Databases Autumn 2019

Data Analysis Project
P1: Schema integration

The relation between event headlines, tourism and wealth in Europe.

January 11, 2020

Nikolai Rutz, Luc Kury

1   Introduction

In this part of the project, we analyse the structure of the data and identify the relevant entities and their relations. Further we combine the ER diagrams and connect them with the alleged relations for which we want to find a correlation. Lastly we deduct a relational schema for the data.

2   Entity relationship diagrams

This section contains the ER diagrams of the single data sources representing their as-is state in the following order: GDELT, Tourism and Income. The GDELT dataset has a few functions and specifications which cannot be represented in the ER diagram. These are:

- One of the actors can be empty. For some events the `Actor1` or `Actor2` attribute can be a blank entry for complex or single-actor situations. `Actor1` usually represents the acting party and `Actor2` represents the affected party.

- The `EventGeo` for actors can can be blank, if a headline did not specify the actor country affiliations.
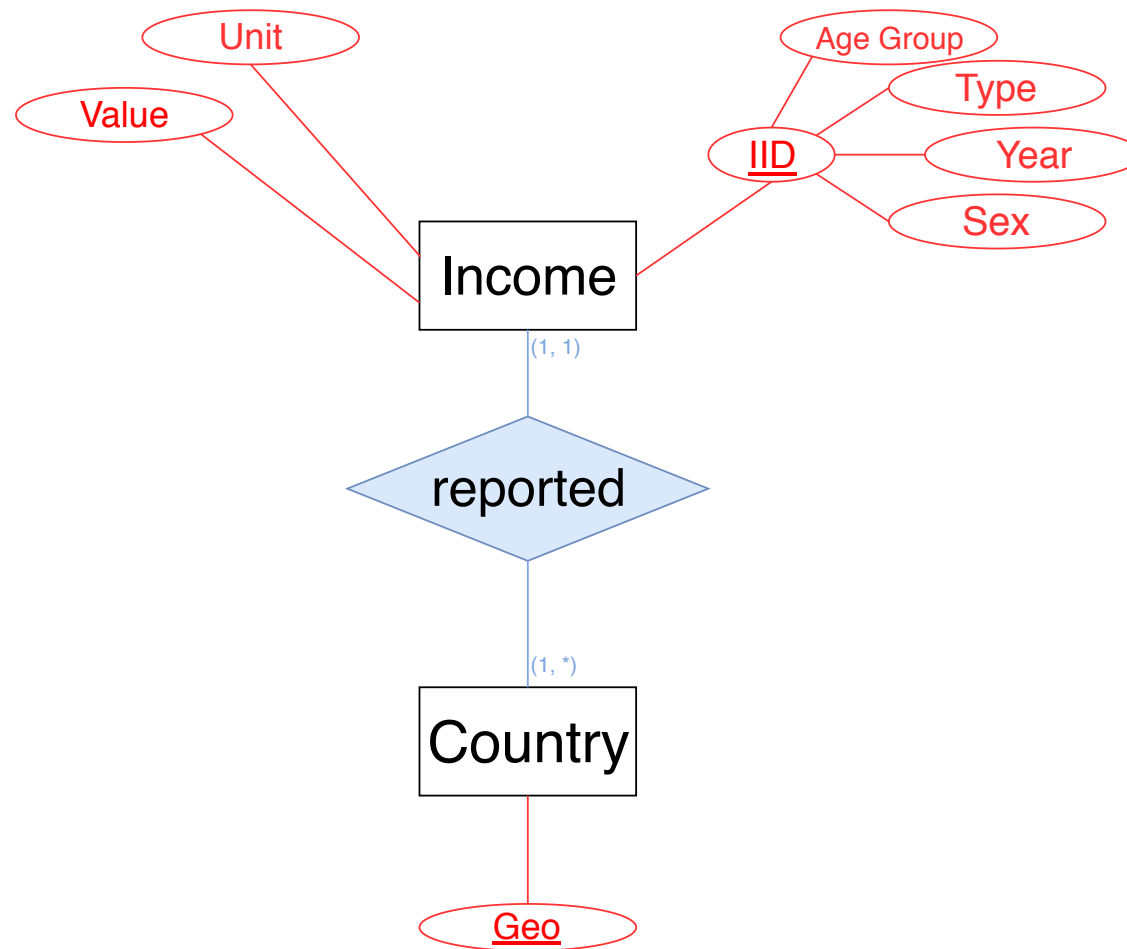
- The `DATEADDED` attribute of the `DATA MANAGEMENT FIELDS` did not exist prior to 2013. This does not affect us, since we do not rely on this attribute to exist.
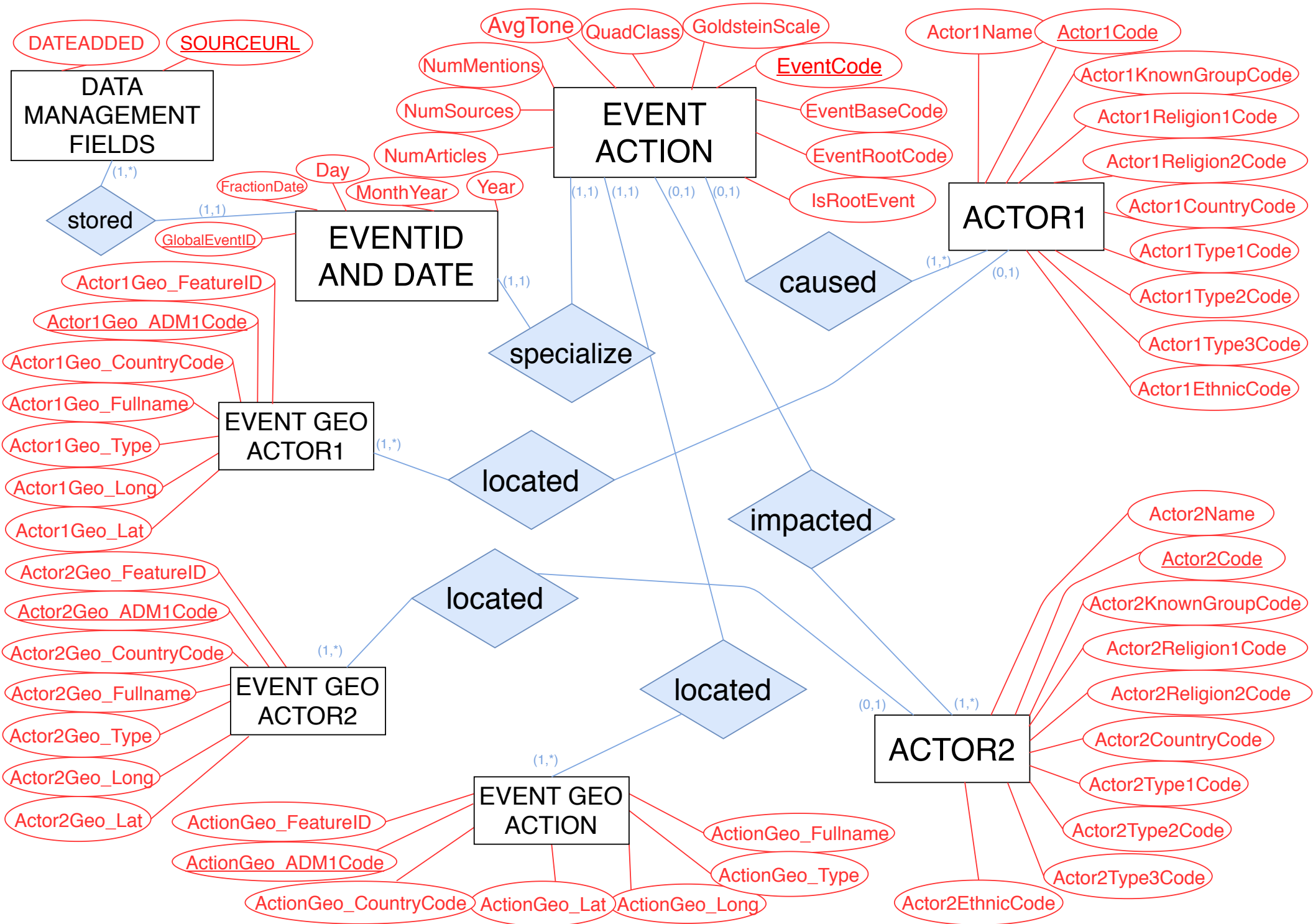
EuroStat - Tourism data
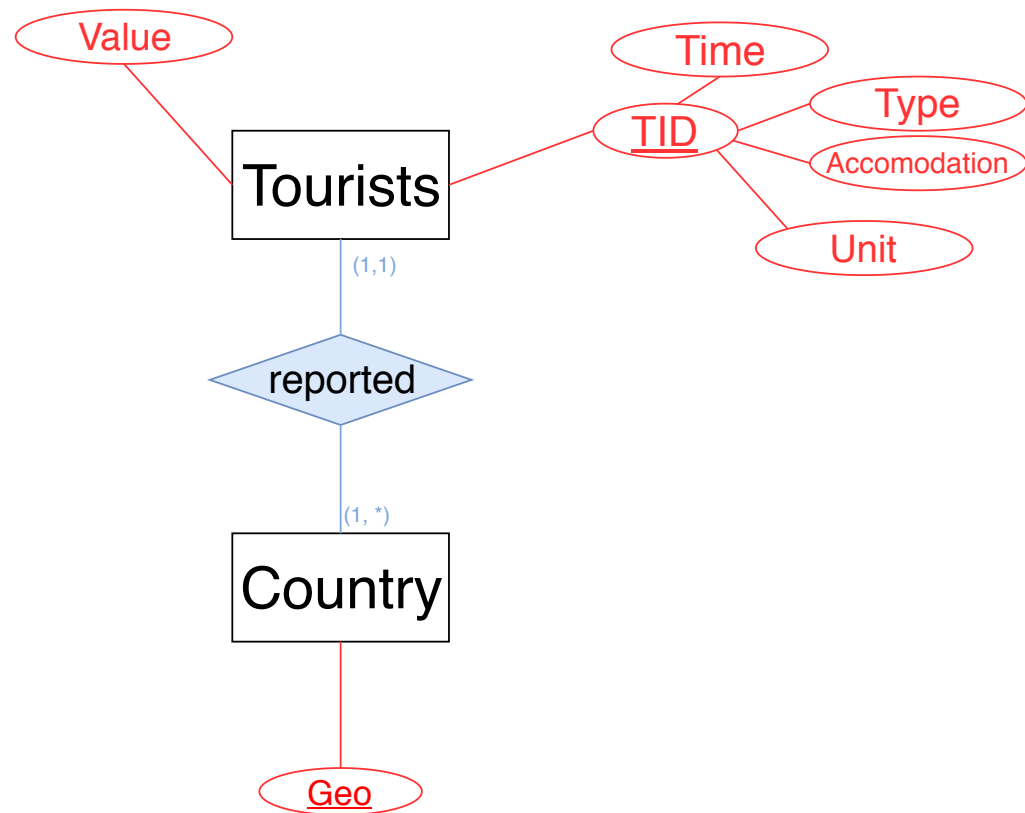
EuroStat - Wealth data

3 Integrated schema

This section is concerned with the ER model of the integrated schema. On the next page you can see the ER diagram of all three data sources combined. From the GDELT data, only the relevant attributes and relations were duplicated, due to space contraints. For the full GDELT schema, please refer to section 2.

Integrated ERM
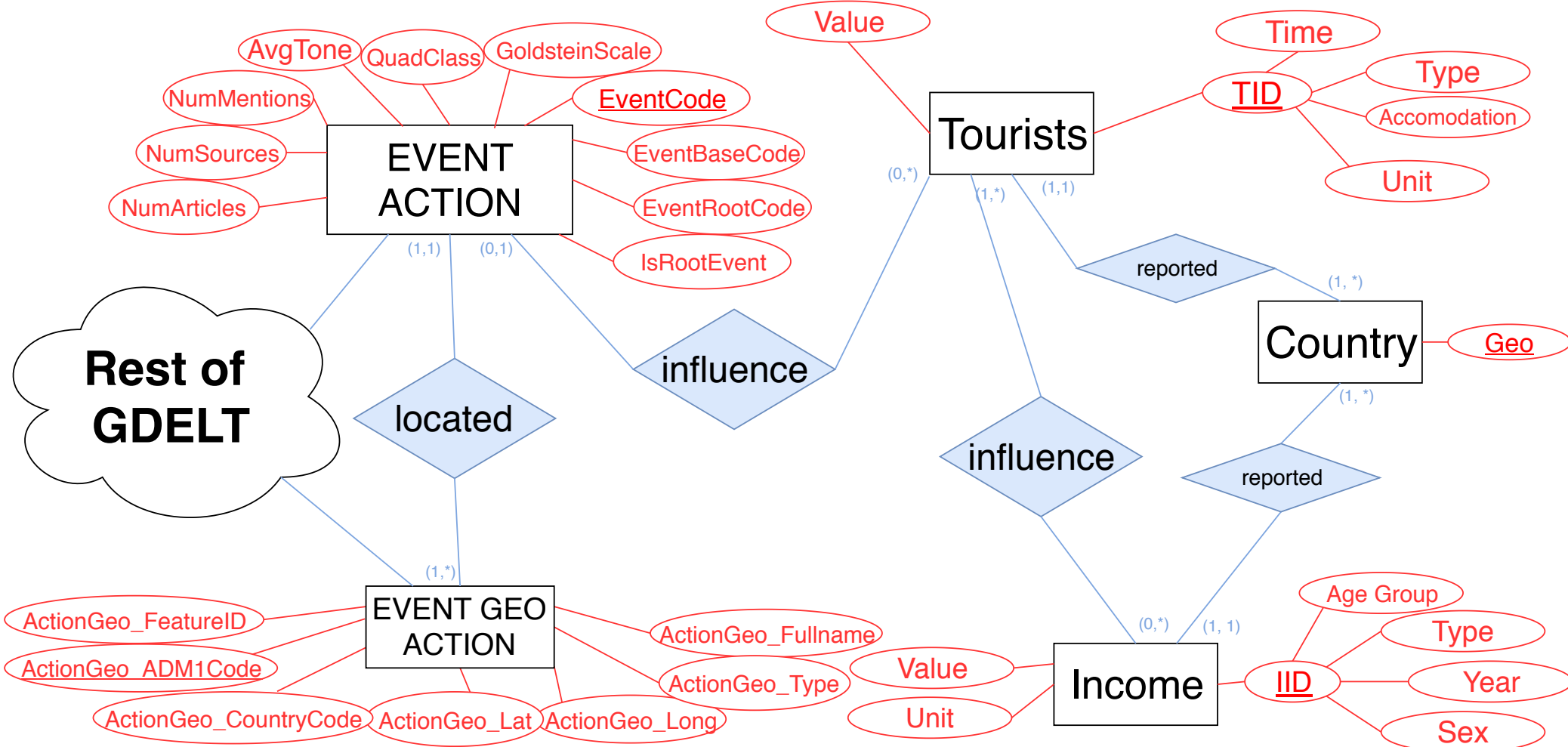
4 Relational schema

In this section we deduct the relational schema for the integrated model. We merged all the relations where possible and defined the necessary primary and foreign keys.

- `EVENTID AND DATE`(<u>GlobalEventID</u>, <u>SOURCEURL</u>, <u>EventCode</u>, FractionDate, Day, MonthYear, Year)

- `DATA MANAGEMANT FIELDS`(<u>SOURCEURL</u>, DATEADDED)

- `ACTOR1`(<u>Actor1Code</u>, <u>Actor1Geo_ADM1Code</u>, Actor1Name, Actor1KnownGroupCode, Actor1Religion1Code, Actor1Religion2Code, Actor1CountryCode, Actor1Type1Code, Actor1Type2Code, Actor1Type3Code, Actor1EthnicCode)

- `ACTOR2`(<u>Actor2Code</u>, <u>Actor2Geo_ADM1Code</u>, Actor2Name, Actor2KnownGroupCode, Actor2Religion1Code, Actor2Religion2Code, Actor2CountryCode, Actor2Type1Code, Actor2Type2Code, Actor2Type3Code, Actor2EthnicCode)

- `EVENT ACTION`(<u>EventCode</u>, <u>Actor1Code</u>, <u>Actor2Code</u>, <u>ActionGeo_ADM1Code</u>, <u>TID</u>, EventBaseCode, EventRootCode, IsRootEvent, GoldsteinScale, QuadClass, AvgTone, NumMentions, NumSources, NumArticles)

- `EVENT GEO ACTION`(<u>ActionGeo_ADM1Code</u>, ActionGeo_FeatureID, ActionGeo_CountryCode, ActionGeo_Lat, ActionGeo_Long, ActionGeo_Type, ActionGeo_FullName)

- `EVENT GEO ACTOR1`(<u>Actor1Geo_ADM1Code</u>, Actor1Geo_FeatureID, Actor1Geo_CountryCode, Actor1Geo_Lat, Actor1Geo_Long, Actor1Geo_Type, Actor1Geo_FullName)

- `EVENT GEO Actor2`(<u>Actor2Geo_ADM1Code</u>, Actor2Geo_FeatureID, Actor2Geo_CountryCode, Actor2Geo_Lat, Actor2Geo_Long, Actor2Geo_Type, Actor2Geo_FullName)

- `Income`(<u>IID</u>, <u>Geo</u>, Value, Unit, IID.AgeGroup, IID.Type, IID.Year, IID.Sex)

- `Tourist`(<u>TID</u>, <u>Geo</u>, Value, TID.Time, TID.Type, TID.Accomodation, TID.Unit)

- `Country`(<u>Geo</u>)

- `influence`(<u>TID</u>, <u>IID</u>)