Decentralised semantics in distributed data ecosystems: Ensuring the structural, definitional, and contextual harmonisation and integrity of deterministic objects and objectual relationships*

Paul Knowles[1,*], Philippe Page[1] and Robert Mitwicki[1]

[1]Human Colossus Foundation, Geneva Switzerland

Abstract

We introduce the concept of decentralised semantics and how the segregation of task-oriented objects within a standard architecture can provide a long-term solution for unifying a data language within (or between) distributed data ecosystems. From that lens, decentralised semantics is ontology-agnostic, offering a harmonisation solution between data models and data representation formats while providing a roadmap to resolve privacy-compliant data sharing between servers, networks, and across sectoral or jurisdictional boundaries. Finally, as an illustration, the concept is applied to internationalisation and the dynamic presentation of transient objects.

Keywords

decentralised semantics, semantic interoperability, data harmonisation, objectual integrity

1. Introduction

The concept of decentralised semantics underpins pairwise peer-to-peer relationships where both parties must agree on which semantic definitions to use without referencing a specific standard. Due to the development of multi-stakeholder distributed data ecosystems, open discussions in two distinct areas: data harmonisation and trust infrastructures, have highlighted the need to advance the concept of decentralised semantics.

1.1. Data harmonisation requires context

Before data can be considered accurate[1], a harmonisation process is necessary. With the advent of decentralised semantics[2], defined in section 2, the harmonisation process can start at the earliest stage of the data management lifecycle at the point of data capture, where the contextual meaning of the data is fresh and more manageable than further into the data lifecycle.

Data without context is pointless. Data without definition is meaningless. Data without structure is unusable.

The Eighth Joint Ontology Workshops (JOWO'22), August 15-19, 2022, Jönköping University, Sweden

*Corresponding author.

paul.knowles@humancolossus.org (P. Knowles)

By most estimates, unstructured data, including everything from documents to images to video and audio streams to social media posts, account for at least 80-90% [3] of the overall digital data universe.

Graph technologies focus on analytics and querying (property graphs) and data integration (RDF graphs)[4] but often disregard the data harmonisation process, leaving no guarantee that the structural, definitional, and contextual integrity of the data is maintained. However, to ensure objectual integrity, data transformation must be controlled and issued by the services that capture data for a specific purpose (i.e., purpose-driven services) rather than those seeking existing data for analytics and insights (i.e., insights-driven services).

The primary objective of segregating task-oriented objects within a standard architecture is to facilitate data (structural), semantic (definitional) and pragmatic (contextual) harmonisation within any distributed data ecosystem, with the consensually-agreed data standards defined by ecosystem data governance.

1.2. Trust infrastructure for a data-agile economy

Decentralised semantics is one of the foundational concepts of a Dynamic Data Economy (DDE)[5], a next-generation data-agile economy offering a new paradigm in digital living, interaction, and growth, with a vision of empowering people and businesses to make better-informed decisions based on insights from accurate harmonised data framed by sound data governance. The DDE promises to facilitate new structures free from existing economic models while providing bridges to existing standards and legacy infrastructure.

The DDE is essentially a decentralised trust infrastructure acutely aligned with the European data strategy[6], where actors have the transactional sovereignty to share accurate information bilaterally. This trust infrastructure will ensure all digital objects' structural, definitional, and contextual integrity, the factual authenticity of any recorded event, and the consensual veracity of purpose-led policy-making.

The DDE conceptual infrastructure gives an equal weight of importance to three core data domains: ”decentralised semantics”, ”decentralised authentication”, and ”distributed data governance”. By differentiating explicitly between these three foundational domains, the definition of a network-agnostic information system, offering the cryptographic assurance of verifiable digital primitives and the human accountability facilitated by socio-economic data governance administrations and frameworks, is possible. Furthermore, decentralising semantics provides objectual integrity with secure bindings to advanced authentication mechanisms, which enable the construction of auditable data governance frameworks.

Decentralising semantics is crucial to unshackle the potential of data reuse through open opportunities while fully respecting jurisdictional rules and human values. Furthermore, pivotal for innovation in analytics, artificial intelligence, or other data-driven applications, the quest for accurate data is essential given the next wave of non-personal industrial data and the proliferation of Internet-of-Things (IoT) devices.

1.3. Plan of the paper

This paper focuses on the power of segregating task-oriented objects within a standard architecture to enable semantic interoperability, data harmonisation (structural, definitional, contextual), internationalisation, and dynamic presentation.

2. Decentralised Semantics

The two use cases introduced above shape the definition of decentralised semantics. The DDE trust framework provides a cryptographically secure ambient infrastructure to underpin distributed data ecosystems where stakeholders have the means to develop data governance administrations to enable bilateral peer-to-peer data exchange in a controlled environment. Both parties can agree on which semantic definitions to use without referencing a specific standard in a pairwise peer-to-peer relationship. Specific to semantic interoperability, the data governance administration defines the data capture requirements for their distributed data ecosystem, including the usage of any preferred ontological models. To achieve this in a simple manner accessible to end-users, an interoperable semantic architecture based on the segregation of task-oriented objects will allow the harmonisation of semantic definitions across different ontological models.

Data semantics[7], defined as the study of the meaning and use of data in any digital environment, underpins the definition of Decentralised Semantic necessary for distributed ecosystems.

Decentralised semantics describes a data modelling methodology of layering and cryptographically binding task-specific objects (overlays) to a standard capture base, which, when combined, defines a complex digital object. The segregation of task-specific overlays enables dynamic semantic interoperability in the construction process of any digital object without compromising the objectual integrity of the semantic structure, its modular components, or the relationship between those objects.

As a result, schemas based on decentralised semantics enable data harmonisation at the data capture level. Peer-to-peer connections access deterministic objects represented as a bundle of specific layers responsible for a given task. Cryptographic content-based identification of each separate overlay creates a deterministic structured bundle even when the overlays are from different jurisdictions.

Authorised data managers can define internationalisation and dynamic presentation in a decentralised but deterministic manner with semantic interoperability built into the underlying architecture.

3. Semantic Interoperability

One of the core characteristics and advantages of decentralised semantics is that different actors from different institutions, departments, sectors or jurisdictions can acquire control of specific task-oriented objects within the same semantic structure by binding an authorisation credential to that object. Semantic interoperability is an essential design characteristic within distributed data ecosystems, allowing multiple actors from different legal entities to participate in complex

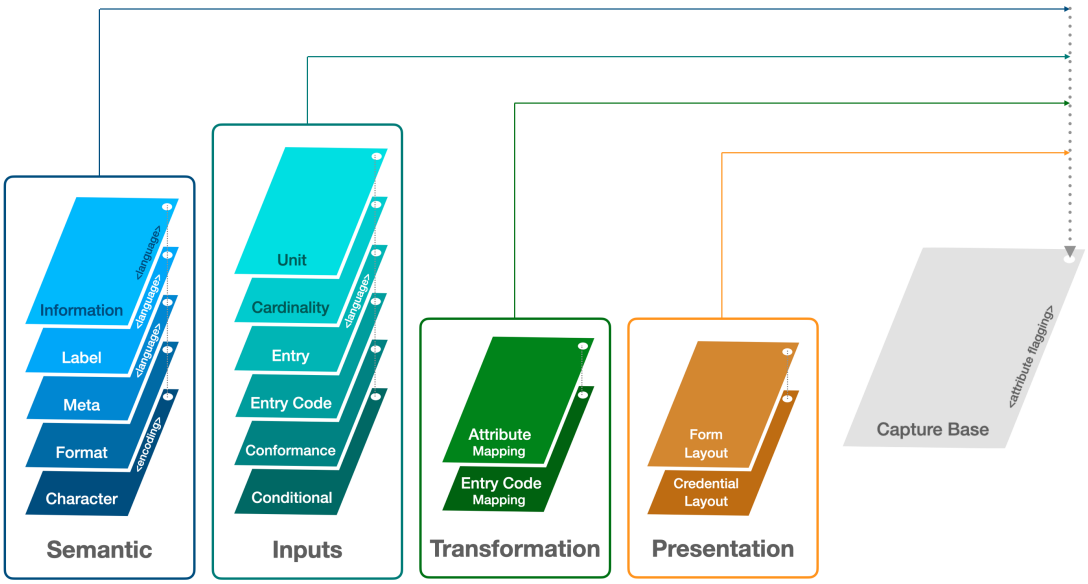

Figure 1: Semantic interoperability. Segregating task-specific objects (overlays) within a standard architecture enables different authorised controllers to update specific structural, definitional, or contextual components of the same semantic structure.

use cases, supply chains, and data flows supported by multi-stakeholder data governance administrations and frameworks. Figure 1 shows different overlay types categorised into the following groups:

- Semantic overlays provide contextual meaning to describe objects and their relationships, including attributes, forms, and schemas.
- Inputs overlays provide predefined inputs for data attestations, including claims, credentials, and records.
- Transformation overlays provide information to convert data from one format or structure to another, such as raw data to processed or unstructured to structured.
- Presentation overlays provide information to display data objects at the application layer, including forms, credentials, contracts, and receipts.

4. Data Harmonisation

Data transformation[8] is a crucial data management requirement for integration, migration, data warehousing, and data preparation, involving converting data from one format to another (e.g., a database file, XML document or Excel spreadsheet). These modifications typically involve converting a raw data source into a cleansed, validated, and ready-to-use format.

Separating task-specific overlays from the defined capture base offers a harmonisation solution between data models and data representation formats and from unstructured to structured

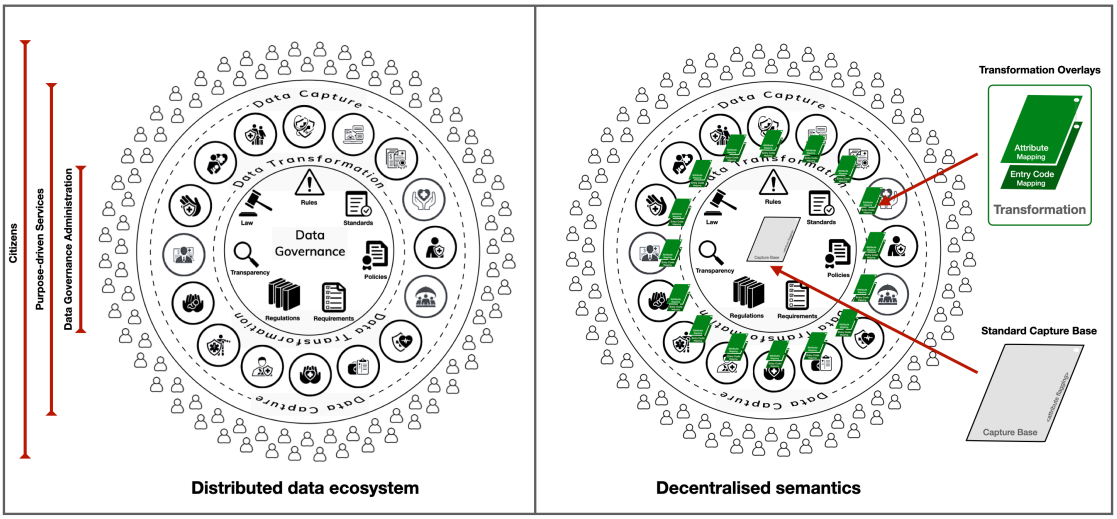

Figure 2: Data harmonisation. Decentralised semantics offer a harmonisation solution between data models and data representation formats and from unstructured to structured data.

data. Data harmonisation involves transforming datasets to fit together structurally while ensuring the definitional and contextual meaning of the source data is uniformly understood by all interacting actors, regardless of how it was collected initially.

The DDE infrastructure enables the development of trust frameworks for distributed data ecosystems, each with consensually-agreed rules, processes, and role delegations, acting as a data-sharing framework for collecting, processing, and maintaining electronic data. As depicted in Figure 2, distributed data ecosystems comprise citizens, service providers, and data governance administrations.

A data governance administration (DGA) is a multi-stakeholder collaborative or consortium working for a common purpose, industry sector, or jurisdiction, assuming responsibility for the consensual veracity of data transactions under its administrative control on behalf of the citizens and legal entities it serves. DGAs sit at the core of any distributed data ecosystem, aligning closely to the role of a "data intermediary" as described in the European Commission's recently proposed Data Governance Act[9], serving as a mediator between those who wish to make their data available and those who seek to leverage that data.

By issuing and controlling a set of proprietary transformation overlays, purpose-driven service providers can securely map source attribute names, entry codes, or unit conversions to a standard capture base defined by an ecosystem DGA. Capture bases provide a target for data harmonisation within distributed data ecosystems. Specifically, a cryptographic thread is established from the transformation overlays to a consensually-defined capture base, ensuring the integrity of those objectual relationships and facilitating a secure means for data harmonisation.

4.1. Application #1: Internationalisation

Internationalisation is the action or process of making something international[10]. The internationalisation of transient digital objects across distributed data ecosystems is essential for

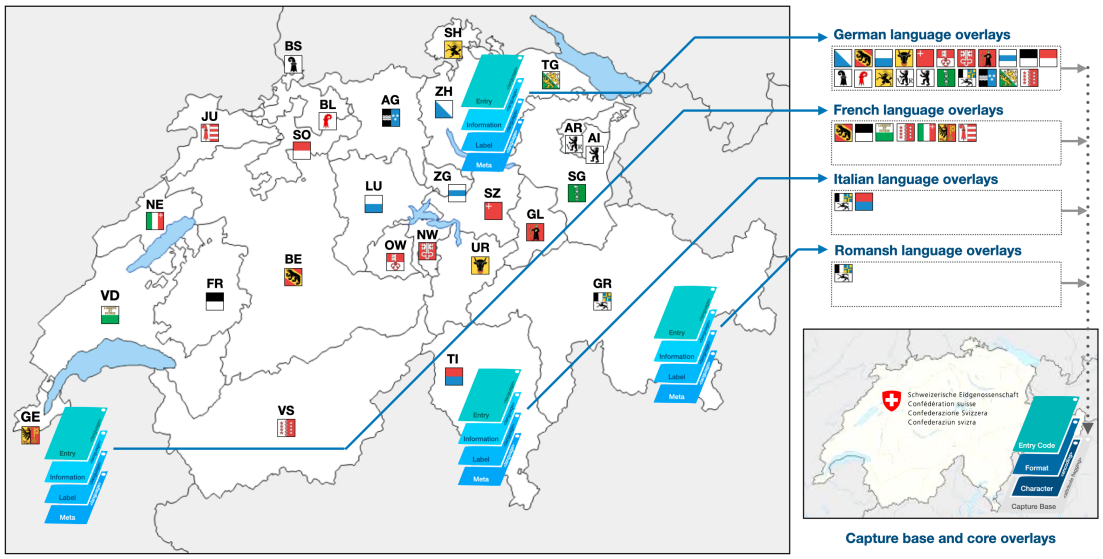

Figure 3: Internationalisation. Switzerland is a quadrilingual country. The decentralised control of language overlays would enable cantonal authorities to translate official documents issued by the Swiss national federal government into their region's official language(s).

service providers to participate in a global market. Traditionally, presenting information for a purpose-driven activity in a language understandable to all recipients has involved replicating digital forms, credentials, contracts, and receipts into various languages based on user preferences. With federated or centralised governance authorities maintaining digital objects in multiple languages, internal data management inefficiencies are common to many organisations, institutions, and governments.

The FAIR (Findable, Accessible, Interoperable, and Reusable) data principles[11] support the reusability of digital assets. Still, many legal entities have difficulty streamlining data management practices and processes to comply with these guiding principles.

Decentralised semantics offer a solution for the internationalisation of digital objects within distributed data ecosystems by enabling various authorised entities to control a different set of language overlays for a particular transient object, such as a digital form, with a DGA defining and issuing a standard capture base and core language-agnostic overlays. Let us take Switzerland as an example of a multilingual country from that lens.

Switzerland is officially quadrilingual, with German, French, Italian, and Romansh as its national languages. However, many other minority languages, such as English, are becoming increasingly important. Fiora et al. showed the importance of the principle of territoriality in their analysis[12]. Decentralised semantics is well suited to represent this principle in digital information systems. Since Switzerland is a federation, the sovereign cantons define their official language according to the primary language spoken by their inhabitants. Figure 3 presents Swiss cantonal borders, the segregation of their primary spoken languages, and how decentralised semantics provide a dynamic solution to internationalisation in a federal environment.

With cantonal participation being an essential ingredient of Swiss-style federalism[13],

separating language overlays from any capture bases and core language-agnostic overlays issued by the federal government would enable a collaborative solution to internationalisation. In this scenario, decentralised semantics allow sets of language overlays to be controlled and maintained by different cantons depending on their primary spoken language. In other words, decentralised control of language overlays would enable regional authorities to manage the official translation of any document issued by a national federal government into their region's official language(s).

The above example is globally scalable, with decentralised semantics enabling the translation of any digital object under established governance while preserving its objectual integrity. More importantly, it significantly impacts objectual inclusiveness within digital systems. Within an ecosystem, decentralised semantics allow for transient object design in a particular language, where additional interoperable language-specific overlays, including those for minority or indigenous languages, can be added dynamically.

4.2. Application #2: Dynamic Presentation

Another characteristic of decentralised semantics is the ability to cryptographically bind presentation overlays to the standard capture base used for authentic data entry. However, in many presentation instances, the legal entity that issues the original capture form may differ from the entity that issues the presentation objects required to produce an associated credential. For example, national passport issuance provides an opportunistic use case to demonstrate the advantages of this particular characteristic.

The International Civil Aviation Organization (ICAO) is a specialised agency of the United Nations tasked with planning and developing standards for safe international air transport[14]. ICAO's primary role is to provide a set of standards that will help regulate aviation worldwide. One of those standards is ICAO Document 9303[15] (endorsed by the International Organization for Standardization (ISO)[16] and the International Electrotechnical Commission (IEC)[17] as ISO/IEC 7501-1[18]), a global standard for machine-readable travel documents (MRTD), including the data capture requirements of a machine-readable passport (MRP). As a result, ICAO is well-positioned to be the primary issuer of a standard capture base and the core overlays required for MRP form inputs, semantics, and presentation.

However, the issuance of any presentation objects needed to produce a national passport with branded design requirements would be under the remit of issuing governmental agencies, including cantonal passport offices in the country and at its embassies or consulates overseas. As an example for Switzerland, the Swiss Government[19] is the authority to act as the primary issuer of presentation overlays to produce a branded Swiss passport, a credential identifying a traveller as a Swiss citizen or national with a right to protection while abroad, and a right to return to Switzerland.

The capture base and overlays are identifiable by Self-Addressing Identifiers (SAID)[20], a particular type of content-addressable identifier based on an encoded cryptographic digest that is self-referential. These identifiers are deterministic. In other words, there is no randomness in the identifier generation process, ensuring the objectual integrity of the digital objects and their relationships.

In this particular use case, authorised Swiss governmental agencies would inevitably store an

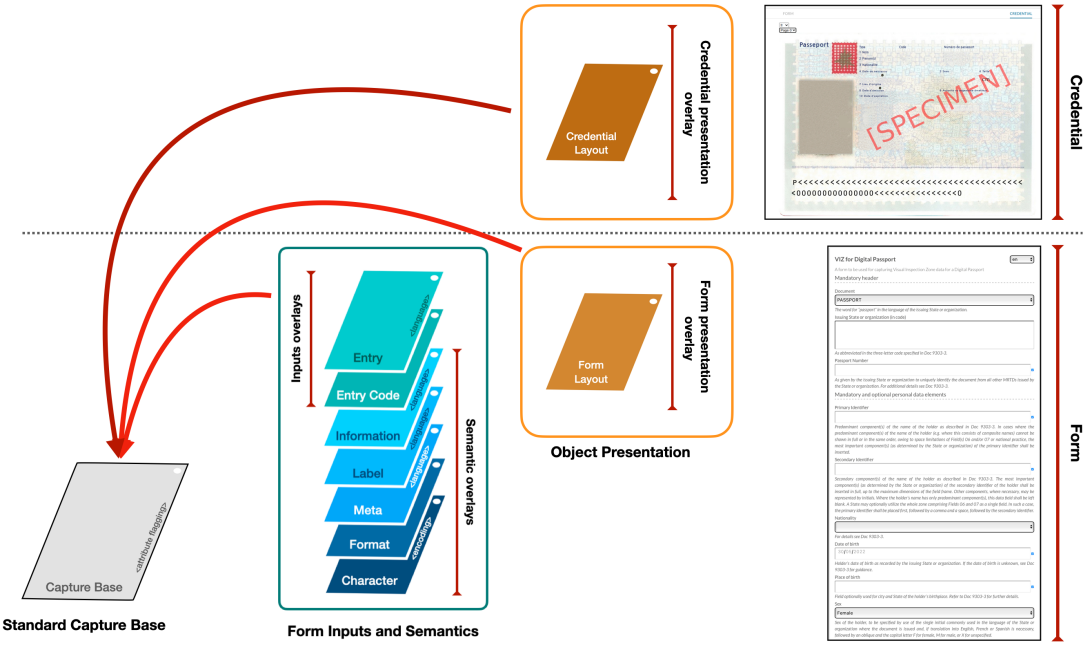

Figure 4: Dynamic presentation. The decentralised control of presentation overlays within a governed ecosystem enables the autonomous rendering of different transient objects cryptographically bound to the same capture base.

instance of the ICAO-issued MRP objects in local repositories. However, the SAIDs of those digital objects would remain unchanged from the original identifiers held in an ICAO repository.

As the object identifiers are deterministic, the dynamic presentation of national passports, in this case, can be established securely by maintaining a cryptographic thread from the presentation overlays to a standard capture base for global standardisation. Note that a national passport is an example of a credential presentation[21]. However, for different use cases, the presentation of other transient object types, such as digital forms, contracts, and receipts, would also benefit from the dynamic issuance of presentation overlays.

5. Conclusions

The Human Colossus Foundation[22] continues to develop open-source components that leverage decentralised semantics through sectoral projects in bio-science research, healthcare, and government, providing the means to harmonise data across each data lifecycle stage. Deterministic semantic modelling ensures the objectual integrity of digital objects and their relationships in every data management process, including data capture, data entry, data transformation, and data presentation, to ensure the same results upon running the model under the same initial conditions. Within DDE-compliant distributed data ecosystems, multiple actors from various institutions participate in complex use cases, supply chains, and data flows supported by multi-stakeholder data governance administrations and frameworks. Decentralised semantics

offers an enhanced data harmonisation solution (structural, definitional, contextual) while ensuring objectual integrity throughout any data lifecycle.

Acknowledgments

This work has been partially funded by: (i.) The Human Colossus Foundation "Data Harmonisation" programme, and (ii.) EU Horizon 2020 NGI grant number 871932 (essif-02). The authors would like to thank the team of developers at Argonauths. Special thanks to Michal Pietrus and Marcin Olichwiruk for their contributions and insights on decentralised semantics.

References

References are in order of appearance in the body of the text.

[1] F. Kim, What is Data Accuracy, Why it Matters and How Companies Can Ensure They Have Accurate Data, 2020. https://dataladder.com/what-is-data-accuracy/

[2] P. Knowles, Overlays Capture Architecture, 2020. https://humancolossus.foundation/blog/cjzegoi58xgpfzwxyrqlroy48dihwz

[3] O. Djuraskovic, Big Data Statistics 2022: How Much Data is in The World?, 2022. https://firstsiteguide.com/big-data-stats/

[4] Oracle, 17 Use Cases for Graph Databases and Graph Analytics, pp. 5. https://www.oracle.com/a/ocom/docs/graph-database-use-cases-ebook.pdf

[5] R. Mitwicki, Genesis of a Dynamic Data Economy, 2020. https://humancolossus.foundation/blog/dde-first-contact

[6] European Commission, European Data Strategy, 2020. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en

[7] J. Reed, What is Data Semantics?, 2022. https://www.languagehumanities.org/what-is-data-semantics.htm

[8] M.K. Pratt, C. Bernstein, Data transformation. https://www.techtarget.com/searchdatamanagement/definition/data-transformation

[9] European Commission, Proposal for a Regulation of The European Parliament and of The Council on European data governance (Data Governance Act), 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN

[10] Oxford Learner's Dictionaries, Internationalisation. https://www.oxfordlearnersdictionaries.com/definition/english/internationalization

[11] M.D. Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship, 2016. https://www.nature.com/articles/sdata201618

[12] M. Fiora, K. Schwarck, Quadrilingual Switzerland – a model for multilingual education?, 2021. https://sepia2.unil.ch/wp/multilingualism-in-society/2021/01/26/quadrilingual-switzerland-a-model-for-multilingual-education/

[13] Mission of Switzerland to the European Union, The cantons play an active role in the Confederation's policy on Europe. https://www.eda.admin.ch/missions/mission-eu-brussels/en/home/key-issues/cantons-role.html

[14] Wikipedia, International Civil Aviation Organization. https://en.wikipedia.org/wiki/International_Civil_Aviation_Organization

[15] ICAO - International Civil Aviation Organization, Doc 9303, Machine Readable Travel Documents, Eighth Edition, 2021. https://www.icao.int/publications/documents/9303_p1_cons_en.pdf

[16] ISO - International Organization for Standardization, https://www.iso.org/home.html

[17] IEC - International Electrotechnical Commission, https://www.iec.ch/homepage

[18] ISO/IEC 7501-1, Identification cards - Machine readable travel documents - Part 1: Machine readable passport, 2008. https://www.iso.org/standard/45562.html

[19] The Federal Council, The portal of the Swiss government. https://www.admin.ch/gov/en/start.html

[20] S. Smith, Self-Addressing IDentifier. IETF - Internet Engineering Task Force, 2022. https://www.ietf.org/id/draft-ssmith-said-02.html

[21] W3C, Verifiable Credentials Data Model v1.1, Presentation, 2022. https://www.w3.org/TR/vc-data-model/#dfn-verifiable-presentations

[22] The Human Colossus Foundation, https://humancolossus.foundation