Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: 0005-1144 (Print) 1848-3380 (Online) Journal homepage: https://tandfonline.com/loi/taut20

Geographic dependency of identity-associated data

Petar Djerasimović

To cite this article: Petar Djerasimović (2018) Geographic dependency of identity-associated data, Automatika, 59:3-4, 340-348, DOI: 10.1080/00051144.2018.1530827

To link to this article: https://doi.org/10.1080/00051144.2018.1530827

Published online: 23 Nov 2018.

AUTOMATIKA
2018, VOL. 59, NOS. 3-4, 340–348
https://doi.org/10.1080/00051144.2018.1530827

REGULAR PAPER

OPEN ACCESS

Geographic dependency of identity-associated data

Petar Djerasimović

Department of Applied Computing, Faculty of Electrical Engineering and Computing, University in Zagreb, Zagreb, Croatia

ABSTRACT
A consequence of the proliferation of private and identity data in a globalized world is the emergence of geographically dependent data representations – data in various systems in the world cannot always be captured and processed in the exact same form and individuals often transform data they are making available to systems so it would conform with locally used scripts, languages or constraining rules. Data representation as well as semantics of that data may vary across language and state domains and subjects associated with that data may in different locations be granted different rights stemming from that data. Current systems are mostly built on an attribute-based model of identity and we propose extending this model to include this naturally occurring dependency. Analysing some of the most popular protocols and standards for identity management in current use we have grouped frequently used identity attributes according to their geographic dependency to illustrate what kind of geographic dependencies can be found. We argue for a simple model where data representation and semantics are dependent of the geographic location of data interpretation. Also, we provide an example of extending an existing protocol of identity data exchange with the introduced geographic aspects.

KEYWORDS
Identity management; IdM; identity model; security; privacy

1. Introduction

The management of personal and identifying information significantly predates the dawn of the digital age. However, the volume and detail in which such data are proliferated today as a consequence of available digital technologies is a new phenomenon noticeably impacting daily life with new opportunities and risks attached. In today's complex environments, many systems and services require personal information of various levels of privacy to perform their function. Consequently, this establishes a diverse set of scenarios that necessitate a controlled organization of information storage and exchange.

Today, most individuals are aware that data about them are being collected openly in the public sector for administrative or healthcare purposes but also more frequently by private entities with vested interest in collecting data on individuals. Be it small shops hoping to grow business by better understanding their customer base or entities harvesting personal information en-masse (and sometimes less openly), most individuals feel entitled to assert some measure of control over their personal information. Accomplishing both of these goals – providing the necessary data to streamline services dependent upon them and protecting such data that are sensitive in nature is one of the key requirements of Identity Management (in further text IdM) systems. Of course, as for any other system handling data, one of the fundamental requirements for and IdM system is to store and communicate data that is as authentic, valid and complete as possible.

Another relatively new (in historic terms) phenomenon is the growing geographic mobility of people, a facet of the process today often referred to as globalization. Professionals and tourists, owning partly to available means of cheap transportation are more able and intent on crossing international borders and taking temporary, prolonged or permanent residency in foreign countries and often foreign cultures where they are frequently required to interact with local entities, both public or private. These may include economic immigrants requesting residency rights, various foreign professionals expecting prolonged contact with domestic administration, like doctors without borders, reporters onsite in foreign territories, professional soccer players on a seasonal lease abroad or simply tourists making a purchase at a shop interested in the shop's newsletter or leaving personal information for warranty purposes.

Since the mentioned proliferation of personal data is globally present and considering the significant differences in official languages of the world, both in scripture and in sound, it is obvious that data on individuals exist in various systems in different forms. Importantly, these data are of different forms and beside simple written values (like a person's name) may include pictures, X-ray scans, fingerprint or other biometric data. A person

CONTACT Petar Djerasimović petar.djerasimovic@fer.hr Department of Applied Computing, Faculty of Electrical Engineering and Computing, University in Zagreb, FER, Unska 3, 10000 Zagreb, Croatia

staying abroad for a prolonged period and expecting to be required to provide information on themselves in a foreign language of significantly different linguistic heritage may pursue an intuitive strategy of temporarily adopting a localized, somewhat transformed representation of their identity data to streamline communication with local data recipients.

These observations indicate that personal data representations inherently depend on geographic localities. Additional geographic dependency stems from the fact that the identity bearers are authorized to different services and subjected to different policies in part based on their current geographic location. This fact may be observed in the currently widely used and globally deployed systems for digital rights management (DRM) governing the use of multimedia content such as film, music, computer games and e-books [1]. Some such systems feature region-dependent policies, also known as geoblocking that allow access to digital content only to users living in certain regions [2].

Therefore, in accordance with the mentioned principle requiring systems to handle data that is as authentic, valid and complete as possible, we conclude that any IdM system applicable across language domains (or technically more feasible: geographic domains) should take this dependency into account. This work explores the possibility to encode geographic aspects of identity in a technically feasible and pragmatic manner with the goal of creating a more complete model of identity and its geographic dependencies.

For clarity, identity data represents not only data identifying individuals but all stored data regarding individuals that are in some measure perceived private or sensitive (like passport numbers, age or medical data) as these data are of concern to respective identity bearers. Also, data regarding personal transactions are considered identity data: it may not be my concern what identifying integer or salt value is stored in a database for my entry (though I do wish that data to be securely handled), but I may be concerned how much information and to whom a system revealed about my purchase that I may perceive as less-than-reputable.

Lately, we are witnessing an intense public debate bordering an uproar [3] originating from misuse of personal data in various forms identifying personal political and other views of social network users. Although not all the relevant details are currently known, the situation does convey the varied nature of personal and sensitive data that may include personal statements as well as pictures, videos and associated metadata transmitted through various digital channels.

During the past decades, systems have been developed based on thorough research into geographic data, their digital representation and operations possible with them. Today, all major vendors of common data storing technologies (i.e. relational databases), both proprietary and open-source include geographic extensions to their database engines. There are also specialized technologies [4] available developed to operate with geographic information whose popularity is also witnessed by the wide array of available APIs for various popular programming platforms. These geographic information systems (GIS) capture real-world phenomena like rivers, city borders and flight paths modelling them with geometric primitives like coordinates, lines, shapes or polygons and process data through operations like projections, intersections or rasterizations.

While identity information does not constitute a first-class citizen of the GIS paradigm, we argue that identity information has a functional dependence on some of these geographic phenomena and should be modelled taking advantage of existing GIS technologies to include this dependence. Furthermore, current identity management systems neglect to treat this geographic dependence causing an incomplete representation of real-world data in those systems leading to issues of inconsistency and fallibility. In this paper, we will demonstrate through several real-world example scenarios some issues that arise from neglecting this inherent geo-dependency in current systems and propose solutions that may be encoded using the existing GIS functionality for storing and processing data.

We wish to stress in advance that we are not proposing persons should bear different formal or public names in different countries or different territories even if some examples in this work may seem to imply this. However, we assert that there already exists a notable difference in representation of some identity attributes in various systems correlated with geographies of those systems and these differences and their origin should be addressed in a systematic manner.

The rest of this paper is organized as follows: in the next chapter, we describe the current model of identity as implicitly defined by current state-of-the-art technologies. In the Chapter 3 we list several real-world scenarios where this way of modelling identities fails to satisfy the necessary requirements as listed in Chapter 2. Then, in Chapter 4 we demonstrate a change of a commonly used attribute of identity and analyse the implications for storage and exchange of such data. In chapter 5 we provide a sample list of attributes from real-world systems representing various groups of attributes regarding their geographic dependency. We believe these groups to form a base of geo-dependency expansion varieties so that any other existing or yet to be included identity attribute will fit in one of the described groups. A short analysis on the impact of this proposed geo-expansion on basic IdM requirements is provided in chapter 6.A conclusion summarizing our work and perceived contribution is given in the last chapter.

2. Attribute-based identity

A lot of introductory, overview, classification and analysis work has been done in the field [5–13], so we will summarize the basic facts relevant to our work: the model of identity implicitly relied on by systems in current usage and the requirements put on those systems that our proposal has an impact on.

IdM systems can be classified according to different criteria. Some authors introduce classifications and taxonomies that illustrate differences originating from different sets of requirements that these systems were designed for (e.g. Ferdous and Poet [9]), while other authors classify different systems according to their architecture (e.g. Cao and Yang [13]). Since our work proposes an expansion of the model of identity these systems use, the essential element present in these classifications is that there is always an identity providing party (IdP) present. In some scenarios it communicates with a single another party within a single security domain, in others it is a part of a more complex multi-party multi-domain communication, but every scenario involves the IdP storing and exchanging identity data. Other parties include service providers requiring identity data (often called relying parties, RP), clients (with somewhat differing meaning across protocols) and identity bearers (usually end-users). The impact of our proposed expansion on these scenarios is outside of the scope of this work, as we concentrated on data that are being transmitted and stored, not on parties involved.

Regarding the identity modelling shared by all the most popular systems currently in the field (through standards or concrete implementations), we will summarize the current state of the art to a claim about the common foundational model of identity in currently used systems: all current systems are built on the attribute-based model. To describe an identity and attach necessary data to it, standards define attributes that represent characteristics of subjects carrying the identity. Depending on the standard, these are usually of a defined type and can be mandatory or optional, single- or multi-valued, they can have additional metadata attached to them (like time intervals of validity, "prime"/"preferred" traits or similar), they can have constrained domains of allowed values within the respective type's domain (through listing of acceptable values, or by having intervals defined for numeric values defined, or by having specially crafted grammars attached constraining the valid string values, or by having other predicates attached, etc.) or just implicit domains defined through their types (e.g. a 32 bit unsigned integer).

The reason for asserting this is that contrary to an intuitive assumption that attributes or observable characteristics are what constitutes an entity and no other representation or encoding of identity information seems possible, other models exist or can be contrived, e.g. hierarchical models or models based on fuzzy reasonings.

Regarding the requirements for IdM systems, many necessary or desirable characteristics have been identified in existing work [8–11] and these regard privacy concerns, trust relations, security and usability of systems, so these will be addressed when analysing the impact of our approach.

Two important principles related to most of these requirements are proportionality and subsidiarity, as formulated by Alpar [10]:

> Proportionality stipulates that the amount of personal data being collected is proportional to the goal for which it is being collected. Subsidiarity demands that the same goal cannot be achieved in a more privacy friendly way

This is why we consider that the impact any proposed IdM system has on these guiding principles mandates special deliberation.

3. Examples of traditional IdM systems' shortcomings

Here we list several real-world scenarios where the described models of identity fail to provide IdM-expected functionality. We consider them common knowledge so we will omit referencing any formalized studies and sources.

When ordering a package from an online retailer abroad based in a different language domain, the purchasing party can never be certain how their shipping address information will propagate through various systems involved in behind-the-scenes B2B communication. For instance, our domicile alphabet aligns with the English alphabet (a de-facto lingua franca in international retail) in all but for a few diacritic letters. It has arguably (according to authors' personal experience from contacts with other local residents interacting with foreign-based online systems) become a common tactic among the local population to make on-the-fly substitutions of similarly looking letters. For instance, the letter "ž" is quite similar to the letter z so it is relatively safe to make a substitution when entering an address containing this letter to an online form. This is commonly done because even if the application that the purchaser directly communicates with supports the required portion of Unicode symbols, one can still sometimes receive (or fail to receive) shipments addressed to streets that contain HTML tags or are similarly malformed – somewhere along the chain of systems bartering data encoding errors occur due incomplete support of localizations. The end result is that various systems are left with somewhat varying records pertaining to the same information about individuals. This, of course, is not constrained just to address information but to any attribute drawing values from a language-associated domain. And we can speculate such issues to be much more prominent when involving exchanging data between systems based in regions of greater language “distance”, i.e. when parties’ languages or their scripts differ more than in this example.

Many countries in the world enact formal policies that regulate the immigration of foreign nationals to their territory. For example, the U.S. Constitution explicitly empowers the U.S. Congress to establish rules of naturalization (which passed a number of immigration acts through its history e.g. [14–16]), the British nationality (a complex institute including currently six classes of nationality [17]) can be acquired in various ways (through birth in the UK territories, birth to British nationals in other territories, naturalization, adoption or registration [18]).

Many countries in the world provide some mechanisms of accepting foreign nationals interested in gaining local temporary or permanent residence (e.g. Also, most countries either explicitly define their official language or set of languages for official use in formal communication and public documents. Since people are usually named according to their local customs and localized names, the question arises about what names they can adopt and how these are represented in issued documents like national identity cards or driver’s licences. One can hardly expect every public servant that will rely on those documents for authentication or similar purposes to understand the myriad of scripts in use today. Aside from the written form, spoken languages have their rules and patterns, so it is not uncommon for Petar to become Peter among English speakers or even farther Ivan to become John (owing to traditional name conversions and name equivalences, in this case stemming from Bible texts and their translations). It is also a commonly observed phenomenon that expatriates returning to their original country of residence after taking on a new name for the aforementioned reasons be referred to by their old names back home, effectively forming a relation between a country and their name (i.e. “In the U.S.A. my name is John, but here everyone calls me Ivan, so I left that on my doorbell”).

Through historic tradition, languages have adopted geographic names into their linguistic corpuses transforming them through locally acceptable transliterations and pronunciations. Cities known to their inhabitants as Wien and 北京 are better known to English speakers as Vienna and Beijing and to Croatian speakers as Beč and Peking. These locality names form parts of various attributes bound to identities, like place of birth, residence, work experience, education, etc.

These examples mostly illustrate language-associated differences in representation of personal characteristics. However, since official languages as well as various official nomenclatures and ontologies that compose values of some attributes are often prescribed by states in charge of some territory, we can transitively conclude that the attribute values actually depend on geographic location. Additionally, many countries observe specificities localized to some regions of special status that impact identity data, e.g. provisioning the public and official use of minorities’ alphabets or languages in regions where a certain minority is especially present. We note that country borders change in some measure through time and so do legislative frameworks impacting some identity attributes, but these constitute temporal dependencies of identity data which are not the scope of this work.

Lately, there has been a notable shift in national professional and academic degree classifications in many countries in an effort to make them more compatible between various countries. The resulting systems are still not completely mapped to each other, and a significant fraction of the workforce still owns degrees from past systems. Bearers of degrees usually looking for work opportunities abroad are forced to attain locally recognizable comparable degrees through processes of nostrification or possible additional education and certification. The usual result is for a person to have different degrees recognized in different areas.

The common trait of these example scenarios is the existence of data for certain identities that are bound to different areas. IdM systems involved in these scenarios, whether in the form of proprietary services acting in B2B scenarios or the form of public identity-concerned legislature (that also constitutes an IdM system) fail to include these dependencies, resulting in misdemeanours of various severities: endangering the delivery of postal packages to the right address, forcing some measure of change of a person’s name or of a person’s city of birth name for reasons of encoding them in a locally usable way or not recognizing foreign degree classifications and forcing their bearers to translate them to locally acceptable on case-by-case basis. We conclude that a system acknowledging the geographic dependency of these attributes would help mitigate such problems.

4. Representing geographic dependency of an attribute

To demonstrate our proposed expansion of the identity model, we will illustrate it through expanding the definition of an attribute common in most systems used today, namely, a person’s address.

The differences in localized attribute values in our example may seem minor, but they illustrate the differences stemming from different phoneme sets across languages and may represent greater differences in locality-based representation of attributes. We are aware of situations where e.g. Cantonese names are transformed through transliteration and romanization to English-sounding names or their bearers take on different names completely (whether formal or informal, e.g. artistic) for better recognition among English-only speakers (e.g. celebrities like Mr John Woo or Mrs Ziyi Zhang). But we omit using them as we believe such examples would be better explored by native speakers of respective languages. Since we are neither sufficiently acquainted with these language traditions nor linguistic experts, we choose examples from our local experience.

For a concrete example, we will use an address in Zagreb that includes an existing street (the street number has been chosen too large on purpose, the full address is non-existent) – "Ulica Ivana Banjavčića" (Ivan Banjavčić's Street). There are several important features in this example. Letters "č" (pronounced similarly to ch in nacho) and "ć" (pronounced similar to t in nature) are specific to some languages. Serbian Cyrillic has symbols for both, but Russian Cyrillic only contains letter , an analog to "č" (Serbian Cyrillic has letter-for-letter equivalents to all Latin letters used both in Croatian and Serbian). Also, "nj" (pronounced similar to n in new in most English pronunciations) is a single letter with its Unicode code point (0 × 01CC for small, 0 × 01CA for capital), but in practice people write it as the two-letter combination "n" and "j" (owning to historic key layouts of keyboards and prior to that mechanical and electrical typewriters). Contrary to this, when writing in Cyrillic scripts that have an equivalent (e.g. Serbian Cyrillic script contains , analogous to nj, Russian does not), people use this single letter, avoiding any two-letter combinations. In addition, when referring to street names, it is customary to shorten them in a well-known way where "Ulica Ivana Banjačića" becomes "Banjačićeva" which is so thoroughly practiced that many official documents include these shortened names. For our example, we will assume that everyone in the city of Zagreb observes this habitual shortening of street names.

As previously mentioned, in a scenario where a person is entering this information into a web-form on an English-based system, our experience suggests there is an established tradition of substituting diacritic letters. This is not consistent with closest-sounding phonemes, as č is regularly substituted with c (in Croatian read like tz in waltz), but more closely follows letter similarity. In contrast to this, when translating some letters, like ć, to other (usually Slavic) languages that do not have this letter, but do have a č (or an analogous symbol), it is often transformed to this, closest-sounding phoneme and letter.

A number of standards in IdM is defined using XML [19], notably SAML [20] and WS-Federation [21]. XML syntax has limitations, but has proven to be of sufficient expressive power for defining protocols of data exchange, and also provides facilities like MTOM [22] and XOP [23] to (relatively) efficiently handle binary data making it acceptable for dealing with identity-associated multimedia data. The provided example modifies the attribute encoded in SAML 2.0 [24] syntax. We choose the attribute `streetAddress` as defined by X.500 and the SAML's X.500/LDAP [25] attribute profile. We will expand an attribute's value into (value, geodomain) pairs that list in which geographic domain which attribute's value is valid (we call them geodomains here to distinguish from traditional attribute's domain, i.e. the set of allowed values of the attribute). The example is based on the illustrative example from SAML X.500 LDAP attribute profile standard and uses commonly used XML namespace prefixes `xsi:` and `xsd:` for XML Schema Instance and XML Schema (defining types and type definition) namespaces as well the `x500:` and `saml:` prefixes to expectedly map SAML assertion definition and SAML's X.500 attribute profile:

```
saml: urn:oasis:names:tc:SAML:2.0:
assertion
x500: urn:oasis:names:tc:SAML:2.0:
profiles:attribute:X500
xsd: http://www.w3.org/2001/
XMLSchema
xsi: http://www.w3.org/2001/XMLSchema
-instance
```

With these, a SAML 2.0-conformant attribute description (most often part of an assertion in SAML) would be formed this way:

```
<saml:Attribute
    xmlns:x500="urn:oasis:names:tc:
    SAML:2.0:profiles:attribute:X500"
    NameFormat="urn:oasis:names:tc:
    SAML:2.0:attrname-format:uri"
    Name="urn:oid:2.5.4.9" Friendly
    Name="streetAddress"
    x500:Encoding="LDAP">
  <saml:AttributeValue xsi:type=
    "xsd:string">
      Ulica Ivana Banjavčića 100
  </saml:AttributeValue>
</saml:Attribute>
```

If SAML allowed for attributes to have localized values, it could define an XML element called e.g. LocalizedAttributeValue with an XML attribute called e.g. `poly` referencing a geographic region this value applies to. A fallback value can be introduced that is valid for all geographic regions not listed explicitly. So the attribute value:

```
<saml:AttributeValue xsi:type=
"xsd:string">
    Banjavčićeva 999
</saml:AttributeValue>
```

becomes:

```
< saml:LocalizedAttributeValue xsi:
type = “xsd:string” gtype:poly =
“45.82,15.82;45.75,15.88;45.75,
16.05;45.78,16.19 ;45.83,16.23;
45.96,16.15“ >
Banjavčićeva 999
< /saml:LocalizedAttributeValue >
< saml:LocalizedAttributeValue
xsi:type = “xsd:string” gtype:ref =
“glist:hr“ >
Ulica Ivana Banjavčića 999
< /saml:LocalizedAttributeValue >
< saml:LocalizedAttributeValue
xsi:type = “xsd:string” gtype:ref =
“glist:ru“ >
Улица Ивана Банявчица 999
< /saml:LocalizedAttributeValue >
< saml:LocalizedAttributeValue
xsi:type = “xsd:string” gtype:ref =
“glist:rs“ >
Улица Ивана Бањавчића 999
< /saml:LocalizedAttributeValue >
< saml:LocalizedAttributeValue
xsi:type = “xsd:string” gtype:ref =
“glist:us“ >
999 Ulica Ivana Banjavcica
< /saml:LocalizedAttributeValue >
< saml:LocalizedAttributeValue
xsi:type = “xsd:string” gtype:default
= true >
Ulica Ivana Banjavcica 999
< /saml:LocalizedAttributeValue >
```

Here we have assumed the existence of a Schema-type document in an XML namespace referenced by the `gtype:` prefix (contriving a concrete namespace seems irrelevant) that contains at least three types from the example. The type `poly` is of base type `string` (according to Schema conventions) and contains geographic coordinates defining a geographic polygon (in the example we draw a hexagon approximating the administrative borders of the city Zagreb in Croatia). This way an attribute values' geodmain can be defined in line with the attribute.

Since listing geodomains this way is cumbersome and expectedly repeating, we provide an alternative way of encoding domains through a list of predefined domains. These would be listed in the namespace referenced by namespace identified by prefix `glist:` (again, for purposes of illustration, contriving a concrete namespace is superfluous) binding territory names (like the used `hr/ru/rs/us`) to shapes expressed through coordinates of their vertices or through curves and other GIS-primitives. This document would probably offer shapes representing established administrative borders (i.e. country borders). To denote concrete countries' territories, we have used countries' respective assigned TLDs in the example.

Overlaps are allowed as illustrated through the example (Zagreb is within Croatia). When deciding what geodomain to use, a simple first-fit algorithm can be used (a somewhat intuitive alternative would be a smallest-shape-fit algorithm, but we do not see this having any significant merit). Therefore, the list of territories in `glist:` can also contain overlapping shapes to be chosen per-instance (e.g. if city borders are included in this list, the first LocalizedAttributeValue might have avoided listing coordinates, and instead used the formulation `gtype:ref = ”glist:zg”` if we adopt ISO3166 codes for the purpose or a similar formulation).

Since an attribute does have to have a valid value in all geographic locations, and providing a complete list of geodomains is unnecessarily cumbersome, we provide a default-style facility through the XML boolean attribute `gtype:default` that is false unless explicitly stated true for a single XML element.

5. Geographic dependency of identity attributes

We have analysed several popular standards dealing with identity data and their attributes according to geographic dependency of their values. The presented table (Table 1) illustrates various groups of attributes through an example. For each group, we identified we have provided a sample attribute from one of the analysed standards, and the standard the attribute is derived from is given through a reference.

The three basic groups of attributes regarding their values' dependency of geography are dependent, non-dependent and semantic-impacting. As illustrated by (Figure 1), we have expanded this set into a set of 8 groups in the table to illustrate the reasons why certain attributes are dependent or not dependent of geographic context in cases where the decision is not straightforward. The last basic group of attributes we call semantic-impacting. These show that identities can bear a different number of attributes in different regions. This is illustrated through a superfluous attribute that we see as an attempt to circumvent the inherent geo-dependency of attributes and provide a means of encoding it more or less within the traditional attribute-based model of identity as well as a combination of attributes that is impacted by geography. We believe that all the attributes from protocols we have analysed fit in one of the categories listed in the provided table.

6. Analysis

As illustrated through the example extension of an existing attribute in an existing protocol (the

Table 1. Classification of attributes according to their geo-dependency.

Attribute name	Group	Notes
email[26]	Non-dependent.	Example of an attribute that is not dependent of geodomain: an email address is absolute in this sense along other, usually "synthetic" data like IM contact info, universal IDs (e.g. STORK's eIdentifier), X.500's supported Algorithm and similar. This also applies to attributes with a well-defined domain in form of a value set where localizing it makes little sense (e.g. per OIC, the locale attribute is a language tag per RFC5646, and such tags should not be localized for simplicity reasons). However, not all synthetic data are absolute as illustrated by some of the following examples.
given_name[26]	Dependent	Representative of attributes dependent on geography for reasons explained in the example section – names act similarly to addresses in that they can be encoded differently and transformed in various measures when introduced to other languages.
maritalStatus[27]	Dependent for differing value-domains in different regions.	Per STORK, the domain of this attribute is "S (Single) / M (Married) / P (Separated) / D (Divorced) / W (Widowed)". Countries may differ in tracking marital statuses (if at all) of their citizens. Similar category is gender: STORK defines it as a M/F attribute, while OpenID allows the use of other values when "when neither of the defined values are applicable".
telephoneNumber[25]	Dependent for technical reasons	A system may choose to represent phone numbers along with appropriate prefixes (dial-out or dial-in prefixes) making a subject's phone number geo-dependent. In similar fashion, a digital certificate can be localized if a subject can provide digital certificates with attribute fields in various languages (e.g. if description-style attributes are translated or country names localized).
photography[26]	Dependent for cultural reasons.	For cultural reasons, subjects may choose to represent themselves through different photography according to the region. This example illustrates culturally based choices not necessitated, but enabled by a geo-dependent system.
dateOfBirth[27]	Non-dependent for technical reasons.	Representative of all date and timestamp-based attributes. Obviously different times are observed in different time zones at any moment, so there is a locality-bound dependence of such attributes. However, time-related data should preferably be stored in a geo-independent format (such as UTC-bound time structures) for simplicity reasons, and any necessary localizations for display or similar purposes should be performed as on-the-fly transformations in UI or other, "final" processing layers. Otherwise, the management of such data (including comparing and coordinating events) becomes too cumbersome.
honorificPrefix/honorificSuffix[28]	Dependent attribute combination	An example of an attribute combination that is dependent of the region, e.g. in English the academic title Ph.D. is usually written as a suffix, while in Croatian the analogous "dr.sc." is written as a prefix.
isAgeOver[27]	Superfluous	STORK 2.0 defines this as an attribute even though it is actually a relation (it requires a parameter and returns a Boolean value indicating if the identity-carrier is in relation with the given number or not). We assume this "attribute" was conceived because of the different legislations in the scope of the STORK project – because a person is allowed a certain right (like voting or purchasing an alcoholic beverage) at a different age in different countries. This approach can be remedied with geo-dependent attributes like mayVote or mayPurchaseAlcoholicBeverage that are simple attributes not requiring any parameters.

address in a SAML 2 assertion), managing geographic dependencies introduces additional complexity. Systems aiming to include this geographic dependency of attribute representations require more data to be processed and additional constraints to be observed (e.g. providing values for all relevant geodomains). This reflects negatively on both the simplicity of a system implementation as well as the efficiency of data processing. The reason for introducing this additional level of complexity is to model data in a more authentic, valid and complete manner. This is a necessary trade-off in any system that requires authentic representation of the real-world phenomenon of geographic dependency of identity attribute representations and meaning.

This higher consistency of data will enable systems to use it for additional tasks not supported by current systems like the ones illustrated in the examples section. The common trait of these is the possibility (or currently impossibility) of delivering data to interested parties in forms they are more accustomed to and can more easily understand. Importantly, this allows for treating the discrepancies already present in the real world (e.g. diacritic signs treatment in various online systems). Some existing real-world scenarios where locality dictates the interpretation of identity attributes are naturally addressed by our approach, as illustrated by our assumed origin of the age-related attribute/relation as explained earlier.

An obvious argument can be made that privacy is reduced by the introduction of systems binding possibly disparate personal data represented in differing forms. An individual might assume their privacy is somewhat heightened if they are using somewhat differing identities in different geographic domains (e.g. they believe their purchase history abroad will not be associated with their domestic purchase history because of differing name representations in domestic and foreign systems). However, this

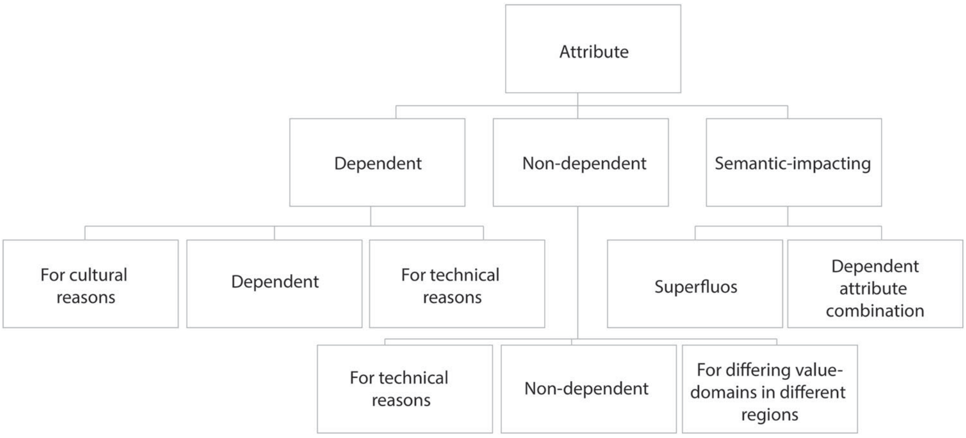

Figure 1. Base classification groups and derived groups.

assumption rests on an ill-advised principle related to security-through-obscurity that can not be relied on in any scenario as evidenced by capabilities of today's intricate data-correlating algorithms stemming from data-mining and other AI approaches. We surmise therefore that benefits provided by more accurate data representation in systems that rely on such data outweigh the privacy diminishment in this sense.

In a similar fashion, it might be assumed that proportionality and subsidiarity principles are undermined by our approach. However, for systems based on attribute-based models these stipulate that a system should not handle more attributes than necessary, not versions or representations of a single attribute: if a public healthcare system is breached and medical data endangered, the damage is even worse if the system additionally stored patient's bank account numbers, not because of storing localized formats of patient's charts.

Trust among parties is built over time. Even in scenarios with policy-based asserted trust, parties failing to deliver arranged functionality will lose trust confided by other communication participants relying on them. As has been noted [10], trust relationships are varied and not adequately addressed by existing systems in many scenarios, but we see the delivery of faulty or incomplete data as one of the factors for potential trust loss. Therefore, a system capable of handling more complete information, including geographic dependency of identity attributes should help avoiding this risk and enjoy higher trust. We see this as a consequence of the better real-world state representation enabled by a system taking geographic dependencies into account: even among people, we generally trust more those who approximate and abstract less and "tell it like it is".

We do not see the security of systems impacted by our proposed changes as these are mostly addressed at other levels of system design than data model itself. There may be some concern about the increase in transmitted and stored data volume and introduction of predictable patterns (e.g. we include additional references to shape sets in the illustrative example earlier) since some attacks scale with such target data, however, we assume that our proposal does not have a noticeable impact here and will consult empirical data on the matter once available.

Usability of IdM systems has been cited [29] as one of the cornerstones of such systems' adoption. This is a valid concern when introducing additional layers of complexity through geographic dependency of attributes as compared to simple name-value pairs. We believe that this can be significantly mitigated through careful design of implementing systems as long as only the necessary dependencies and values are introduced, e.g. no user wishes to control the localization of their attributes for every country or region possible – these should be reduced to the necessary set only. Also, users should be wherever possible assisted by automated means to handle added complexity. Nevertheless, usability is negatively impacted by our approach and part of the trade-off with the other mentioned concerns.

7. Conclusion and future work

In the introduction to this work, we have described some general trends that necessitate the use of organized identity management and more precisely some trends dictating new challenges to identity management in a cross-border context. We have introduced several simple scenarios that we believe illustrate the need to encode geographic dependency of identity representations. We have suggested a straightforward way of encoding this dependency through an example using one of the most popular protocols for exchange of identity data today – SAML 2.0. Our proposal to expand the current systems is inspired by GIS-related research and available technologies. Additionally, through examining several of the most popular standards representing various fields where identity management is present (OIC for web-SSO, SAML-referenced X.500/LDAP for enterprise IdM, SCIM for lately very researched field of cloud computing and STORK 2 representative of public efforts in the EU) we have classified concrete attributes in use today regarding their geo-dependency. We believe these to be a representative sample of identity attributes in use today. Three main groups were identified and described, as well as an additional layer of classification to help distinguish the classification of attributes. We concluded with an analysis of the expected impact of our proposed expansion on the basic requirements of IdM systems as they were identified in existing literature.

Our future work will explore the use of current GIS technologies for the storage of identity data since the data storage is a fundamental part of an IdM system. After that we intend to analyse in more detail the data exchange in the existing scenarios and provide necessary mechanisms for extending these into exchanges of geo-dependent data thus creating a full geographically sensitive system.

Disclosure statement

No potential conflict of interest was reported by the author.

References

[1] Azad MM, Ahmed AH, Alam A. Digital rights management. Int J Comput Sci Netw Security. 2010;10(11): 24–33.

[2] Dingledy F, Matamoros AB. What is digital rights management? Digital rights management: the librarian's guide. Lanham (MD): Rowman & Littlefield; 2016.

[3] Facebook scandal "hit 87 million users" [Internet]. London (UK): *The BBC*; [update 2018 April 4; cited 2018 April 5] Available from: http://www.bbc.com/news/technology-43649018

[4] ArcGIS [Internet]. Redlands (CA): Esri; [cited 2018 April 5]. Available from: https://www.arcgis.com/

[5] van Do T, Jørstad I. The ambiguity of identity. Identity Management. 2007;2007:3.

[6] Vapen A, Carlsson N, Mahanti A, et al. A look at the third-party identity management landscape. IEEE Internet Comput. 2016;20(2):18–25.

[7] Torres J, Nogueira M, Pujolle G. A survey on identity management for the future network. IEEE Commun Surv Tutorials. 2013;15(2):787–802.

[8] Birrell E, Schneider FB. Federated identity management systems: A privacy-based characterization. IEEE Secur Priv. 2013;11(5):36–48.

[9] Ferdous MS, Poet R. A comparative analysis of identity management systems. Proceedings of the 2012 International Conference on High Performance Computing and Simulation (HPCS); 2012 Jul 2–6; Madrid, Spain. p. 454–461.

[10] Alpar G, Hoepman JH, Siljee J. The identity crisis. Security, privacy and usability issues in identity management. J Inf Syst Secur. 2013;9(1):23–53.

[11] Dhamija R, Dusseault L. The seven flaws of identity management: usability and security challenges. IEEE Secur Priv. 2008;6(2):24–29.

[12] Beltran V. Characterization of web single sign-on protocols. IEEE Commun Mag. 2016;54(7):24–30.

[13] Cao Y, Yang L. A survey of identity management technology. Proceedings of 2010 IEEE international conference on Information Theory and Information Security (ICITIS); 2010 Dec 17–19; Beijing, China. p. 287–293.

[14] Bankston CL III. Immigration and Nationality Act of 1965; 2013.

[15] Stokes R. Immigration Reform and Control Act of 1986; 1987.

[16] Reid H. Comprehensive Immigration Reform Act of 2007.

[17] Types of British nationality [Internet]. London (UK): Government of UK. [cited 2018 June 9th]. Available from: https://www.gov.uk/types-of-british-nationality

[18] Become a British citizen [Internet]. London (UK): Government of UK. [cited 2018 June 9th]. Available from: https://www.gov.uk/becoming-a-british-citizen

[19] Bray T, Paoli J, Sperberg-McQueen CM, et al. Extensible markup language (XML). World Wide Web Journal. 1997;2(4):27–66.

[20] Hughes J, Maler E. Security assertion markup language (saml) v2. 0 technical overview. OASIS SSTC Working Draft sstc-saml-tech-overview-2.0-draft-08; 2005;29–38.

[21] Lockhart H, Andersen S, Bohren J, et al. Web services federation language (WS-Federation). Web services security specification; 2006.

[22] SOAP Message Transmission Optimization Mechanism [Internet]. Cambridge (MA): W3C; [cited 2018 July 3]. Available from: https://www.w3.org/TR/2005/REC-soap12-mtom-20050125/

[23] XML-binary Optimized Packaging [Internet]. Cambridge (MA): W3C; [cited 2018 July 3]. Available from: https://www.w3.org/TR/2005/REC-xop10-20050125/

[24] Cantor S, Kemp IJ, Philpott NR, et al. Assertions and protocols for the oasis security assertion markup language. OASIS Standard. March 2005;2005:1–86.

[25] Cantor S, Saml X. 500/LDAP attribute profile schema. OASIS SSTC; 2005.

[26] Sakimura N, Bradley J, Jones M, et al. OpenID Connect Core 1.0 incorporating errata set 1. The OpenID Foundation, specification.

[27] Ribeiro C, Leitold H, Esposito S, et al. STORK: a real, heterogeneous, large-scale eID management system. Int J Inf Secur. 2018;17(5):569–585.

[28] Hunt P LIK, Khasnabish B, Nadalin A, et al. System for Cross-domain Identity Management: Definitions, Overview, Concepts, and Requirements. 2015.

[29] Cameron K. THE LAWS OF IDENTITY, Kim Cameron's Identity Weblog; 2005.