Put Relational Databases in Your Data Curation Toolbox

Author: **Jenna Jordan**

School of Information Sciences,

University of Illinois at Urbana-Champaign

SCAN ME

jennajordan.me

ILLINOIS

iSchool

19 - 23 OCTOBER 2019

CROWN MELBOURNE, AUSTRALIA

82ND ANNUAL MEETING

See this project on Github:

Jupyter notebooks • original data • transformed datasets • SQLite database • documentation & references

SCAN ME

https://github.com/jenna-jordan/international-relations-database

Abstract

This poster makes a case for the use of **relational databases as a data curation tool**, especially for datasets that are published separately but can be used together due to a *common identifier scheme* and *shared attributes*.

The Correlates of War datasets are used as an illustrative example to show how the *normalization process* results in a design with greater **data reusability**, while *check constraints* and *foreign key constraints* can improve **data quality**.

Catch Data Entry and Versioning Errors for Data Quality

Unlike a spreadsheet, a (well-designed) database does not allow data to be entered that violates the constraints.

These constraints force you to **check your assumptions** about the data, and *enforce those assumptions systematically and automatically*.

Wrong dates: when you forget to press tab

WarNum	StartMonth1	StartDay1	StartYear1	EndMonth1	EndDay1	EndYear1	StartMonth2	StartDay2	StartYear2	EndMonth2	EndDay2	EndYear2
585	10	-9	1866	10	-91866		-8	-8	-8	-8	-8	-8
623	2	29	1894	5	6	1894	9	14	1894	11	28	1894
682	1	6	1919	5	1919		-8	-8	-8	-8	-8	-8

✗

↓

WarNum	StartMonth1	StartDay1	StartYear1	EndMonth1	EndDay1	EndYear1	StartMonth2	StartDay2	StartYear2	EndMonth2	EndDay2	EndYear2
585	10	-9	1866	10	-9	1866	-8	-8	-8	-8	-8	-8
623	2	28	1894	5	6	1894	9	14	1894	11	28	1894
682	1	6	1919	5	-9	1919	-8	-8	-8	-8	-8	-8

✓

Check constraints and *strict datatypes* can catch these errors

```
StartDate      Date,
EndDate        Date,
StartYear      NUMBER(4),
StartMonth     NUMBER(2),
StartDay       NUMBER(2),
EndYear        NUMBER(4),
EndMonth       NUMBER(2),
EndDay         NUMBER(2),
```

Wrong IDs: when you forget to update every changed entity

number	year	month	gainer	entity	loser
427	1909	3	200	822	800
452	1914	5	200	822	822

Entity #	Name	Begin Year	End Year	Ending Political Status
820	Malaysia (Malaya)	1946	1957	Became colony of 200
821	Federated Malay States	1816	1896	Became part of 8201
823	Sabah (North Borneo)	1816	1888	Became part of 835
824	Sarawak	1816	1841	Became part of 835
8221	Johore	1914	1942	Became protectorate of 200
8222	Kedah	1821	1841	Became part of 800
8223	Kelantan	1909	1942	Became protectorate of 200
8224	Perlis	1816	1841	Became part of 800
8225	Trengganu	1909	1942	Became protectorate of 200

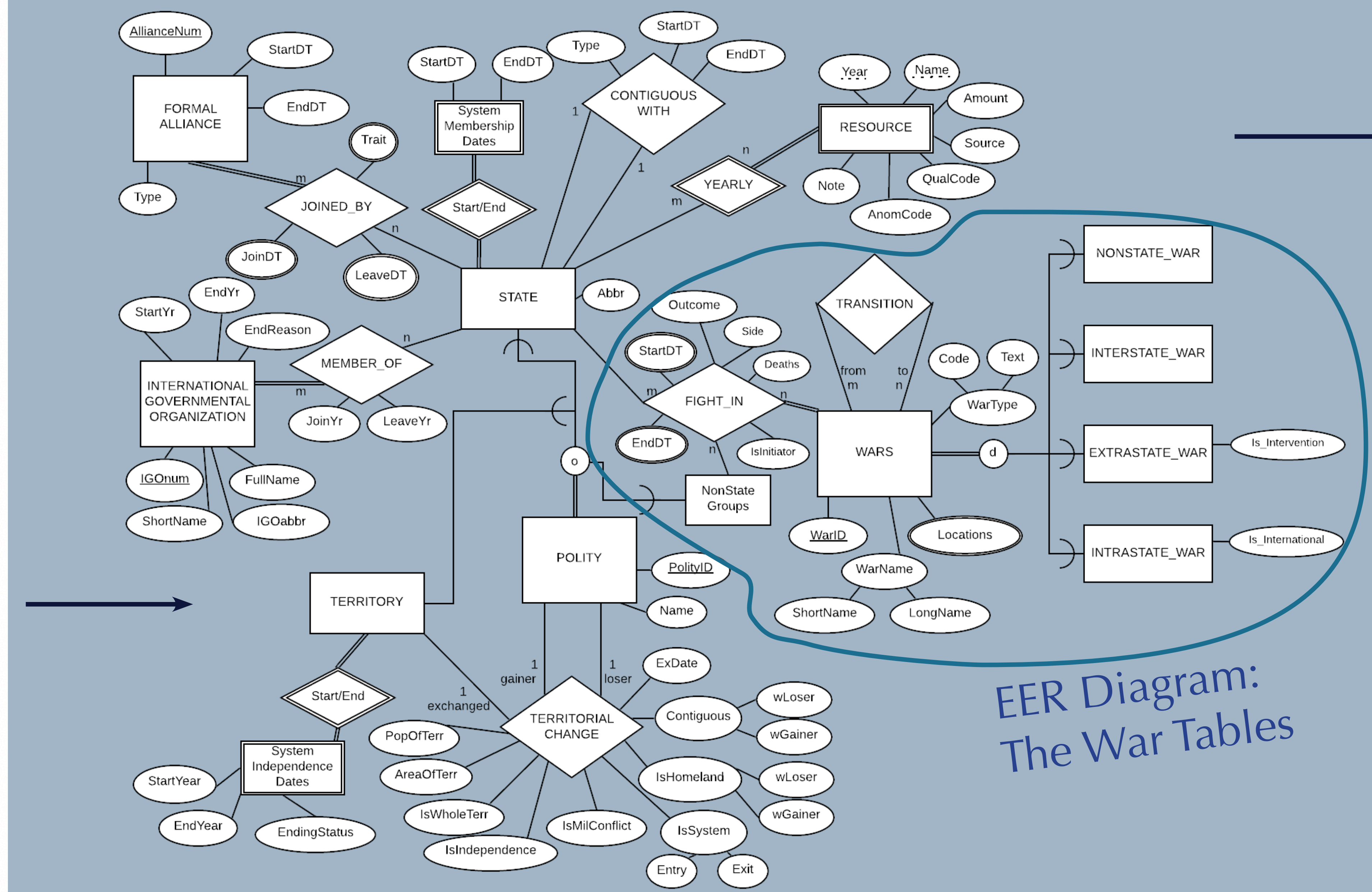
Foreign key constraints can catch these discrepancies

```
CONSTRAINT TERRGAINER_TO_POLITY FOREIGN KEY (Gainer)
   REFERENCES POLITY (PolityID),
CONSTRAINT TERRLOSER_TO_POLITY FOREIGN KEY (Loser)
   REFERENCES POLITY (PolityID),
CONSTRAINT TERR_TO_POLITY FOREIGN KEY (TerritoryID)
   REFERENCES POLITY (PolityID)
```

Tidy up and Normalize for Data Reusability

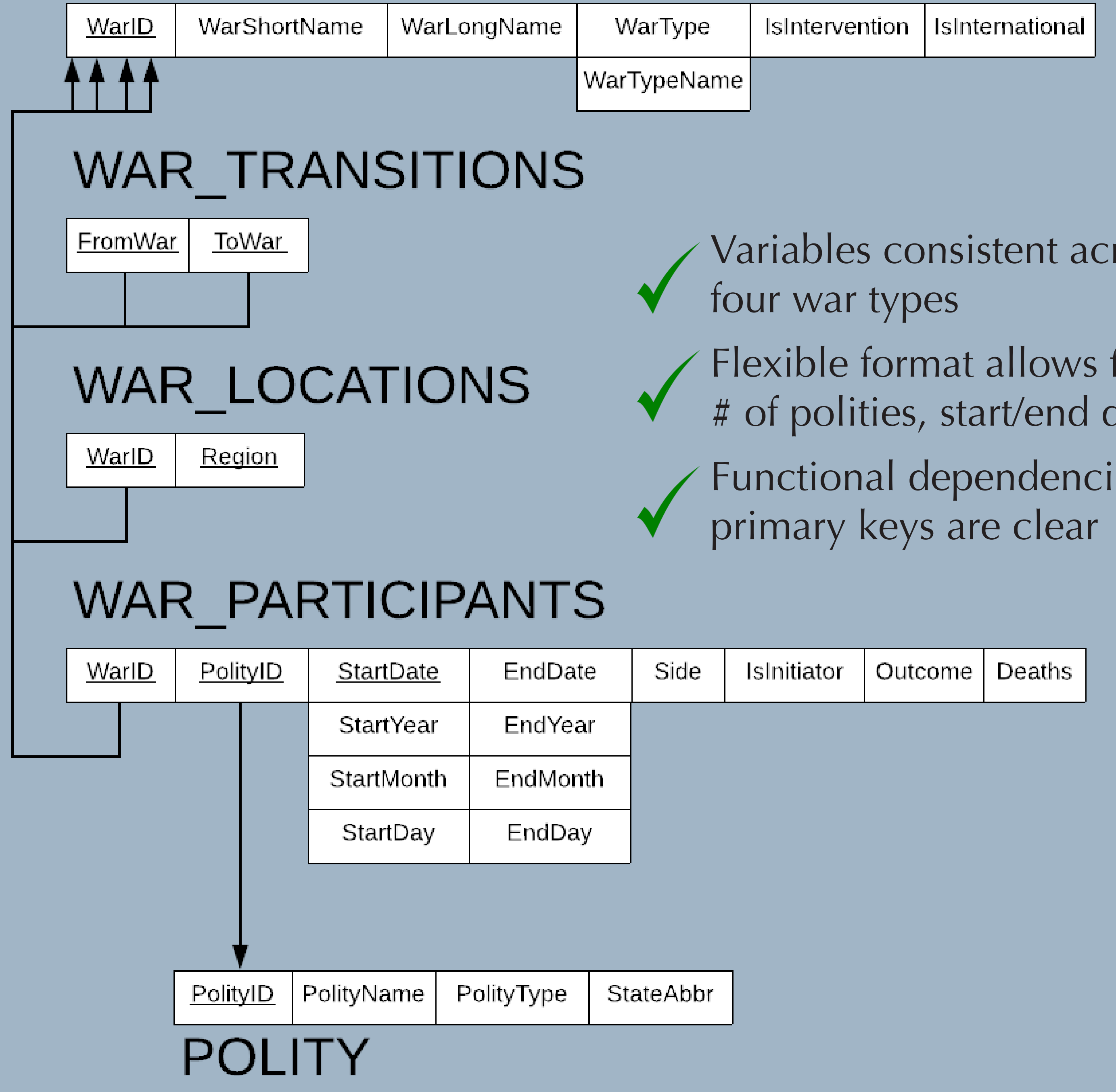
The Original 4 CoW War Tables

Inter-State War	WarNum	WarName	WarType				
	ccode	StateName	Side				
	StartMonth1	StartDay1	StartYear1	EndMonth1	EndDay1	EndYear1	
	StartMonth2	StartDay2	StartYear2	EndMonth2	EndDay2	EndYear2	
	TransFrom	TransTo	Initiator	Outcome	WhereFought		
	BatDeath	Version					
Intra-State War	WarNum	WarName	WarType	Intnl			
	CcodeA	SideA	CcodeB	SideB			
	StartMonth1	StartDay1	StartYear1	EndMonth1	EndDay1	EndYear1	
	StartMonth2	StartDay2	StartYear2	EndMonth2	EndDay2	EndYear2	
	TransFrom	TransTo	Initiator	Outcome	WhereFought		
	SideADeaths	SideBDeaths	Version				
Extra-State War	WarNum	WarName	WarType	Interven			
	ccode1	SideA	ccode2	SideB			
	StartMonth1	StartDay1	StartYear1	EndMonth1	EndDay1	EndYear1	
	StartMonth2	StartDay2	StartYear2	EndMonth2	EndDay2	EndYear2	
	TransFrom	TransTo	Initiator	Outcome	WhereFought		
	BatDeath	NonStateDeaths	Version				
Non-State War	WarNum	WarName	WarType				
	SideA1	SideA2	SideB1	SideB2	SideB3	SideB4	SideB5
	StartMonth	StartDay	StartYear	EndMonth	EndDay	EndYear	
	TransFrom	TransTo	Initiator	Outcome			
	SideADeaths	SideBDeaths	TotalCombatDeaths	Version			
War List	War Name	Start Year	War Type & Number				



EER Diagram: The War Tables

War Tables in the Relational Model



WAR

WarID	WarShortName	WarLongName	WarType	IsIntervention	IsInternational
			WarTypeName		

WAR_TRANSITIONS

FromWar	ToWar

WAR_LOCATIONS

WarID	Region

WAR_PARTICIPANTS

WarID	PolityID	StartDate	EndDate	Side	IsInitiator	Outcome	Deaths
		StartYear	EndYear				
		StartMonth	EndMonth				
		StartDay	EndDay				

POLITY

PolityID	PolityName	PolityType	StateAbbr

- ✓ Variables consistent across all four war types
- ✓ Flexible format allows for any # of polities, start/end dates
- ✓ Functional dependencies and primary keys are clear