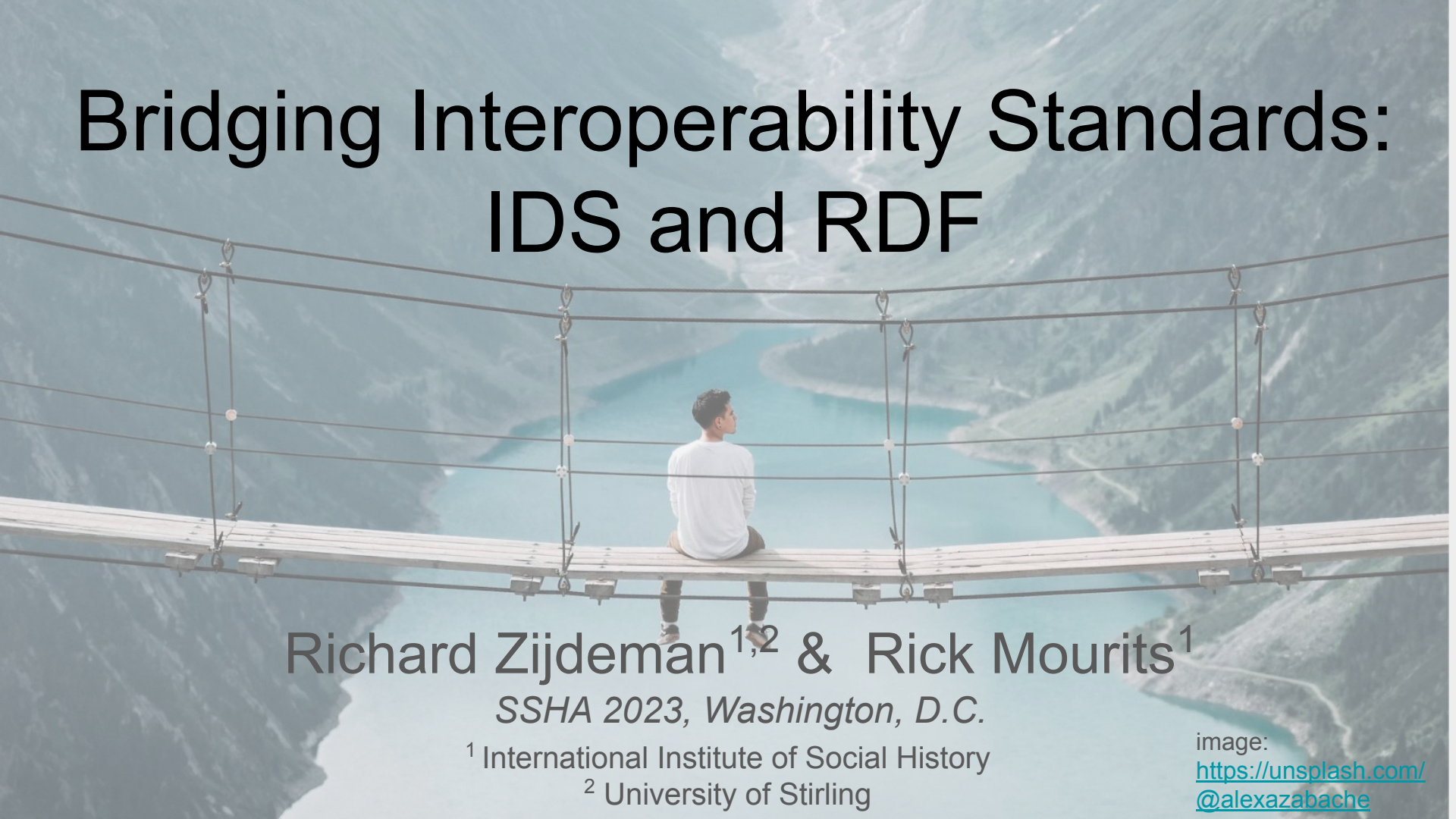
Bridging Interoperability Standards: IDS and RDF



Richard Zijdeman[1,2] & Rick Mourits[1]

SSHA 2023, Washington, D.C.

[1] International Institute of Social History
[2] University of Stirling

image: https://unsplash.com/@alexazabache

How to make datasets interoperable and reusable?

- FAIR principles for scientific data (Wilkinson et al. 2016)
- Already a question long before FAIR

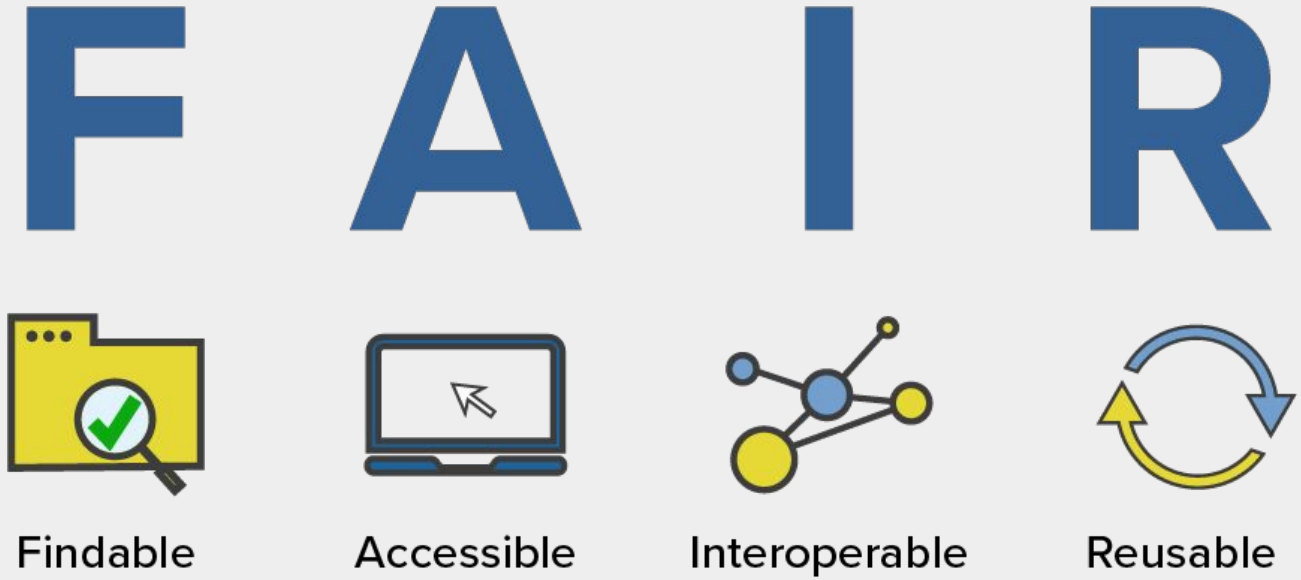


image source:NIH

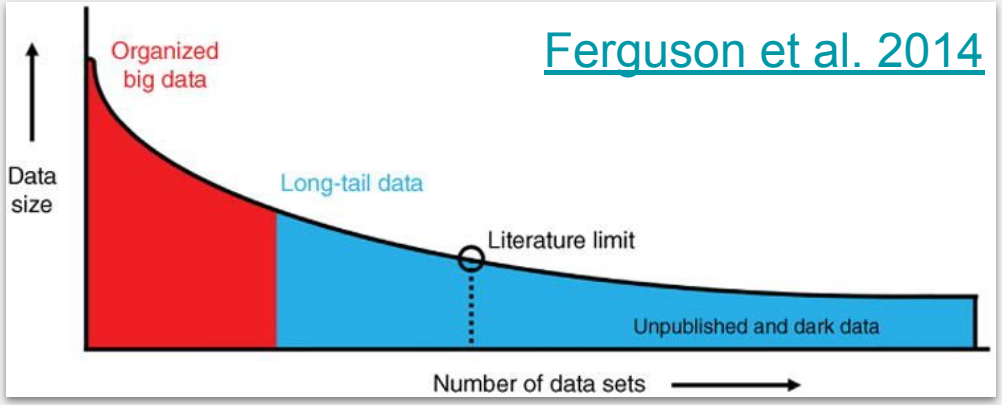
Two solutions

- Standardise csv's
  - NAPP, MOSAIC, IPUMS
  - Useful for standardising information from a similar source, e.g. census
  - Less ideal for combining different source types
- Use Entity Attribute Value model
  - 38355397 "Event" "Marriage"
  - 38355397 "Role" "Bride"
  - 38355397 "Marriage_date" "1818-03-27"
  - 38355397 "Age" 22
  - 38355397 "Occupation" "Dienstmeid"
  - 38355397 "HISCO" 54020
  - 38355397 "partner" 37810246

Why Entity-Attribute-Value?

To save the long tail of small data!

- e.g. data on housing prices, health, income
- too small to find safe guard under ‘wings of IPUMS-like initiatives’
- important for in-depth analysis of human behaviour
  - specific characteristics not available via certificate registers or census data
  - heterogenous, so allows for studying topic from multiple angles



Ferguson et al. 2014

Two applications of EAV model

- Intermediate Data Structure
  - Standardizes variables using metadata table
  - Solves EAV problem using time stamp
  - Applied in multiple databases & https://hlcs.nl/specialissue2

- Resource Description Framework (~= Semantic Web, Linked Open Data)
  - Standardizes variables using URIs (the Web)
  - Solves EAV problem using schemas
  - Applied to host Dutch civil registry & match millions of certificates using burgerLinker

EAV in IDS

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation	Day	Month	Year
38355397	Marriage certificate	M_date	*NA*	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_SEQUENCE	1	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_LOCATION	*NA*	11165	Event	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_STANDARD	dienstmeid	*NA*	Reported	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_HISCO	54020	*NA*	Reported	Exact	27	3	1818

EAV in IDS

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation	Day	Month	Year
38355397	Marriage certificate	M_date	*NA*	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_SEQUENCE	1	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_LOCATION	*NA*	11165	Event	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_STANDARD	dienstmeid	*NA*	Reported	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_HISCO	54020	*NA*	Reported	Exact	27	3	1818

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation
38355397	Marriage + Death certificate	BIRTH_DATE	*NA*	*NA*	Assigned	Age-based
38355397	Marriage + Death certificate	SEX	Female	*NA*	*NA*	*NA*

Start_day	Start_month	Start_year	End_day	End_month	End_year
28	3	1796	21	2	1797
NA	*NA*	*NA*	*NA*	*NA*	*NA*

EAV in RDF

“Mar_12347682”	bio:hasDate	“1818-03-27”
“Mar_12347682”	bio:hasPartner	38355397
“Mar_12347682”	bio:hasPartner	37810246
38355397	pico:hasRole	“Bride”
37810246	pico:hasRole	“Groom”
38355397	prov:hadPrimarySource	“Mar_cert_112347682”
37810246	prov:hadPrimarySource	“Mar_cert_112347682”
“Mar_cert_112347682”	schema:locationCreated	11165
“Mar_cert_112347682”	schema:additionalInfo	“Marriage certificate”
“Mar_cert_112347682”	schema:dateCreated	“1818-03-27”

Combine the strengths of both EAV models

- Via RDF, IDS data can be retraced to the source data
- Enhances provenance and explicates IDS decisions (e.g. on date ranges)
- New GLAM data are becoming available in RDF every day (and we’re missing out)

- Unlike RDF, IDS contains expert vocabularies to standardize variables
- Various key datasets for SSH research are available in IDS (not RDF)

IDS in RDF

Requires an explicit event to make rephrase it as RDF

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation	Day	Month	Year
38355397	Marriage certificate	M_date	*NA*	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_SEQUENCE	1	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_LOCATION	*NA*	11165	Event	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_STANDARD	dienstmeid	*NA*	Reported	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_HISCO	54020	*NA*	Reported	Exact	27	3	1818

How do we know that this is one event?

1. We have the same person identifier (Id_I)
2. And an identical timestamp (Day, Month, Year)

To model IDS in RDF, this implicit assumption needs to be made explicit

IDS in RDF

Requires an explicit event to make rephrase it as RDF

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation	Day	Month	Year
38355397	Marriage certificate	M_date	*NA*	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_SEQUENCE	1	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_LOCATION	*NA*	11165	Event	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_STANDARD	dienstmeid	*NA*	Reported	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_HISCO	54020	*NA*	Reported	Exact	27	3	1818

- IDS:event IDS:hasTimeStamp “1818-03-27”
- IDS:event IDS:hasSource “Marriage certicate”
- IDS:event IDS:sequence 1
- IDS:event IDS:hasHistoricalLocation 11165
- IDS:event IDS:OccupationStandard “dienstmeid”
- “dienstmeid” IDS:hasHISCO 54020

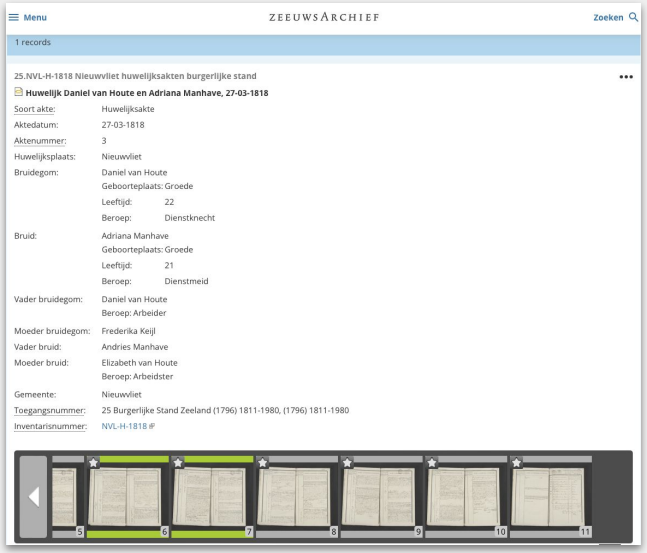
IDS in RDF

It now is possible to state in what format 11165 is coded and retrieve info on it

Id_I	Source	Type	Value	Value_Id_C	Date_type	Estimation	Day	Month	Year
38355397	Marriage certificate	M_date	*NA*	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_SEQUENCE	1	*NA*	Event	Exact	27	3	1818
38355397	Marriage certificate	MARRIAGE_LOCATION	*NA*	11165	Event	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_STANDARD	dienstmeid	*NA*	Reported	Exact	27	3	1818
38355397	Marriage certificate	OCCUPATION_HISCO	54020	*NA*	Reported	Exact	27	3	1818

- IDS:event — IDS:hasHistoricalLocation — 11165
- 11165 — IDS:hasLocationScheme — “AMCO”
- 11165 — AMCO:hasName — “Nieuwvliet”
- 11165 — AMCO:inProvince — “Zeeland”
- 11165 — AMCO:inCountry — “Netherlands”

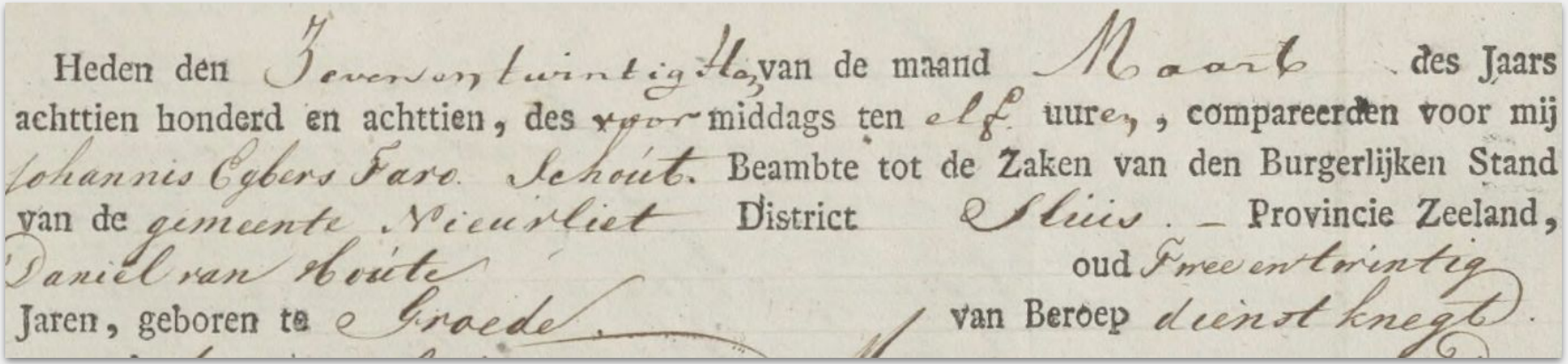
IDS in RDF



And even possible to roundtrip:

- link back to the source
  - credit source (archive)
  - quality (direct check on information correctness)

IDS:EVENT SDO:URL
<https://hdl.handle.net/21.12113/5884A85510884D14A643A259E6C242AB>


Heden den Zevenentwintigste van de maand Maart des Jaars achttien honderd en achttien, des voor middags ten elf uure, , compareerden voor mij Johannis Cybers Faro. Schout. Beambte tot de Zaken van den Burgerlijken Stand van de gemeente Nieuwvliet District Sluis . – Provincie Zeeland, Daniel van Houte oud Vreeentwintig Jaren, geboren te Groede van Beroep dienstknegt.

To summarise

- Entity Attribute Value (EAV) formats are highly suited to standardise data
- IDS and RDF are two implementations of EAV
- By transforming IDS into RDF:
  - IDS becomes compatible with other GLAM data
  - The provenance of IDS improves
  - Implicit assumptions in IDS become clear
  - Encoding schemes for place names can be more clearly specified
  - IDS can directly link to borrowed standardized vocabularies, such as HISCO
  - Other people can more easily use and refer to the standardized IDS vocabularies
- In other words, a lot of extra functionality for us data peeps
- In the (near) future makes it possible for people to travel between data sets!

Questions / comments?

mail: richard.zijdeman at iisg.nl
X (twitter): ‘rlzijdeman’