that may occur when serializing NGSI-LD data between peers.
- **INT-3:** Verification of integrity should be independent of the serialization format itself.

7.7 Trust between Multiple Federated Stakeholders

The following desired security features are related to trust in federated scenarios, as described in clause 6.7:

- **FED-1:** Security capabilities are modular. For example, some Context Sources may not have full "security capabilities" (e.g. be able to evaluate all kinds of policies): they could delegate trust to the Distribution Brokers, such that some policies are evaluated there.
- **FED-2:** Users and applications should be able to access the federation using single identity and authentication information, regardless of the entry point.
- **FED-3:** NGSI-LD should provide consistent authorization and access control policies and rules across the federation.
- **FED-4:** Since NGSI-LD Attributes of the same NGSI-LD Entity can be fragmented across different peers of the federation and different domains, data integrity mechanisms should be able to cope with fragmentation.
- **FED-5:** The key distribution techniques and infrastructure (e.g. for signature verification) should be differentiated, depending on the level of trust among peers, and whether peers are within the boundary of the federation or not.

7.8 Multi-tenancy

The following objectives are related to multi-tenancy as described in clause 6.8:

- **MT-1:** Access control, as discussed in clause 7.3, is expected to support isolation also for the NGSI-LD access control policies themselves.

This also means that the access to the policies themselves has to be protected, i.e. that the meta-level policies have to be defined on a tenant basis.

Annex A:
Use Cases supporting security provisions for NGSI-LD API

A.1 Motivation

In Annex B of ETSI TS 102 165-1 [i.8] the role of motivation with respect to system protection and attack is addressed. The following text addresses what may be considered as a set of motivation paradoxes:

- The likelihood of an attack:
  - If a threat is highly motivated, an attack can be considered imminent.
  - If a threat is unmotivated, no attack can be anticipated.
- The value of the asset, monetarily or otherwise, to either the attacker or the asset holder:
  - An asset of very high value is likely to motivate an attack.
  - An asset of little value is unlikely to motivate an attack.
- The expertise and resources with which an attacker is willing to effect an attack:
  - A highly motivated attacker is likely to acquire sufficient expertise and resources to defeat the measures protecting an asset.
  - An attacker with significant expertise and resources is not willing to effect an attack using them, if the attacker's motivation is low.

In each case, there is no probabilistic means of determining the role of motivation in mounting an attack. However, in assessing threat potential, it is essential to consider motivation in order to minimize the effect of motivation on the attacker.

As it is largely impossible to state with any certainty that a system will be attacked, it is however equally impossible to state with any certainty that protection is essential. With protection, a system when attacked will be resistant to that attack in some measure.

A.2 Use case: Emergency Situation in Smart Buildings

The potentials of smart buildings have been put into perspective as the basic building block of smart cities. These buildings are usually fully equipped with different kinds of sensors and actuators such as: access control mechanisms (e.g. based on smart cards) in each door, presence sensors, luminosity sensors, humidity sensors, HVAC and lighting systems to name a few. In addition, information about the users of the building are usually registered in the system, which can comprehend not only generic details such as *Name*, *Role* or *Area*, but also *sensitive information*, such as if the user *has mobility restrictions*.

The usefulness of having a selective access control mechanism based on the attributes of the identities stored in the system is presented in this use case. This technology is also used together with an encryption mechanism which allows for a secure broadcast of information.

In the case of an extraordinary event, which requires the evacuation of the users of a smart building, the access to sensitive information can be very useful to prioritize the order of the evacuation.

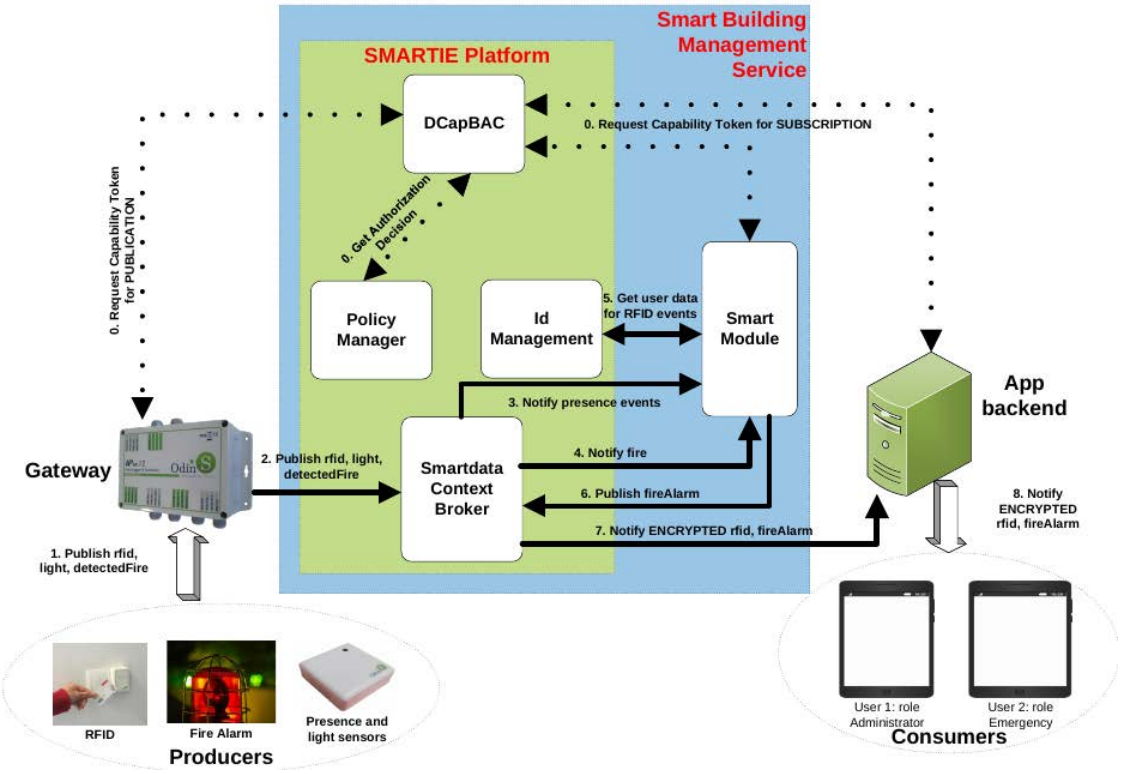

Figure A.2-1: Emergency Management in Smart Buildings

In case of a fire, the emergency staff will have access to sensitive information of the users, this information together with their location (obtained or estimated through the use of the deployed sensors inside the building). Moreover, the information is transmitted in a secured way, which allows only the legitimate consumers to access the information. Figure A.2-1 shows the different modules and how they interact, in case a fire has been detected.

A.3 Use case: Processing Medical Data and eHealth Applications

Interconnecting siloed repositories of clinical and genomic data is becoming more and more crucial [i.11] to infer new insights and carry out research on the data sets by querying and joining information coming from different sources. Three fundamentally different approaches can be conceived for this application:

a) Centralized approach: move data to central repository and let it process the data; but trust is expected to be fully delegated to it, so that it represents a single point of trust and a potential security and privacy threat.

b) Decentralized approach: data is kept on-premise and accessed through a peer-to-peer network; there is no need to trust other parties because private data never leaves the clinical site unless pseudonymized, and clinical sites can enforce full control on data access; but maintaining the infrastructure is costly.

c) Hybrid approach: distribute the trust among a set of different "storage and processing" units (for instance at the national level), which form a secure, federated and interoperable network that investigators can query for research purposes as if it were a single unified database.

Because of their shortcomings neither a) nor b) has been fully adopted by the healthcare sector.

Requirements

- Process the distributed data, so as to obtain (pseudononymized) information about patients matching various medical criteria. Aggregate data based on medical criteria.
- Trust decentralization, with no single point of failure.

- **End-to-end data protection at rest, in transit and during computation:** data is encrypted at the clinical site and query results can be decrypted only by the investigator issuing the query.
- Threat model.
- Storage and Processing Units (SPU) are *honest-but-curious* parties. SPUs can be compromised by internal or external adversaries that do **not tamper with the data-sharing protocol** but can try to infer sensitive **information about the patients from the data stored** at their premises and from the data being processed during the protocol itself. As a result, SPUs cannot be trusted by clinical sites and they do not trust each other, either.
- Investigators are potentially *malicious-but-covert* adversaries. An investigator can try to legitimately use the system in order to infer sensitive information about the patients (without being discovered) by performing consecutive queries and exploiting the information leaked by the end-results. For example, a malicious investigator with some background information about a given individual will try to infer the presence of such an individual in a sensitive cohort (e.g. patients who are HIV-positive) or even reconstruct a subset of her medical record.
- Clinical sites are trusted parties.
- Investigators cannot collude with SPUs
- SPUs can collude with the other SPUs.

Mapping to NGSI-LD

The MedCo [i.11] system fits the federated NGSI-LD architecture:

- Clinical sites are the Context Producers.
- Investigators are the Context Consumers.
- Storage and Processing Units are the Brokers, acting both as Central Brokers of a domain and as Context Sources.
- All of the clinical sites that refer to the same SPU represent an NGSI-LD domain formed by the Context Producers with their Central Broker. Each Central Broker registers as a Context Source to a Context Registry.
- There is the need for an additional component, a Context Registry, which keeps track of all SPUs (Central Brokers) that are part of the federation. Since in MedCo all SPUs potentially have interesting information, queries are always broadcast to all of the SPUs. Hence the Context Registry is going to register each Central Broker as a Context Source, and it can be seen as trivially distributed to each SPU in the MedCo.
- A Federation Broker is introduced, acting as a unified entry point for Investigators to address queries to. This is not necessary in MedCo for the same reason as above, i.e. all queries are broadcasted to every SPU, regardless of the entry point.

Figure A.3-1 captures both the threat model and the mapping with NGSI-LD.

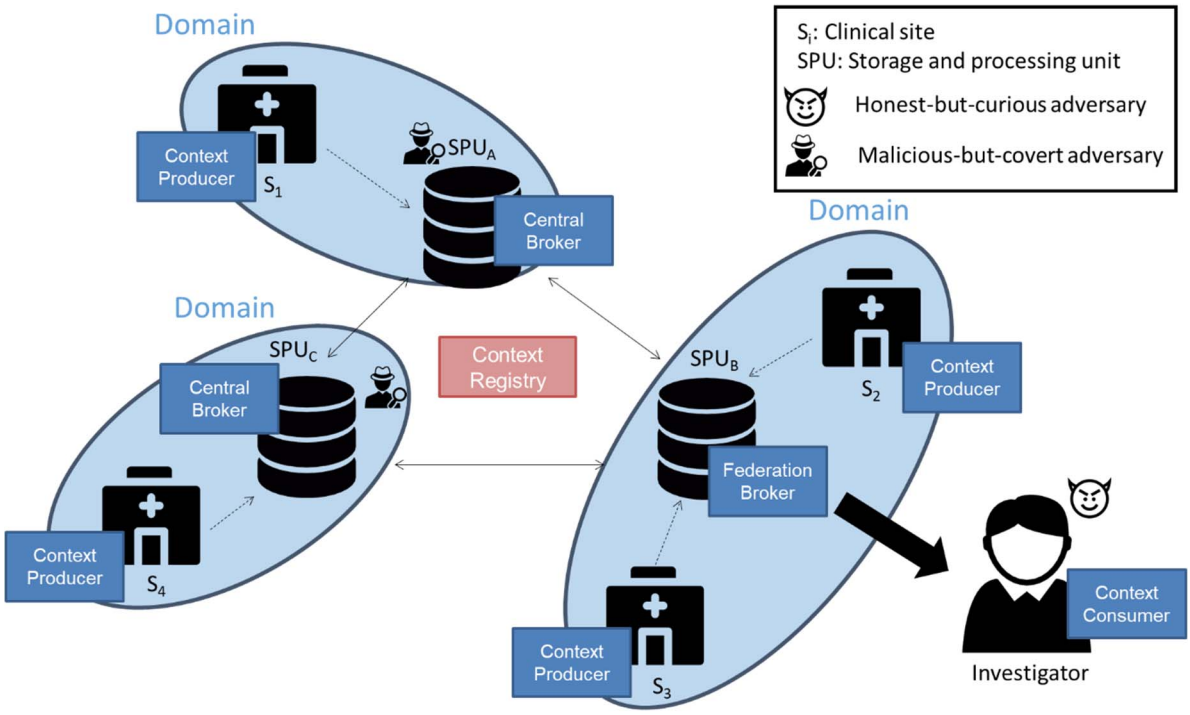

Figure A.3-1: Threat model and mapping with NGSI-LD

Discussion and input to NGSI-LD

In MedCo, interoperability is achieved by forcing all data to adhere to a common format, which is syntactically based on the Entity-Attribute-Value concept (EAV) that is widespread in clinical research systems such as i2b2 [i.12]. In terms of semantics, there are a patients table and a facts table, which contains observations on the patients. This is perfectly aligned with the NGSI-LD data model.

From the security and privacy standpoint, clinical sites want to protect the mapping between patients and facts, while still allowing some processing and aggregations of data. Several techniques and solutions are applied in MedCo, in parallel, to enable that:

- Clinical sites perform data pseudonymization before pushing it to the storage and processing units.
- Clinical sites generate a set of both dummy patients and dummy observations, in order to achieve uniform distribution in the facts table.
- The system uses DTE (Deterministic Encryption) techniques to perform equality searches on encrypted data in the same way as it would on the plaintext. The system uses homomorphic encryption that supports computation on encrypted data.
- Depending on the type of query and the Investigator's privileges, each SPU can obfuscate the encrypted computation results, by homomorphically adding noise, in order to guarantee differential privacy.

In federated big data scenarios, for security and privacy reasons, data processing, aggregation, analysis is often to be carried out in a trusted way, far from the data consumers. Data consumers may be allowed to get, for instance, the COUNT number of patients affected by a certain condition in a certain area, but may be forbidden to get the full list themselves and perform the GROUP BY on their own. Hence, the GROUP BY operation has to be carried out by the SPU.

This means that Brokers would have to be able to perform more complex operations than currently supported in the NGSI-LD API, in order to fully implement a scenario such as MedCo.

Furthermore, a Broker could incorporate standard procedures to perform differential privacy while serving results.

A.4 Use case: International data integration strategy for Earth System Grid Federation

The Earth System Grid Federation (ESGF) [i.13] is an international collaboration to create open-source software and infrastructure that empowers the study of climate science and exa-scale climate data. The goal is to design and implement an international integration strategy for data, database and computational architecture.

Requirements

- Facilitate traceability of scientific results: for data providers, being able to control and **track utilization** and **receive appropriate credit** for their contributions; for data users who expect easy discovery and access to data, including the information necessary to **understand and use the data** in an appropriate manner.
- Provide **quality of the data indication** when the data served is coming from individual scientists' data sets and it may not be as reputable and authoritative as those published by modelling or data centres.
- Preserve data producers' visibility in science applications.
- Provide services to either allow users to select and retrieve only the specific data subsets of interest, or provide the analysis capability within the infrastructure.
- Provide federation of services: a user should be able to search, discover, download, and analyse data hosted at different centres as if they were served from a single location.
- Provide unified access control: a user or a software client should not be asked to authenticate or be authorized separately at all centres. Rather, the system infrastructure should support Single Sign On.
- Individual administration of local resources: at the same time, a centre or project should be able to define its own policies for accessing a certain class of resources.
- No single point of failure: the infrastructure is to be designed in such a way that interruption of services at one centre will have minimal or no impact on the services offered by other centres.

Threat model

- Vulnerable Node may join the federation, and be compromised from the inside.

Mapping to NGSI-LD

The ESGF system nicely fits the distributed NGSI-LD approach.

- Each Node contains climate data, and acts as a Context Source that keeps data without moving it outside.
- Scientist are Context Consumers.
- The XML Registry (which contains service endpoints and SSL public keys for all Nodes in the federation) can be mapped to the Context Registry.

Since ESGF acts more like a peer-to-peer network, there is no need to introduce a Distribution Broker that would act as a single entry-point for querying the system. Each Context Consumer asks the Registry for relevant sources instead, and then manages the query distribution and aggregation of results on its own, based on the information received from the Registry.

Figure A.4-1 explains the mapping.

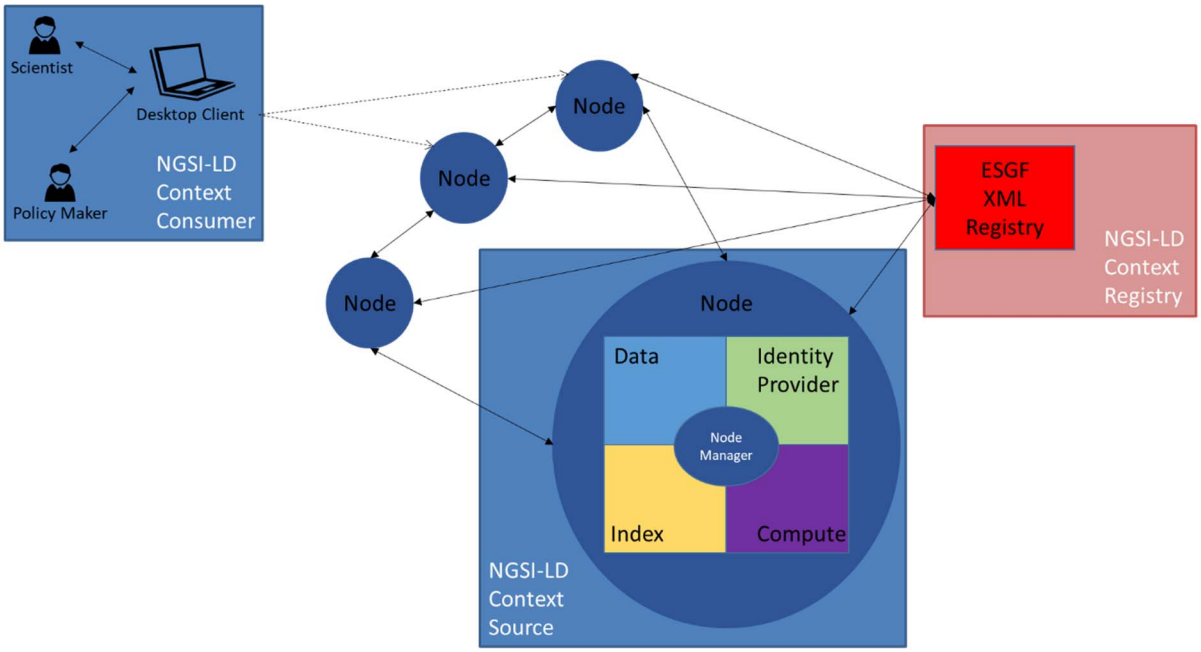

Figure A.4-1: ESGF Context System Architecture

Discussion and input to NGSI-LD

The requirements of ESGF point to two clear directions that are relevant to NGSI-LD:

- There should be some capability to capture cross-domain descriptive and usage information about the data sets.
- There should be the capability to guarantee and verify provenance of data sets.

History

Document history		
V1.1.1	March 2022	Publication