Humanitarian data models for deduplication in cash coordination - Internal briefing note

August 2024

Background
The issue about deduplication in humanitarian cash coordination
What is deduplication actually about?
    Variations of deduplication
    Deduplication and adjudication processes
    Data standards: Core set of data semantics and data syntactic
    Governance frameworks
What drives deduplication? What hinders deduplication?
    Drivers
    Challenges and bottlenecks
What are potential ways for uptake of data standards for deduplication?
In short: Summary and follow-up
Bibliography

Background

During the last few years, the Dignified Identities in Cash Assistance (DIGID) Consortium[1] and the Collaborative Cash Delivery (CCD) Network[2] commissioned several analyses which identified a need for greater data sharing and interoperability through commonly agreed data models for deduplication in cash coordination. Around the same time, the IASC's Data Responsibility Working Group (DRWG) formed the Cash Task Team (CTT)[3] to identify key issues that cash practitioners face in conducting data management activities, as well as initiatives (guidance and support) developed in response to these issues. Through a practitioner's survey, undertaken by the CALP Network and CTT working sessions, the group also identified deduplication and adjudication issues that are consistently raised by cash implementers.

The three entities (DIGID, CCD, DRWG/ CTT), thus, agreed to facilitate a sector wide discussion with key stakeholders to gather learnings about

1) commonly agreed data models for deduplication of aid recipients, cash assistance packages, referrals, and
2) commonly used business rules for deduplication and adjudication processes.[4]

Reports, opinion pieces and relevant documents were reviewed and 20 key informant interviews (KIIs) with four humanitarian stakeholder groups of UN Agencies, technology/ service providers, NGOs and humanitarian coordination fora were conducted between June and August 2024. In addition, two CCT meetings were held in June and September 2024[5]. The information and opinions gathered throughout this consultation process are summarized in this internal briefing note.

The issue about deduplication in humanitarian cash coordination

DIGID analyses and numerous other reports including CALP's State of the World's Cash Report 2023[6] reveal a need for data sharing and interoperability to **reduce the duplication of cash recipients** and **facilitate referrals** between CVA programmes, organizations and social protection schemes.[7] DIGID's landscape mapping referenced deduplication as one of the most relevant use cases for cash practitioners. However, the mapping of CVA data sharing and interoperability use cases also revealed little to no publicly available evidence that backs-up the importance of deduplication, including the level of duplicated assistance caused by recipient error or fraudulent practices. Estimates refer to on average 5% level of

[1] The DIGID Consortium is a group of Red Cross Members and INGOs comprising the International Federation of the Red Cross (IFRC), Norwegian Refugee Council (NRC), Norwegian Red Cross and Norwegian Save the Children.
[2] The CCD Network represents 14 INGOs including World Vision, Oxfam, Mercy Corps and others.
[3] The CTT was formed at the request of the IASC Cash Advisory Group (CAG) and comprises representatives from across the humanitarian cash assistance space with an interest and experience in data management, from a range of geographies and types of organizations.
[4] See Concept Note: Data standards and business rules for adjudication of duplication
[5] The feedback of the CTT meeting scheduled in September 2024 is not reflected in this document.
[6] The importance of improved deduplication practices by identifying targeting challenges was identified as the second largest risk associated with CVA (43.5%). Inflation and/or currency depreciation (48.5%) was identified as the main risk, followed by fraud/corruption/diversion of assistance (41.6%) and protection for recipients (41.5%) (CALP 2023: 116).
[7] see DIGID reports on interoperability related matters, CALP 2022, Kreidler et al. 2022, DCF N.d.

duplicates while anecdotal evidence and rumors talk about 40-70% in humanitarian operations like Yemen and Somalia.[8]

DIGID's report "Investigating safe data sharing and systems interoperability in humanitarian cash assistance" further highlighted the need for data standardization.[9] Data standards, understood as **data semantics**[10] and **data syntactic**[11], play a foundational role in deduplication and, more broadly, interoperability. There are few humanitarian data standards that currently inform different purposes and processes of deduplication. Most standards, however, focus on non-personal data for facilitating the coordination of humanitarian assistance (e.g., 4Ws, HXL, p-codes, etc.).

The report also talks about the challenges to agree on the **governance framework** for managing relevant data standards: *"Defining the actual fields and descriptions is likely straightforward. Much more difficult will be determining the optimal process for stakeholder engagement and decision-making, identifying the appropriate institutional home for the schema, and developing the right incentive structures to spur adoption"* (Pon, B. et al. 2023: 32). There is a common understanding that data standards build the foundation for data sharing and interoperability in general, deduplication in particular. They, however, need to be defined along relevant business rules, including data semantics for data collection, data quality and data accuracy requirements, data syntactic for identifying potential duplicates, process flows, roles and responsibilities for handling potential duplicates and adjudication, as well as legal frameworks to inform data protection, data security and privacy aspects.

What is deduplication actually about?

In humanitarian cash assistance and cash coordination, deduplication is usually referred to the process of comparing lists of intended cash recipients to identify and/or eliminate any intentional and unintentional duplicates.[12] Commonly agreed assistance packages as well as predefined data points of personal data are used to flag the number of total duplicates including potential duplicates that must be adjudicated. Adjudication, herein, is described as the manual or automated process that allows for the administration of potential duplicates; in other words, the decisions taken once an individual has been identified in more than one humanitarian interventions' dataset. Questionable matching results, the potential duplicates, are flagged and, in most contexts, forwarded to human experts for manual adjudication.

The overall objective of deduplication varies depending on the context, its purpose, use case or variation as well as stakeholder group. Deduplication usually aims at

[8] Worthington, R. et al. 2023; KIIs.
[9] Pon, B. et al. 2023.
[10] Data semantics are standards that help applications establish a common vocabulary with similar enough definitions. It *"involves the development of vocabularies, code lists, and models to describe data elements and data sharing processes to ensure a common understanding between or among parties"* (Pon, B. et al. 2023: 11). Examples include dates in the same date formats, indicators that are measured the same way, groups that are aggregated in the same groupings (see WHO. N.d.).
[11] Data syntactics are standards that help applications establish a common grammar with similar enough data formats. It *"involves the description of the exact format of data that's being shared by parties in terms of grammar and format"* (Pon, B. et al. 2023: 11). This includes spreadsheets using XLS, CSV, or Sheets; images using PNG, JPEG, or GIF (see WHO. N.d.).
[12] Adjusted from Pon, B. et al. 2023: 15.

- Improving transparency and providing greater accountability to affected people, partners, host governments, donors and other decision-makers by ensuring the right person gets the right benefit.
- Contributing to more effectiveness and greater impact of humanitarian assistance.
- Supporting cost efficiency by avoiding “double dipping”.
- Reducing required data storage space and increased data processing speed.[13]

The purpose of deduplication is generally defined by programme staff who design and manage CVA programmes and operations. Deduplication is, however, no default option to implement in every programme to be implemented in every programme as deduplication can be resource and time intensive. There might be contexts where duplicate recipients are less crucial or less likely, for example, in case of

- low levels of duplication[14]
- individual family or household members enrolled in different programmes
- one-off cash distributions or short/er duration
- geographic targeting
- community-based targeting.

→ It is thus recommended that data standards for deduplication are defined in accordance with an objective, purpose and context.

Variations of deduplication

In the DIGID Landscape Mapping four variations or purposes of deduplication were identified:

1. **Deduplication with common unique identifier codes**
   Deduplication is based on a common identifier code for all people. The identifier might refer to a foundational ID (e.g., national ID number, tax number) or a functional ID (e.g., ID provided by UNHCR, UID generated through BB). Bank account number or mobile phone number could also be used as a unique identifier, provided that they are not shared with other family or household members.

2. **Deduplication without common identifier codes**
   In contexts where people have no functional or foundational ID, different types of ID documents or proxies are used to verify their identity. A form of validation process is required which is usually supported by local actors including implementing partners, community members, authorities and others.

3. **Deduplication of family or household data**
   Many humanitarian organizations follow a household approach, i.e., data is associated with a family or a household generally identified with the head of household and an alternate. The data of individual family or household members is usually not registered.

[13] This aspect was not mentioned by any respondent. The impact might be considered minimal and less important for humanitarian stakeholders, mainly due to low(er) data volume and storage needs in comparison to other sectors. It is, however, an integral part of similar discussions in other sectors (e.g., health).
[14] The example of IRC in MENA was referenced. IRC used an algorithm to anticipate the level of duplications before investing in setting-up costly deduplication mechanisms. The approach was considered sufficient by the donor. [Note: The example was not re-confirmed by IRC.]

4. **Deduplication with biometrics**
   Fingerprints, face, iris or voice might be collected to verify a person's identity. If an organization uses biometrics or not depends on the programme design, organizational policies as well as organizational capacities and capabilities to manage biometrics.

During this assignment, respondents raised the need to further specify or refine the purpose or use cases of deduplication. Above variations were thus differentiated into four use cases. The terms used below reflect the wording repeatedly used by respondents.

A. **Deduplication of cash assistance packages**
   To simplify deduplication of cash recipients the harmonization of cash assistance packages is crucial. This includes a coordinated approach to targeting (e.g., household setup, eligibility/ selection criteria, sector-/ multi-purpose cash), the modality (transfer values, frequency and duration, delivery mechanisms/ platforms) as well as programme cycles (e.g., time from registration and deduplication to the actual transfer of cash). The more parameters are harmonised or coordinated, the easier deduplication of cash recipients becomes. Each parameter is, however, very context- and/or organisation-specific.[15]

B. **Deduplication of registration or identities**
   This process aims at identifying whether an individual or household has been registered more than once by different organizations. Deduplication might happen at the point of registration or post-registration and pre-distribution.

C. **Deduplication of cash recipients**
   This variation was also referred to as 'deduplication for the purpose of cash assistance'. Here, deduplication aims at identifying whether an individual or household receives overlapping cash assistance based on the definition of commonly agreed cash assistance packages.

D. **Referral of cash recipients**
   This variation focuses on referrals between CVA actors and CVA programmes. It needs to be considered in conjunction with feedback and protection mechanisms.

Respondents agreed that each variation or purpose is highly context-specific, depending on the operational setup, organizational mandates and programme objectives. Coordinated cash assistance packages (A) were considered a prerequisite and useful for deduplicating cash recipients. In order to deduplicate identities and cash recipients, partners would need to agree on harmonizing the cash assistance or packages for complementary cash assistance. For example, the lack of definition of household composition was frequently mentioned as a challenge alongside targeting approaches towards households or individual household members receiving MPCA which is coordinated by in-country Cash Working Groups (CWG) and/or Sector Cash coordinated by respective Clusters or Sector. Tech vendors highlighted the challenge of setting-up data collection and deduplication systems on behalf of a CWG in contrast to customizing a system for a consortium that agree on common porgramme parameters or cash assistance packages.

[15] see also Cash Working Group Ukraine. 2024; Tonea, D. al. 2022.

Data standards for deduplicating registration or identities (B) were considered as straightforward, easy to agree on and easy to manage. Most actors use more or less the same data standards for registering and deduplicating identities: names, data of birth (DOB), location, etc. Data standards for the deduplication of cash recipients (C) were perceived as more complex, i.e. crucial but very difficult when numerous players are involved. Referrals of cash recipients (D) were considered important but not raised by many respondents. All respondents highlighted that data standards, no matter the purpose, need to be discussed as part of an overall deduplication process and governance framework.

Some tech providers also mentioned the need to differentiate between the different setups and different layers of deduplication:

- Deduplication serves one partner, addressing programme- or organization-internal purposes.
- Deduplication serves multiple partners that work in a consortium, e.g., CCD Network.
- Deduplication aims at the wider cash ecosystem in a specific country or context, e.g., coordinated by a CWG.

Deduplication is further informed by the specific tech solution or platform. In case of consortia and CWGs, tech providers commonly agreed that it would be easiest if one organization functions as the lead agency.[16] Ideally, only one solution is applied by different organizations, even though the use of APIs is not considered a major issue or bottleneck.[17]

Deduplication and adjudication processes

Many respondents provided anecdotal evidence about deduplication requirements. For example, in Somalia, humanitarian organizations use 10 different types of functional ID in the absence of a foundational ID, with another additional 10 to 20 functional IDs that are sometimes used (e.g., pre-war passports or passports from Southern authorities).[18] Others referred to common spelling mistakes when Arabic or Cyrillic names are recorded in a data system that is based on the Latin alphabet. Common birth dates or phone numbers that are used by multiple family members were also mentioned. To reflect any context-specific requirements, data standards are as important as business rules that define deduplication processes, roles and responsibilities. Literature and respondents hence consider the following data points and/or process steps important for humanitarian organizations to agree on:

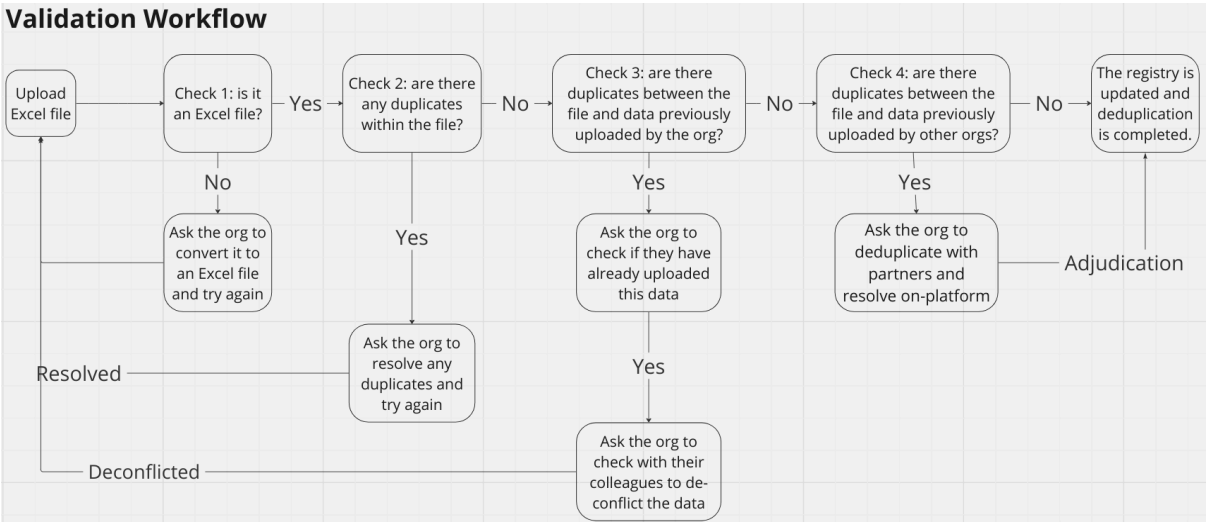
- Common definition of ‘duplicate’
- Common cash assistance packages (e.g., frequency of cash disbursement, transfer value, caseload, type of assistance)
- Common data points to collect at time of registration (e.g., first - last name, DOB, location, gender) and common format for data entry including the transcription of names or required data fields (e.g., first - middle - last name, date of birth), possible values for those data fields (e.g., DD/MM/YYYY or age) and the data object relationships (e.g., max number of members of a household)
- Type(s) of foundational or functional ID to prioritize (e.g., tax number in Ukraine)
- Data points and parameters/ criteria for identifying potential duplicates, for example, enrolled HH focal point, HH size, identification element like ID number, phone

[16] The main rationale is to avoid signing different agreements with multiple partners.
[17] For example, the API for Genius Chain and CommCare.
[18] Worthington, R. et al. 2023b.

number, household ID, enrolment date, cash assistance date, delivery status, country of enrolment, DOB, source system
- Common data formats (e.g., CSV, JSON, JPG)
- Minimum levels to classify potential duplicates as unique and maximum levels to classify duplicates as duplicate as well as the threshold - identities with matching scores between the minimum and maximum levels - to classify any need for adjudication
- In case of match(es) identified, the actor to retain and exclude data from registration and/or enrolment lists.

The CCD Network, for example, suggest the following deduplication and adjudication process to be contextualized by its members:

- *"Each organization digitally registers their specific recipients for the CCD project collecting the mandatory data [...]. Each organization also collects additional data according to their own organizations policies and guidelines.*
- *Each organization ensures there are not duplicate records in their database of CCD project recipients.*
- *Before sharing the new record with the CCD community database (hosted by the lead agency), each organization queries the community database asking 'does a duplicate of new beneficiary record with mandatory data XYZ exist?'*
- *CCD Community database returns an answer either NO or POSSIBLY.*
    - *If NO, the new record is added to CCD Community database.*
    - *If POSSIBLY, the CCD community database's response includes the name(s) of the orgs that have registered a beneficiary that looks like a duplicate.*
- *Verification and resolution of duplicates would happen by a meeting (in person or by telephone) among organizations with potential duplicate(s).*
- *After verification and resolution is finalized,*
    - *If the record is a duplicate, one organization removes record from CCD project file*
    - *If the record is not a duplicate, both organizations keep their records and are included in the CCD community database."*[19]



[19] CCD. 2020. CCD Deduplication Process. A Narrative Description. https://docs.google.com/document/d/1OkS2RCqweUlTynnkGffuJ5nt243RNW-Q/edit

Source: CCD & DIGID South Sudan. May 2024. Deduplication Business Rules.

The system-wide cash coordination during the Ukraine response offers another example. The lack of a common definition of a 'duplication' was considered critical. As a result of the slow operational response, affected people had used different self-registration tools before receiving a first transfer or re-registered to add another family member.[20] CALP 2022 also reported *"a disconnect between how humanitarians deduplicate only heads of households (HH), and the reality on the ground - where all HH members have their own tax ID (HHs do not have single IDs). This can enable people from the same family to register multiple times with different agencies"* (Tonea, D. et al. 2022: 7). The CWG Task Team 3 was thus requested to develop a deduplication and adjudication SOP to

- *"map the decision-making process behind the deduplication system, including: key programmatic decisions around harmonization of cash programmes that are to be deduplicated and those that are not (for ex. one-offs)*
- *map processes and timeframes of participating agencies from registration to payments*
- *develop the decision-making tree during the deduplication process*
- *add a case management component and troubleshooting (for example to accommodate beneficiaries that want their data removed from the system, etc.)*
- *map any data sharing involved in the process"* (ibids.: 15).

These decisions were used to inform the BB set-up in Ukraine.

→ Deduplication and adjudication rules are considered as important, if not more important as data standards. The rules are informed by the different steps for deduplication and adjudication. While few examples for standardizing deduplication processes exist, adjudication processes are less standardized and managed manually with the support of local actors.

Data standards: Core set of data semantics and data syntactic

There are few humanitarian data standards including personal and non-personal data points that inform deduplication processes: Minimum Core Dataset of the UN Common Cash Statement (UNCCS)[21], CCD's draft standards for deduplication and referrals[22], HDX and its Common Operational Datasets (CODs)[23] including P-CODES and admin levels for geolocations and others, HXL for information management[24], Washington Group Questions on Disability Inclusion[25] or ISO[26]. The IASC and Clusters manage their own data standards and principles for coordinating collective action (e.g., Collective AAP Framework) sector-related activities (e.g., 4Ws) which are built on existing standards and internal best practices.

---

[20] see Tonea, D. et al. 2022; Worthington, R. et al. 2023a.

[21] https://www.unhcr.org/media/minimum-core-assistance-delivery-dataset-affected-populations

[22] https://www.collaborativecash.org/_files/archives/1693a9_7b99c7d095f94bafae243de42c256a41.zip?dn=Working%20Draft%20of%20Data%20Standards.zip

[23] https://data.humdata.org/dashboards/cod?

[24] https://hxlstandard.org/

[25] https://www.cdc.gov/nchs/data/washington_group/WG_Short_Measure_on_Disability.pdf

[26] https://www.iso.org/home.html

At the moment, there are no data standards agreed at the global level for either the deduplication of registration and cash recipients nor referrals. The most crucial ones are the UNCCS Minimum Core Dataset (2019) and the CCD's draft data standards (2024) which, in comparison to other standards, also include personal data:

- The **UNCCS Minimum Core Dataset** was developed as part of the common cash statement interoperability work stream[27] and builds on the UNHCR/WFP Global Data Sharing Addendum. It is a core dataset that aims at the effective delivery of cash transfers in interagency environments, data interoperability, harmonization of statistical output, the reduction of duplication of data collection, and application of robust data protection principles. Its purpose is to facilitate humanitarian assistance provided by UN Agencies. The dataset includes a set of core data points labeled as minimum data for household record, metadata requirements, definitions and the secure format of sharing.
- The **CCD draft data standards** for deduplication of registration and assistance as well as referrals were developed for and with CCD members operating across the Ukraine Response and South Sudan. They build on other standards such as UNCCS and provide a set of core data and metadata, including descriptions and equivalence with UNCCS. The data standards are understood as a work in progress and *"not as a final product that cannot be updated or adapted"*[28]. The standards are licensed under Creative Commons Attribution-ShareAlike 4.0 International[29], can be adapted, built upon and distributed.

Literature and respondents further referred to the usefulness of health and social protection standards like Health Level Seven (HL7)[30], Digital Convergence Initiative (DCI)[31] or Open Identity Exchange (OIX)[32].

Respondents further agreed that any data standard would need to be flexible to allow customization and contextualization of data semantics and syntactics.[33] Most humanitarian organizations use common data sets for the deduplication of registration and identities making data standards easy to agree on. The deduplication of cash recipients, however, are more complex and difficult to standardize. It was therefore suggested to rather focus on principles and best practices to guide the process of deduplicating cash recipients by recommending deduplication rules, including data dictionaries for data collection and deduplication, processes, roles and responsibilities.

The Playbook on Digital Social Protection (2024) provides several purposes and use cases for using data dictionaries including, for example, common definitions, reference units, and reference periods. Therein, they describe a data dictionary as "a repository that contains descriptions of all data objects consumed or produced by the software. An organized listing of all data elements that are pertinent to the system, with precise, rigorous definitions so that both user and system analyst will have a common understanding of inputs, outputs,

[27] UNCCS is supported by OCHA, UNHCR, UNICEF, WFP, and is built on 3 pillars or work streams: Procurement, joint programming and interoperability.
[28] CCD. N.d.
[29] ttps://creativecommons.org/licenses/by-sa/4.0/
[30] https://www.hl7.org/
[31] https://github.com/spdci or https://standards.spdci.org/standards/v/wip-social-registry-v1.0.0/social-registry/1.-crvs
[32] https://openidentityexchange.org/networks/87/item.html?id=700
[33] see also Pon, B. et al. 2023.

components of stores, and (even) intermediate calculations" (World Bank et al. 2024. 99). They consider data dictionaries as a "fundamental building block" (ibids. 55f) of a digital social registry providing, for example, the definition of the assistance unit (household or individual) and code books for household members and their relationships - see table below:

Sample code book for household members	Sample code book for relationships
1 - Head of Family 2 - Spouse 3 - Mother 4 - Father 5 - Son 6 - Daughter 7 - Brother 8 - Sister 9 - Domestic help	1 - Married 2 - Polygamous 3 - Married polygamous 4 - Never married 5 - Living together 6 - Divorced or separated

→ Data standards for the deduplication of registration and identities are easy to agree on. For the duplication of cash recipients, data principles are more relevant.

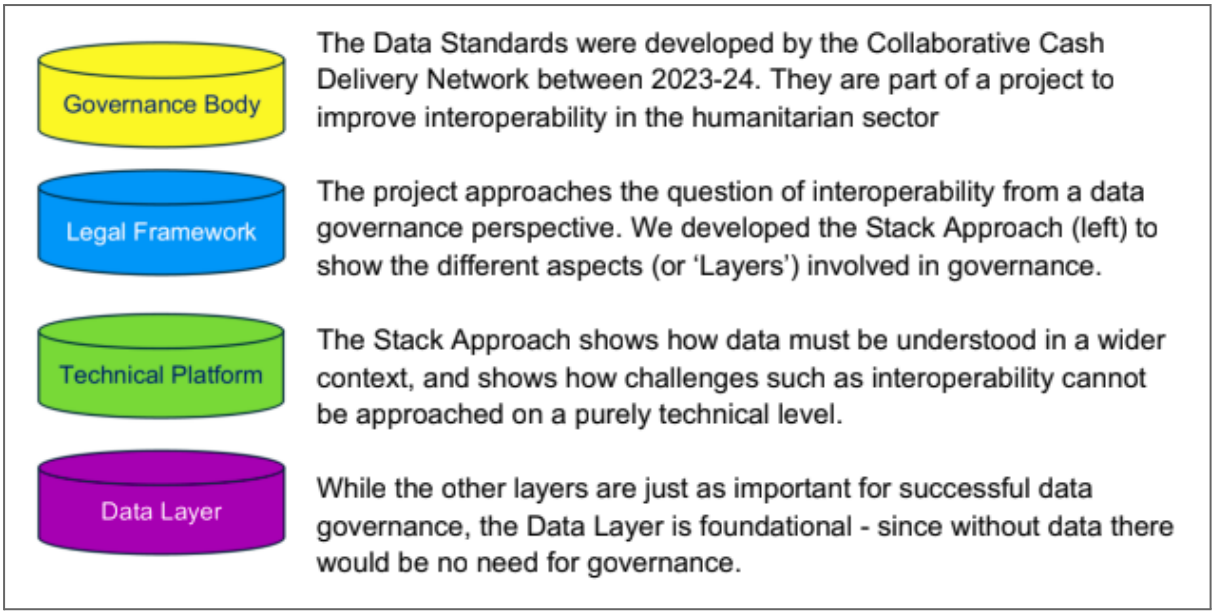
Governance frameworks

Governance for deduplication purposes was mainly considered as part of cash coordination and data sharing agreements that are informed by different data protection regimes. No respondent focused on data governance in detail but indirectly referred to a stack model.

CCD and DIGID experiences provide insights and learnings from applying a stack model to address different aspects of deduplication and more broadly interoperability Based on the stack model, data standards are considered a foundational layer for data governance, followed by technical, legal and governance layers.[34]

CCD Stack model

[34] CCD. N.d.



Source: CCD. N.d. Data Standards for Interoperability: Guidance Note: 2.

DIGID Stack Model

Figure 1. Interoperability 'stack'

System interoperability	Organizational	Ecosystem	User protection
Legal Data sharing agreements and other contracts; regulations at local, national and regional level, including GDPR	To what extent can data standardization lead to more standardized and efficient legal regimes (e.g. DSAs)?	Could different operating models for interoperability create new fiduciary or stewardship roles?	How can privacy policies, access to data, redress mechanisms, and other protections be maintained?
Governance Operations and processes, including workflows, data management, informed consent, user access, security policies	Is it possible to standardize workflows/ SOPs or decision trees around referrals or deduplication?	To what extent can this initiative engage technology vendors to support key system functionality?	How can informed consent and use authorizations be standardized and persist with data as it is shared?
Semantic Definition of data fields and possible values (e.g., gender), relationship/ hierarchy of data objects (e.g. household <> individual)	As likely requires more flexibility than technical standards, will semantic standards sit at country level? Regional?	How might these related to existing SDO (standards development organization) efforts in other sectors?	How can the design consider user input and support localization (e.g. alternative definitions of household)?
Technical Device hardware, communication protocols, data formats (e.g. XML, JSON, HXL), storage, APIs	What is the right entity to drive alignment of technical standards? A new working group?	Assuming an open-source stack, how is a community built to design and support ongoing development of technical components?	How will user needs be prioritized to ensure that the technology used does not exclude already marginalized individuals?

Adapted from European Commission, "New European Interoperability Framework"

Source: Pon, B. et al. 2023: 12.

Other, non-humanitarian actors, highlight the importance of trust building to be considered as part of any governance framework and data strategy, including:

- external factors such as laws, regulations and norms, contracts, penalties, standards, codes of conduct, ethical and organizational design, organizational governance; and

- internal factors such as reputation, competence and skills, presence of a pre-existing trust relationship;[35] as well as
- regular feedback loops to monitor effects and impacts of relevant processes.[36]

→ Data standards need to reference relevant deduplication rules/ processes and governance framework.

What drives deduplication? What hinders deduplication?

Drivers

The following aspects were mentioned as potential drivers and enablers for deduplication:

- Sector-wide funding cuts to foster discussions about cost efficiency and effectiveness
- A mapping of currently used tech/ algorithm (e.g., Elasticsearch, fuzzy matching)
- A mapping of existing gaps of currently used data standards
- A mapping of topics about deduplication
- Open source models as a more organic form of collaboration, (e.g., common software repositories).

Challenges and bottlenecks

There are different challenges which were identified as bottlenecks for operationalizing deduplication.

Operational challenges

Relevant data is ideally collected at the onset of a humanitarian crisis. While this, in principle, is a possible option for sudden onset emergencies, the bulk of cash assistance happens in protracted crises where data has been already collected by different organizations, and is managed in different databases using different solutions and customer interfaces designed for different mandates, targeting criteria and caseloads. While the type of data points are informed by organizational mandates, eligibility criteria and caseloads, all stakeholders struggle with issues around data quality and data accuracy.

A digital public infrastructure at national level or social protection scheme to tap into is rarely existing and thus forcing humanitarian organizations to build parallel data systems that are used for organizational purposes, notably identity and transfer management. In comparison to the health sector which generally focuses on individual case management, many humanitarian organizations practice HH-level targeting and assistance, and usually do not collect individual data of all HH members.

Humanitarian organizations that process personal data of cash recipients are considered as data owners. This role comes with a number of responsibilities to protect people's data and privacy on the one hand and to provide data rights for cash recipients, including the right to access their data or the right 'to be forgotten', on the other. Many humanitarian organizations

[35] Open Data Institute. 2021: 56, 48.
[36] https://connectedbydata.org

face difficulties to operationalize those data subject rights[37]. Concepts of data stewardship or data portability are little known and rarely explored.[38]

Few respondents who closely work with local actors also raised concerns about their partners not having the resources, capacities and capabilities to access sophisticated data systems or cash delivery mechanisms/ platforms. They rely on international peers to support open source technology and develop new ODK functionalities based on relevant (industry) standards and principles.

Legal challenges
Different data protection legislation such as national or regional data protection regimes are a major bottleneck for humanitarian organizations and tech providers, causing major delays and governance issues. Time, resources and capacities for establishing legal frameworks in form of data-sharing agreements are considered as one of the main operational bottlenecks for deduplication. Global templates such as the UNHCR/WFP Data Sharing Addendum, the CCD Data Sharing Agreement are available but lack sufficient clarity, knowledge and experience to contextualize and operationalise data protection on the ground (e.g., segregation of duties, data access, data retention etc.).[39]

Governance challenges
Governance challenges are considered another major bottleneck for operationalizing deduplication. Respondents agreed that there is 'no natural home' for this type of discussion, leaving deduplication as an issue to be handled by different organizations in parallel. It would need an organization who 'owns' and drives the problem.

Technical challenges
Most respondents consider technical challenges to be minimal. Technical solutions are rather impacted by all other challenges and thus easily perceived as a bottleneck.

What are potential ways for uptake of data standards for deduplication?

As a way forward to discuss and agree on data standards for deduplication, respondents suggested:

- Refer to best practices: Tech providers reference and share commonly agreed and/or applied data standards with its users. Standards do not have to be vetted but operationally useful and successfully practiced by a group of organizations such as the DIGID Consortium or CCD Network members.
- Build the case: More publicly available evidence and case studies are needed and shared through different fora such as CALP.
- Develop de/duplication indicators: Indicators that report on the # of organizations involved in deduplication or xx % levels of duplication against xx USD of cash

[37] All data subject rights are prescribed in the EU General Data Protection Regulations (GDPR).
[38] Fore more information, see Currion, P. 2022. Data Portability and Digital Identity in Humanitarian Aid: A Desk Review. CCD; ibid. 2022b. Safe Passage: Options for Data Portability in the Humanitarian Sector. CCD; Duechting, A. 2022. Digital Accountability. CHA. See also
CCD. N.d. A case for collaborative ownership
https://docs.google.com/document/d/1gALGSu24sDSPSJvwgcWDMpcAokaj25dc/edit
[39] Worthington, R. et al. 2023.

assistance are included in 4Ws and/or HRP/HNO, coordinated and reported on by CWGs.
- Work with champions: The identification and work with operational champions is key when referring to any best practices or case studies and reporting on indicators.
- Use the momentum: Respondents mentioned the sector-wide funding cuts to again boost discussions about cost efficiency and effectiveness.

The uptake of HXL as a standard[40] and the Digital Convergence Initiative (DCI)[41] could serve as an example for successful uptake. The success factors can be summarized as follows:
- Collect evidence to demonstrate the benefit of better data coordination to decision-makers, incl. cost benefit, levels of deduplication.
- Identify an owner of the problem who feels responsible, has the time and capacity to work on a solution and facilitate the process. Then, conceptualize the idea, structure it well and present it to others (inside/outside the sector).
- Identify experts and establish a working group or steering committee, technical team and programme team to build consensus, test, adjust, identify and work with champions. Meet on a regular and ad hoc basis.
- Know your end users and follow a user-centric design. Use concepts and language that people understand and know already.
- Regularly share learnings (e.g., meetings, conferences, blogs, webinars, etc.) and ask feedback (e.g., feedback loops, surveys, etc.).

→ The discussion about global standards needs to go hand in hand with in-country discussions.

In short: Summary and follow-up

As a way forward to discuss and agree on data standards for deduplication, respondents suggested that
- data standards for deduplication are defined in accordance with its objective, purpose and the specific operational context.
- deduplication and adjudication rules are considered as important, potentially more important than data standards. The rules are informed by the programme or cash assistance packages, different process steps and governance frameworks. While few examples for standardizing deduplication processes exist across the sector (e.g., Ukraine), adjudication processes are less standardized and usually managed manually with the support of local actors.
- data standards for the deduplication of registration/ identities are easy to agree on; for the duplication of cash recipients and referrals, data principles and guidance are more relevant.
- the development of global standards or guidance needs to go hand in hand with in-country discussions (e.g., at CWG or cluster-level), be built on case studies and best practices, be applied in a flexible manner, shared amongst different stakeholders, and use easy-to-understand and easy-to-apply messages (e.g., using conditional examples).
- tech providers are involved in shaping the discussion.

[40] Warner. T. A. 2016.
[41] Digital Convergence. 2024.

- local actors are reflected and involved in the discussion and capacitated by sharing best practices and making easy-to-understand knowledge and easy-to-use tools accessible.

The majority of respondents agreed that flexible, concrete and concise guidance based on evidence and/or use cases and user-friendly design is needed to guide humanitarian practitioners in deduplication questions. The guidance should be simple and include easy-to-understand messages and tools.[42] Below statements could serve as an entry point.

Why does deduplication matter?

Deduplication is usually considered as a process/ mechanism for cost efficiency. It aims at supporting humanitarian actors to be more efficient and effective by improving collaboration through identifying duplicates, coordinating assistance packages, referring people in need to mandated organizations and contributing to creating choice models for them. At the same time, deduplication has a moral implication by providing cash assistance to people affected by crises in a more dignified, transparent and accountable manner and through safe data sharing between organizations and affected individuals.[43]

Why does the humanitarian system need data standards for deduplication?

Data standards are the foundation for improved cash coordination, data sharing and interoperability. Research conducted by DIGID and a gap analysis of existing data standards conducted by the CCD Network revealed that there are currently no standards to support operational deduplication processes for humanitarian cash coordination. Hence, a discussion about data standards for improved cash coordination is needed.

What deduplication rules apply?

Deduplication rules are about clearly defined parameters, processes and solutions/ tools to identify, flag and manage potential duplicates of cash recipients. The process of flagging potential duplicates is usually automated while verification of identities and adjudication are conducted manually with the support of frontline workers, local communities and authorities.

What is adjudication about?

Adjudication describes the manual or automated process that allows for the administration of potential duplicates. Questionable matching results such as potential duplicates are flagged and usually forwarded to human experts for manual adjudication in the data system. The objective of adjudication is to assist the deduplication procedure when an absolute decision on the uniqueness of an identity cannot be made.

[42] A good example for easy-to-understand messages is the SDG Cookbook on Effective and Ethical Data Sharing, see Global Partnerships for Sustainable Development Data. 2024. Effective and Ethical Data Sharing at Scale. https://www.data4sdgs.org/effective-and-ethical-data-sharing-scale
[43] see DIGID Interoperability Initiative. https://interoperability.ifrc.org/projects/interoperability/

Appendices

Interview questions

The interviews were based on a set of key questions summarised in an interview guide shared with all interviewees and adjusted depending on the stakeholder group.

1. Core set of commonly agreed data semantic standards for deduplication of aid recipients, assistance packages and referrals

- Kindly have a look at the working draft of data standards for deduplication of cash recipients and referrals developed by and for CCD members and let us know what you think about this data model which builds on existing standards and was commonly agreed by CCD/DIGID consortium partners. How useful are these standards for your own organisation? Would you apply commonly agreed data models/ standards if widely used and recommended across the humanitarian sector?
- Process: Do you use any CVA data standards across your organisation? How do you translate these standards into your operations/systems (e.g., roles and responsibilities, decision-making, etc.)? What would change if the humanitarian sector recommends the use of commonly agreed data standards? What is missing and important to keep in mind?
- Solutions: What type of solutions are you using? What are your learnings and recommendations?
- Data points: What are your minimum or required set of data fields? Are they similar to the ones identified/suggested by CCD? How useful do you find this data model for your own organisational purpose(s)?
- Data governance: WIth relevance to deduplication, what does or should the data governance model look like? Who is/ needs to be involved in decision-making processes and how?
- Accountability: To your own experience, what is or should be the main driver for widely used data models/ standards? Who should host or govern the standards? How do you engage local actors and aid recipients in the process?
- Outlook: In your opinion, what is or should be the pathway for agreeing on commonly used data models/ standards for deduplication of cash recipients? What is missing?

2. Commonly used adjudication business rules

- Process: What are your experiences in conducting deduplication (context, process, stakeholders, challenges etc.)? Do you have any business rules for deduplication of cash recipients? What are the use cases? What are the business rules for adjudication used in your organisation (if any)?
- Decision-making: Who is/ needs to be involved in decision-making processes? How do you accommodate local/ context-specific decisions? How do you translate these business rules against your own data management solution (if at all)?
- Accountability: How do you engage local actors and aid recipients in the process? Who should facilitate this discussion about recommended adjudication business rules?
- Outlook: In your opinion, is there a need for commonly agreed / used business rules for adjudication? If yes, how to agree on these rules? If no, why not?

List of interviewees

Organisation / Entity		**interview date**
AIDONIC	Tech vendor	10.07.2024
AidKit	Tech vendor	15.07.2024
CAFOD	NGO	26.06.2024
CCD	Network	multiple informal discussions
DIGID	Network	multiple informal discussions
Dimagi / CommCare	Tech vendor	16.07.2024
DRWG CTT	IASC	multiple informal discussions
Global Food Security Cluster	Cluster	29.05.2024
Genius Tags	Tech vendor	15.07.2024
IASC TF 2	IASC	23.05.2024
Independent		27.06.2024
IOM	UN	06.08.2024
IOM	UN	18.07.2024
Kobo / ODK	Tech vendor	24.07.2024
LMMS	Tech vendor	26.06.2024
RedRose	Tech vendor	17.07.2024
UNHCR	UN	27.06.2024
UNICEF	UN	10.07.2024
WFP	UN	07.06.2024
WFP	UN	31.07.2024
WFP	UN	16.07.2024

List of references

A bibliography and full list of useful references and resources can be found here.

CALP. 2023. State of the World's Cash. https://www.calpnetwork.org/wp-content/uploads/2023/11/The-State-of-the-Worlds-Cash-2023-1.pdf

Cash Working Group Ukraine: Registratio, Deduplication and Interoperability Task Team. 2024. TT3 Meeting Minutes. https://response.reliefweb.int/ukraine/cash-working-group-cwg/cash-working-group-task-team-3-registration-and-deduplication

CCD. N.d. Data Standards for Interoperability: Guidance Note. https://drive.google.com/drive/folders/1x6UNt79Qeutlk25ZDREXoSm80Isikv9K

Currion, P. 2024. Working Draft of Data Standards. CCD. https://www.collaborativecash.org/_files/archives/1693a9_7b99c7d095f94bafae243de42c256a41.zip?dn=Working%20Draft%20of%20Data%20Standards.zip

Digital Convergence. 2024. WIP- Integrated Beneficiary Registry v1.0.0. https://standards.spdci.org/standards/v/wip-integrated-beneficiary-registry-v1.0.0

Donor Cash Forum. N.d. Donor Cash Forum Statement and Guiding Principles on Interoperability of Data Systems in Humanitarian Cash Programming. https://www.calpnetwork.org/wp-content/uploads/ninja-forms/2/DCF-Interoperability-Statement-FINAL.pdf

Kreidler, C. and Rieger, N. 2022. Increasing the use of humanitarian cash and voucher assistance. Opportunities, barriers and dilemmas. CALP. https://www.calpnetwork.org/wp-content/uploads/2022/12/Increasing-the-use-of-humanitarian-cva-opportunities-barriers-and-dilemmasupdated.pdf

Open Data Institute 2021. The economic impact of trust in data ecosystems. https://theodi.org/insights/reports/the-economic-impact-of-trust-in-data-ecosystems-frontier-economics-for-the-odi-report/

Pon, B., Lyamuya, A., Schoemaker, E. and Nunez, H. 2023. Investigating save data sharing and systems interoperability in humanitarian cash assistance. IFRC / DIGID. https://cash-hub.org/wp-content/uploads/sites/3/2023/11/DIGID-Interoperability-Investigating-Safe-Data-Sharing-and-Systems-Interoperability.pdf

Tonea, D. and Palacios, V. 2022. Registration, Targeting and Deduplication: Emergency Response inside Ukraine. Thematic paper. CALP. https://www.calpnetwork.org/wp-content/uploads/2022/09/Registration-Targeting-and-Deduplication-Emergency-Response-inside-Ukraine-Thematic-paper-1.pdf

Warner, A.T. and Obrect, A. 2016. Standardizing humanitarian data for a better response. The Humanitarian eXchange Language. ALNAP Case Study. https://www.urban-response.org/system/files/content/resource/files/main/alnap-innovation-hxl-case-study.pdf

WHO. N.d. Go.Data Interoperability Toolkit. https://worldhealthorganization.github.io/godata/interoperability/

World Bank and ISPA. 2024. Playbook on Digital Social Protection. Delivery Systems. https://socialprotection.org/discover/publications/playbook-digital-social-protection-delivery-systems-towards-dynamic-inclusio-0

Worthington, R. and Duechting, A. 2023a. Enabling Dignified Humanitarian Assistance Through Safe Data Sharing. Landscape Mapping. IFRC / DIGID. https://interoperability.ifrc.org/.

Ibids. 2023b. Use Case 1: Deduplication of people, families or households. IFRC / DIGID. https://cash-hub.org/wp-content/uploads/sites/3/2023/05/DIGID-Interoperability-Deduplication-of-people-families-or-households.pdf

Ibids. 2023c. Use Case 3: Individual referral. Sharing data on a person with a partner, donor or government for follow-up services. IFRC / DIGID. https://cash-hub.org/wp-content/uploads/sites/3/2023/05/DIGID-Interoperability-Individual-referral-Sharing-data-on-a-person-with-a-partner-donor-or-government-for-follow-up-services.pdf