2/6/2026

Van Weken naar Seconden

1CijferHO Data Eindelijk Toegankelijk

2/6/2026

Ash Sewnandan & Tomer Iwan

Data is de nieuwe standaard

In 2025 is data-gedreven werken geen luxe meer

1CijferHO

Wat is het?

Centrale database voor alle HO-cijfers in Nederland

- Beheerd door DUO
- Data vanaf 1991
- Alle inschrijvingen, uitval, diploma's

Wat kan je ermee?

De basis voor evidence-based beleid

- Beleidsanalyses
- Benchmarking
- Trendanalyses
- Voorspellende analyses

Maar...

Van aanvraag tot bruikbare data

Weken tot maanden ⏱️

Wat neem je mee?

Waarom

Het probleem begrijpen

Hoe

De oplossing zien

Wat

Zelf aan de slag

Van weken → seconden

Maar als deze data zo waardevol is...

Waarom gebruikt niet iedereen het?

Je krijgt van DUO een ZIP

Bestandsbeschrijving
PDF/tekst met specs

Decodeer bestanden
Lookup tables

Main bestanden
De data zelf

Lijkt simpel, toch?

ZIP Bestandslijst

```
📁 1cijferho_2023.zip
  ├── 📄 bestandsbeschrijving_HO_2023.pdf
  ├── 🔑 decodeer_geslacht.txt
  ├── 🔑 decodeer_nationaliteit.txt
  ├── 🔑 decodeer_opleiding.txt
  ├── 💾 inschrijvingen_2023.dat
  ├── 💾 diplomagegevens_2023.dat
  └── 💾 uitstroom_2023.dat
```

[Screenshot placeholder - voeg zip-lijst.png toe in public/screenshots/]

De realiteit: Main bestanden

```
0100123456789012345678 90NLAMSTERDAM20231001M...
0100987654321098765432 10JAUTRECHT  20231001V...
0200567812345678123456 78NROTTERDAM20231001M...
0300876543218765432187 65JADEN HAAG20231001V...
0100345678903456789034 56NUTRECHT  20231001M...
```

Lange strings. Geen headers. Geen separators.

Alles zit tegen elkaar geplakt 😵

Bestandsbeschrijving Reality

```
BESTANDSBESCHRIJVING INSCHRIJVINGEN 2023
Kolom Start Lengte Type Beschrijving
1 1 3 N Instellingscode
2 4 9 A Onderwijsnummer
3 13 1 A Geslacht (M/V/X)
4 14 20 A Woonplaats
5 34 8 N Geboortedatum (YYYYMMDD)
6 42 1 A Hoofdinschrijving (J/N)
...
[veel meer rijen met onduidelijke formatting]
```

Je moet dit handmatig extraheren 😫

[Screenshot placeholder - voeg beschrijving-chaos.png toe]

De oude workflow

1. **ZIP downloaden** — 5 min
2. **Bestandsbeschrijving lezen** — uren
3. **Schema's extraheren & configureren** — uren
4. **Data inladen (read_fwf)** — 30+ min
5. **Valideren & debuggen** — weken

WEKEN WERK

of helemaal niet gedaan

Herkenbaar?

Wie heeft dit weleens geprobeerd?

Van Weken

naar Seconden

Onze aanpak

Parse

Bestandsbeschrijving automatisch

Extract

Schema's zonder mens

⚡

Load

Efficiënt, niet read_fwf

Validate

Corrigeer automatisch

❌ VOOR

1. **Bestandsbeschrijving lezen**
   uren
2. **Schema extraheren**
   uren
3. **Data inladen**
   30+ min
4. **Valideren & debuggen**
   weken

✅ NA

1. **Parse**
   <1 sec
2. **Extract**
   <1 sec
3. **Load**
   <8 sec
4. **Validate**
   automatisch

WEKEN
Handmatig · Foutgevoelig

SECONDEN
Automatisch · Gevalideerd

Stap 1: Parse

Bestandsbeschrijving automatisch lezen

```
parse_file_description()
extract_schema_tables()
validate_specifications()
```

📄 PDF/tekst wordt gelezen

🔍 Tabel-structuur wordt herkend

✨ Alles wordt automatisch geëxtraheerd

Van chaos naar structuur

Geen handmatig werk meer

Stap 2: Extract

Schema's zonder menselijke interventie

```
build_column_definitions()
map_data_types()
handle_encoding()
```

📋 Kolommen worden gedefinieerd

🏷️ Data types worden gemapt

🔤 Encoding wordt afgehandeld

Van chaos naar:

Hier zijn je kolommen, types, posities

Stap 3: Load

Efficiënt laden (niet read_fwf!)

```
stream_fixed_width_data()
parallel_processing()
memory_optimization()
```

🌊 Streaming in plaats van bulk

⚡ Parallel processing

💾 Memory geoptimaliseerd

Van 30+ minuten naar:

< 10 seconden

Stap 4: Validate

Automatische checks & correcties

```
validate_data_integrity()
cross_reference_decode_files()
flag_anomalies()
```

✅ Data integriteit wordt gevalideerd

🔗 Decodeer bestanden worden gekoppeld

⚠️ Anomalieën worden gemarkeerd

Geen debuggen meer

Warnings bij problemen

Voor iedereen

🖥️ Via UI

Streamlit Interface

📁 Drop je files

🖱️ Klik

✅ Klaar

Geen code nodig

⚙️ Programmeerbaar

```
from dair import load_1cijferho

df = load_1cijferho(
"path/to/zip"
)
```

Klaar!

Integreer in je workflow

Privacy & Veiligheid

Volledig lokaal

Draait op jouw machine

Geen data uitwisseling

Verlaat nooit je computer

Open source

Transparant & verifieerbaar

Gratis

Voor iedereen

Zero trust. Jij blijft eigenaar.

Genoeg gepraat

Laten we het laten zien

Live Demo Time

[Switch naar Streamlit applicatie]

Demo: Wat je zag

Upload

ZIP files droppen

Process

Automatisch parsen

Analyze

Direct inzicht

< 10 seconden ⚡

Wat komt er?

Web interface

Geen installatie nodig

API endpoints

Integreer in je systemen

Meer analyses

Built-in visualisaties

Meer databronnen

Koppeling met andere datasets

Hoe te gebruiken?

📥

Download

🔗 GitHub repository

📦 pip install

🚀 Direct starten

📖

Documentatie

📚 Handleiding

💡 Voorbeelden

❓ FAQ

Links komen op laatste slide

GitHub · Docs · Contact

Recap

✅ Probleem: weken wachten op data

✅ Oplossing: automatisch parsen

✅ Resultaat: seconden in plaats van weken

✅ Privacy: volledig lokaal

✅ Toegang: gratis & open source

Vragen?

Bedankt!

Van Weken naar Seconden

📧 [email@instelling.nl]

🔗 github.com/[repo]

📚 docs.dair-project.nl

Ash Sewnandan & Tomer Iwan