

Paper PP22

Overlays data capture architecture (ODCA): Providing a standardized global solution for data capture and exchange

Paul Knowles, Dativa, Reading, UK

ABSTRACT

The post millennial generation has witnessed an explosion of captured data points which has sparked profound possibilities in both Artificial Intelligence (AI) and Internet of Things (IoT) solutions. This has spawned the collective realization that society's current technological infrastructure is simply not equipped to fully protect *personally identifiable information* (PII) or to entice corporations to break down internal data silos, streamline data harmonization processes and ultimately resolve worldwide data duplication and storage resource issues.

The FAIR Data Principles are a set of guiding principles in order to make data findable, accessible, interoperable and reusable (Wilkinson et al., 2016). These principles provide guidance for scientific data management and stewardship and are relevant to all stakeholders in the current digital ecosystem.

In line with the FAIR principles, data harmonization and interoperability processes between internal departments and functions is a high priority for companies but the current cognitive framework available for global data capture is hampered by limitations to the foundational data object architecture.

Overlays Data Capture Architecture (ODCA) facilitates the creation of a unified data language for content-based networks by introducing a set of properties to define the semantics of interoperable data objects within stable schema constructs. By introducing a dynamic component to static data structures, object interoperability and synergy across different systems can be achieved. The architecture provides the means to harmonise data within assurance frameworks, whilst driving the creation of a unified data language, thereby enabling the evolution of a decentralised data economy.

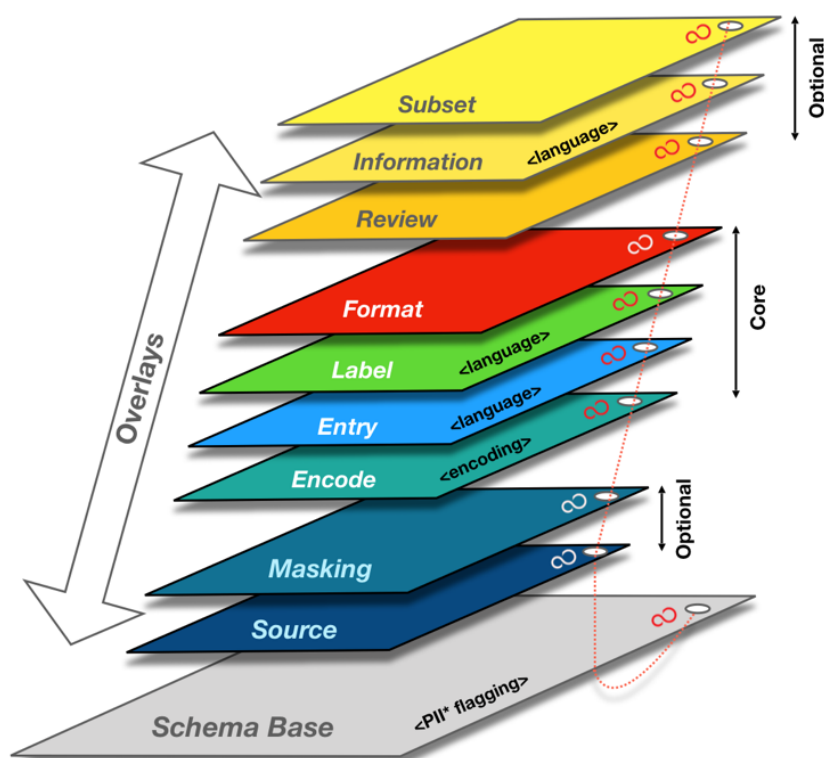


Fig.1 — Multiple overlays developed by different organizations to provide a set of metadata that adequately describes a single set of data.

OVERLAYS DATA CAPTURE ARCHITECTURE (ODCA)

ODCA is a standardized global solution for data capture and exchange which protects *personally identifiable information* (PII) data and provides a positive alternative to current architectures.

A schema, a machine-readable definition of the semantics of a data structure, is typically created as a single data object. However, ODCA represents a schema as a multi-dimensional object consisting of a stable schema base and linked overlays, data objects that provide additional extensions, coloration, and functionality to the base object. Any sponsor can use a pre-existing schema base and build their own suite of linked overlays to add extra context to transform how information is displayed to a viewer or to guide an agent in how to apply a custom process to schema data.

ODCA was primarily devised for data object interoperability and privacy compliant data sharing. The architecture promises to significantly enhance the ability to pool data more effectively in terms of simplicity, accuracy, and allocation of resources. The degree of separation between schema bases and overlays allows multiple parties to use the same base objects for similar data capture requirements thus providing a standard base from which to decentralize data.

ODCA offers many advantages, including:

1. *Data pooling.* Decoupling can occur at any time as overlays are linked objects. With all coloration stored in the overlays, combining data from related sources becomes much easier. Overlays can be removed from the base objects before the data merging process begins and reapplied to ensure consistent coloration post data pooling.
2. *Stable schema bases.* Most schema updates tend to be done at the application layer. In the case of ODCA, all extension, coloration, and functionality definitions are applied in the overlays. This enables issuers to edit one or more of the linked objects to create simple updates rather than having to reissue schema bases on an ongoing basis.
3. *PII encryption.* Using the Blinding Identity Taxonomy (BIT) as a reference [see section below], issuers can flag PII attributes in the schema base. With PII attributes flagged at the base object layer, all corresponding data can be treated as sensitive throughout the data lifecycle and encrypted or removed at any stage. This ensures that data protection of personal data is guaranteed as entity identification is impossible.
4. *Data decentralization.* ODCA enables schema base definitions to remain in their purest form thus providing a standard base from which to decentralize data. Organizations wishing to contribute data to a decentralized data lake for consented third-party usage can capture data using generic open source schema bases. This ensures that data standardization is done prior to any data lake migration.

BLINDING IDENTITY TAXONOMY (BIT)

Internet-related services and social media companies founded in the late '90s and early '00s triggered a digital hoarding revolution with large amounts of personal data captured and stored in corporate data silos under centralized control. Silicon Valley's largest technology companies have subsequently seen revenue figures soar through deployed marketing tools built to enable third party vendors to target subsets of individuals according to dynamic criteria searches. Revenue models have tended to rely heavily on these targeted advertising mechanisms which has encouraged an unparalleled corporate drive for data ownership.

Corporate responsibility regarding data privacy should have escalated in parallel with the relentless drive for data ownership but, until the General Data Protection Regulation (GDPR) came into force in May 2018, strong legislation was not in place to prevent the unethical distribution of personal data.

GDPR is a set of laws that provide a legal framework for the data protection and privacy of all individuals within the European Union (EU) and the European Economic Area (EEA) whilst also addressing the export of personal data outside the EU and EEA areas. It aims to empower individuals by improving their right to self-determination regarding their personal data. However well intentioned, GDPR does not sufficiently define a concrete list of PII elements which is problematic when it comes to tech implementation and personal data processing.

The Blinding Identity Taxonomy (BIT) was created to define a list of elements that could potentially unblind the identity of a person, an organization, or a thing.

- Names (incl. First Names, Last Names, Full Names, Entity Names)
- Physical Addresses
- E-mail Addresses
- Telephone Numbers
- Postal Codes
- Personal Software Application Handles (e.g. Skype, Slack, Hyperledger Chat, etc.)
- Profile Pages
- Passport Numbers
- Social Security Numbers
- National Insurance Numbers
- Driving License Numbers
- Vehicle Registration Numbers
- Bank Account Numbers
- Credit (or Debit) Card Numbers
- Personal Identification Numbers (PIN)
- Private Keys / Master Keys
- Symmetric Keys
- Public Keys
- Link Secrets
- Employee Identifiers
- Account Identifiers
- Governmental Identifiers
- Membership Identifiers (e.g. Trade Union Membership, etc.)
- Institutional Identifiers (e.g. Private Health Care Identifiers, etc.)
- Case Identifiers (e.g. Case ID Numbers, Benefit Plan Participation Identifiers, etc.)
- User Identifiers (e.g. User IDs, Logins, etc.)
- Passwords
- Signatures
- Digital Certificates
- Photos
- Videos
- Images
- Vocal Sound Bites
- Dates and timestamps (e.g. Date of Birth, transaction dates, etc.)*
- Genetic Identifiers (incl. chromosomal, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) data)
- Biometric Identifiers (incl. voiceprints, iris scans, facial imaging and dactyloscopic (fingerprint) data)
- Internet Protocol (IP) Addresses
- Media Access Control (MAC) Addresses
- Service Set Identifiers (SSID) (incl. local WiFi SSIDs)
- Bluetooth Device Addresses (BD_ADDR)
- Locational Information (incl. Global Positioning System (GPS), 3 word address, etc.)
- Cookie Browser Identifiers
- Radio Frequency Identifiers
- IoT Identifiers (incl. smart meter data)
- International Mobile Equipment Identity (IMEI)
- International Mobile Subscriber Identity (IMSI)
- Social media interactive elements, posts and comments (incl. likes, emojis and polling results)
- Free-Form Text Fields / Unstructured Data**

* Note: Not all captured dates will reveal identity but some will so, if in doubt, encrypt.

** Defn.: Text which does not have a given structure, nor which is entered in any specific format. Note: All free-form text fields should be encrypted.

Fig.2 — The Blinding Identity Taxonomy is a list of 48 PII elements that could potentially unblind the identity of a person, an organization, or a thing.

The BIT is one of those critical pieces of behind-the-scenes plumbing that is expected to fundamentally improve data protection of personal data as deployment rates in both traditional and distributed ledger technology (DLT) domains rise.

In terms of ODCA implementation, issuers can reference the BIT and flag PII attributes in the schema bases. With PII attributes flagged at the base object layer, all corresponding data can be treated as sensitive throughout the data lifecycle and encrypted or removed at any stage, making identifying individuals impossible and thus guaranteeing their privacy.

The BIT resides with Kantara Initiative, a non-profit industry consortium and professional trade association dedicated to advancing technical and legal innovation and trust framework operations related to digital identity management and data privacy.

The latest version of the BIT is available at:

<https://kantarainitiative.org/confluence/display/infosharing/Blinding+Identity+Taxonomy>

BACKGROUND

In conjunction with the exponential rise of data capture, Satoshi Nakamoto's groundbreaking white paper "Bitcoin: A Peer-to-Peer Electronic Cash System" was published in November 2008 triggering a peer-to-peer (P2P) computing revolution where files and transaction proofs could be shared directly between network nodes without the need of a central server. In 2009, Bitcoin became the first cryptocurrency to utilize a decentralized ledger to keep a record of all transactions taking place across a P2P network. It was not until the launch of Ethereum in July 2015, that the foundational technology, "blockchain", would feature smart contract functionality giving rise to a golden age of DLT solutions that continue to mould a decentralized data economy.

DLT solutions will continue to drive uniform data processing mechanisms, verifiable proof of consent, secure data portability and self-sovereign identity (SSI). With ODCA, a standardized global solution for data capture and exchange, community driven data standards, interoperable data capture objects and PII encryption capability can also be realized.



It is ultimately a combination of these ingredients that will enable an improved ontology-driven approach to data management allowing data to be decentralized and better AI and IoT solutions to be built for societal benefit.

ODCA DATA OBJECTS

Rather than a schema being created as a single data object, ODCA represents a schema as a multi-dimensional object consisting of a schema base and linked overlays. Each data object serves a specific function in the overall schema definition which, when amalgamated, provide a set of metadata that adequately describes a single set of data.

A schema base contains its own decentralized identifier (DID), a new type of identifier that is globally unique, resolvable with high availability, and cryptographically verifiable. In order for an overlay to be linked to a schema base, a *hashlink*, a cryptographic hyperlink referencing a specific endpoint, must be provided in the metadata block of the overlay. In other words, a linked overlay should always contain a hashlink referencing the schema base for coupling purposes.

SCHEMA BASE

A schema base is a stable base object that defines a single set of data in its purest form thus providing a standard base from which to decentralize data.

Apart from any metadata relating to the object, attribute names and types are defined in the schema base. The construct also facilitates a PII schema object which allows the issuer to flag PII attributes. With PII attributes flagged at the base layer, all corresponding data can be treated as sensitive throughout the data lifecycle and encrypted or removed at any stage thus protecting the identity of the data subject.

The DID of the schema base is contained in the metadata block and allows that base object to remain immutable and interoperable in both traditional and distributed ledger technology (DLT) environments.

```
{
  "@context": "https://odca.tech/v1",
  "name": "Demographics",
  "type": "spec/schema_base/1.0",
  "description": "A schema to capture statistical data about the characteristics
of a population, such as age, gender and race of the people within it",
  "classification": "GICS:35202010",
  "issued_by": "",
  "attributes": {
    "BRTHD": "Date",
    "AGEIC": "Number",
    "AGEU": "Text",
    "SEX": "Text",
    "ETHNIC": "Text",
    "INDALK": "Boolean",
    "ASIAN": "Boolean",
    "RACESP": "Text",
    "BLACK": "Boolean",
    "ISLAND": "Boolean",
    "WHITE": "Boolean",
    "RACEUNK": "Boolean"
  },
  "pii_attributes": [
    "BRTHD"
  ]
}
```

Fig.3 — Example of a schema base. Attribute names and types are defined in the schema base. The construct also facilitates a PII schema object for flagging PII attributes.

SOURCE OVERLAY

A source overlay is an optional linked object that can be used to specify an endpoint where dynamic data or linked attachments can be located. A hashlink referencing the associated schema base is contained in the metadata block of the source overlay, acting as a coupling point to link the two data objects.

MASKING OVERLAY

A masking overlay is an optional linked object that can be used to add data masks to schema attributes that have been flagged as PII in the schema base. A hashlink referencing the associated schema base is contained in the metadata block of the masking overlay, acting as a coupling point to link the two data objects.

ENCODE OVERLAY

An encode overlay is a core linked object that is used to define the deployed character set encoding (e.g. UTF-8, ISO-8859-1, Windows-1251, Base58Check, etc.) to be used in the schema definition. A hashlink referencing the associated schema base is contained in the metadata block of the encode overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/encode/1.0",
  "description": "Character set encoding for Demographics",
  "issued_by": "",
  "default_encoding": "utf-8",
  "attr_encoding": {
    "BRTHD": "utf-8",
    "AGEIC": "utf-8",
    "AGEU": "utf-8",
    "SEX": "utf-8",
    "ETHNIC": "utf-8",
    "INDALK": "utf-8",
    "ASIAN": "utf-8",
    "RACESP": "utf-8",
    "BLACK": "utf-8",
    "ISLAND": "utf-8",
    "WHITE": "utf-8",
    "RACEUNK": "utf-8"
  }
}
```

Fig.4.1 — Example of an encode overlay specifying *UTF-8* as the character set encoding to be deployed. UTF-8 is a variable width character encoding capable of encoding all 1,112,064 valid code points in Unicode using one to four 8-bit bytes.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/encode/1.0",
  "description": "Character set encoding for Demographics",
  "issued_by": "",
  "default_encoding": "utf-8",
  "attr_encoding": {
    "BRTHD": "Shift-JIS",
    "AGEIC": "Shift-JIS",
    "AGEU": "Shift-JIS",
    "SEX": "Shift-JIS",
    "ETHNIC": "Shift-JIS",
    "INDALK": "Shift-JIS",
    "ASIAN": "Shift-JIS",
    "RACESP": "Shift-JIS",
    "BLACK": "Shift-JIS",
    "ISLAND": "Shift-JIS",
    "WHITE": "Shift-JIS",
    "RACEUNK": "Shift-JIS"
  }
}
```

Fig.4.2 — Example of an encode overlay specifying *Shift-JIS* as the character set encoding to be deployed. Shift-JIS is a character encoding for the Japanese language.

ENTRY OVERLAY

An entry overlay is a core linked object that is used to add predefined field values in a specified language to schema attributes. A hashlink referencing the associated schema base is contained in the metadata block of the entry overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/entry/1.0",
  "description": "Field entries for Demographics",
  "issued_by": "",
  "role": "Clinician",
  "purpose": "Patient Response",
  "language": "us_EN",
  "attr_entries": {
    "AGEU": [
      "YEAR"
    ],
    "SEX": [
      "Male",
      "Female"
    ],
    "ETHNIC": [
      "Hispanic or Latino",
      "Not Hispanic or Latino",
      "Not Reported",
      "Unknown"
    ],
    "RACESP": [
      "Chinese",
      "Taiwanese",
      "Asian Indian",
      "Korean",
      "Malaysian",
      "Vietnamese",
      "Other Asian"
    ]
  }
}
```

Fig.5.1 — Example of an entry overlay specifying *us_EN* (Language=English, locale=United States) as the language of choice. Predefined field values are defined in an entry overlay.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/entry/1.0",
  "description": "Field entries for Demographics",
  "issued_by": "",
  "role": "Clinician",
  "purpose": "Patient Response",
  "language": "jp_JP",
  "attr_entries": {
    "AGEU": [
      "年"
    ],
    "SEX": [
      "男性",
      "女性"
    ],
    "ETHNIC": [
      "ヒスパニックまたはラテン系",
      "ヒスパニック系またはラテン系ではない",
      "報告なし",
      "未知の"
    ],
    "RACESP": [
      "中国語",
      "台湾人",
      "アジア人インド人",
      "韓国語",
      "マレーシア人",
      "ベトナム人",
      "その他のアジア人"
    ]
  }
}
```

Fig.5.2 — Example of an entry overlay specifying *jp_JP* (Language=Japanese, locale=Japan) as the language of choice. Predefined field values are defined in an entry overlay.

LABEL OVERLAY

A label overlay is a core linked object that is used to add labels in a specified language to schema attributes and categories. A hashlink referencing the associated schema base is contained in the metadata block of the label overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/label/1.0",
  "description": "Category and attribute labels for Demographics",
  "issued_by": "",
  "language": "en_US",
  "attr_labels": {
    "BRTHD": "Date of Birth",
    "AGEIC": "Age",
    "AGEU": "Age unit",
    "SEX": "Sex",
    "ETHNIC": "Ethnicity",
    "INDALK": "American Indian or Alaska Native",
    "ASIAN": "Asian",
    "RACESP": "If race is Asian, specify origin",
    "BLACK": "Black or African American",
    "ISLAND": "Native Hawaiian or Other Pacific Islander",
    "WHITE": "White",
    "RACEUNK": "Unknown"
  },
  "attr_categories": [
    "_category-1_"
  ],
  "category_labels": {
    "_category-1_": "Race (select all that apply):"
  }
}
```

Fig.6.1 — Example of a label overlay specifying *us_EN* (Language=English, locale=United States) as the language of choice. Labels for schema attributes and categories are defined in a label overlay.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/label/1.0",
  "description": "Category and attribute labels for Demographics",
  "issued_by": "",
  "language": "ja-JP",
  "attr_labels": {
    "BRTHD": "生年月日",
    "AGEIC": "年齢",
    "AGEU": "年齢単位",
    "SEX": "性別",
    "ETHNIC": "民族性",
    "INDALK": "アメリカインディアンまたはアラスカ原住民",
    "ASIAN": "アジア人",
    "RACESP": "人種がアジア人の場合、起源を指定する",
    "BLACK": "黒人またはアフリカ系アメリカ人",
    "ISLAND": "ハワイ先住民またはその他の太平洋諸島系住民",
    "WHITE": "白い",
    "RACEUNK": "未知の"
  },
  "attr_categories": [
    "_category-1_"
  ],
  "category_labels": {
    "_category-1_": "人種 (該当するものをすべて選択)："
  }
}
```

Fig.6.2 — Example of a label overlay specifying *jp_JP* (Language=Japanese, locale=Japan) as the language of choice. Labels for schema attributes and categories are defined in a label overlay.

FORMAT OVERLAY

A format overlay is a core linked object that is used to add formats and field lengths to schema attributes. A hashlink referencing the associated schema base is contained in the metadata block of the format overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/format/1.0",
  "description": "Attribute formats for Demographics",
  "issued_by": "",
  "attr_formats": {
    "BRTHD": "DD/MM/YYYY",
    "AGEIC": "4",
    "AGEU": "6",
    "SEX": "6",
    "ETHNIC": "22",
    "INDALK": "1",
    "ASIAN": "1",
    "RACESP": "30",
    "BLACK": "1",
    "ISLAND": "1",
    "WHITE": "1",
    "RACEUNK": "1"
  }
}
```

Fig.7 — Example of a format overlay. Formats and field lengths for schema attributes are defined in a format overlay.

REVIEW OVERLAY

A review overlay is an optional linked object that can be used to enable third-party comments to initial field entries (e.g. for auditors, reviewers, compliance officers, etc.). A hashlink referencing the associated schema base is contained in the metadata block of the review overlay, acting as a coupling point to link the two data objects.

INFORMATION OVERLAY

An information overlay is an optional linked object that can be used to add instructional, informational or legal prose in a specified language to assist the data entry process. A hashlink referencing the associated schema base is contained in the metadata block of the information overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/information/1.0",
  "description": "Informational items for Demographics",
  "issued_by": "",
  "role": "Clinician",
  "purpose": "Patient Response",
  "language": "en_US",
  "attr_information": {
    "AGEIC": "Fixed unit: Years"
  }
}
```

Fig.8.1 — Example of an information overlay specifying *us_EN* (Language=English, locale=United States) as the language of choice. Instructional, informational or legal prose to assist the data entry process are defined in an information overlay.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/information/1.0",
  "description": "Informational items for Demographics",
  "issued_by": "",
  "role": "Clinician",
  "purpose": "Patient Response",
  "language": "jp_JP",
  "attr_information": {
    "AGEIC": "固定単位: 年"
  }
}
```

Fig.8.2 — Example of an information overlay specifying *jp_JP* (Language=Japanese, locale=Japan) as the language of choice. Instructional, informational or legal prose to assist the data entry process are defined in an information overlay.

SUBSET OVERLAY

A subset overlay is an optional linked object that can be used to create a schema subset. A hashlink referencing the associated schema base is contained in the metadata block of the subset overlay, acting as a coupling point to link the two data objects.

```
{
  "@context": "https://odca.tech/overlays/v1",
  "schema_base": "hl:35WbXY8PCa42U2RJRhsYLL3Z7bE3xZzuXoJ72qbwyowo",
  "type": "spec/overlay/subset/1.0",
  "description": "Schema subset for Demographics",
  "issued_by": "",
  "attributes": [
    "BRTHD",
    "AGEIC",
    "AGEU",
    "SEX"
  ]
}
```

Fig.9 — Example of a subset overlay. A schema subset is defined in a subset overlay.

SENSITIVE OVERLAY (HOLDER ONLY)

In contrast to other overlay types which are assigned by an issuer, a sensitive overlay is an optional object assigned by the data holder that can be used to flag user-defined sensitive attributes. For example, *gender* is not defined as a PII element in its most common presentation of *male* or *female* as, in isolation, it cannot identify an individual. However, Thailand has 18 different gender identities that are recognized in the local lexicon and, as such, may be deemed as sensitive to a Thai citizen. In this case, a sensitive overlay could be coupled to a data vault on a personal device or a repository held by a trusted agent to flag the element.

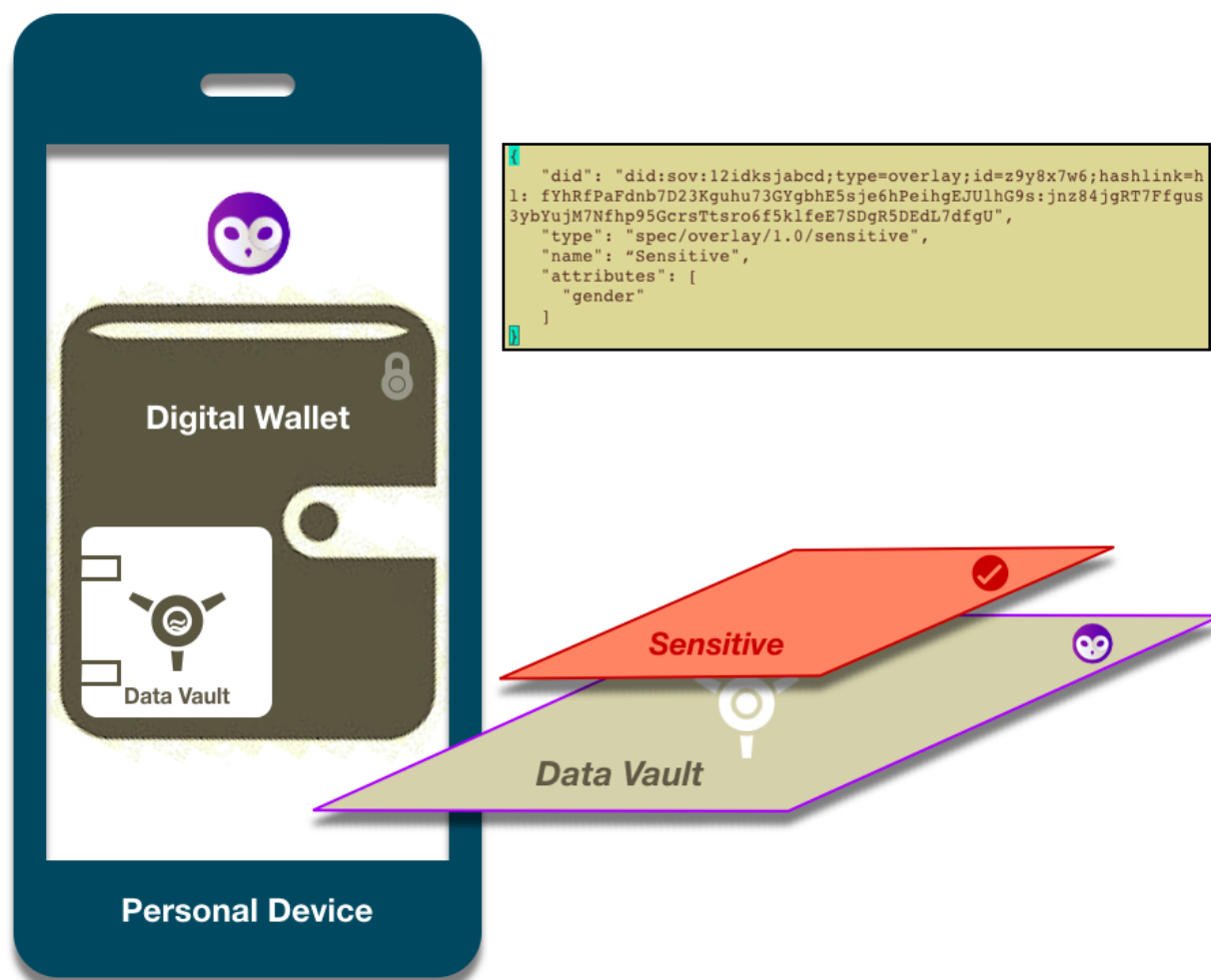


Fig.10 — Example of a sensitive overlay. User-defined sensitive attributes are defined in a sensitive overlay by the data holder. Rather than coupling the object to a schema base, a sensitive overlay can be coupled to a data vault on a personal device or a data repository held by a trusted agent.

CONCLUSION

ODCA significantly enhances the ability to pool data more effectively in terms of simplicity, accuracy, and allocation of resources. The degree of separation between schema bases and overlays allows multiple parties to use the same base objects for similar data capture requirements thus providing a standard base from which to decentralize data.

The architecture enables better data harmonization practices both internally within large organizations and externally under trusted assurance frameworks, seeding a unified data language for richer statistical analysis in the process.

ODCA aims to provide a standardized global solution for data capture and exchange which can support the development of pooled decentralized data lakes containing non-PII health data for societal benefit.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: Paul Knowles

Company: Dativa Limited

Address: Thames Tower, 30 Garrard St.

City / Postcode: Reading, RG1 1NR, UK

Email: paul.knowles@dativa.com

Web: <https://www.dativa.com>

Brand and product names are trademarks of their respective companies.