Linked Open Social Data for Scientific Benchmarking (Supporting Information document)

Renato Fabbri[a,1,*], Osvaldo Novais de Oliveira Junior[a,1]

[a] *São Carlos Institute of Physics, São Paulo University, Brazil*

Abstract

This is a Supporting Information document which exposes ontological diagrams and auxiliary tables for the Linked Open Social Data (LOSD) database. The main document of the article is in [1].

Keywords: Big Data, Data Mining, Benchmark Data, Facebook, Twitter, IRC, Email, Complex Networks, Text Mining

Contents

*Corresponding author
Email addresses: fabbri@usp.br (Renato Fabbri), chu@ifsc.usp.br (Osvaldo Novais de Oliveira Junior)
[1] *URL:* http://www.ifsc.usp.br/

1. General guidance

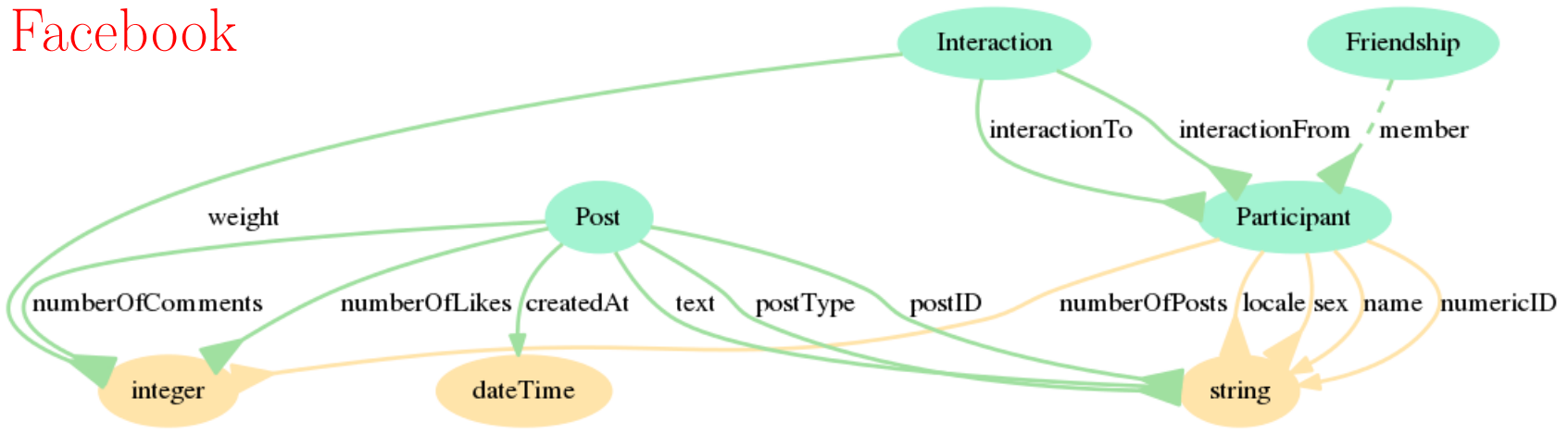
In this document we provide diagrams for the provenances in the LOSD: Facebook, Twitter, IRC, Email, ParticipaBR, Cidade Democrática and AA. Each provenance diagram was broken in two, one presents the relations among main classes (blue nodes) and data types (orange nodes), the other presents metadata on the snapshot. Every class instance is related to the snapshot instance by the triple `class_uri po:snapshot snapshot_uri`. Such triples are omitted for simplicity. Due to the large number of relations, the rendering of diagrams are automatized and displays some overlaps. Even so, the images are useful for grasping what is in current LOSD and for conducting explorations. Edges in the diagrams have:

- green color if representing an OWL existential class restriction (all individuals from the class present at least one triple with the property as predicate);
- inverted nip if representing an OWL universal class restriction (all individuals presenting triples with the property as predicate are from the class);
- full edges (non-dashed) if representing a functional property axiom (there is at most one triple with the property as the predicate for each individual).

Furthermore, this document ends with two sets of tables, one with counts of triples, participants, edges/interactions/relations and characters, the other with references for snapshot groups, such as wikipedia or contact links.

2. Facebook data

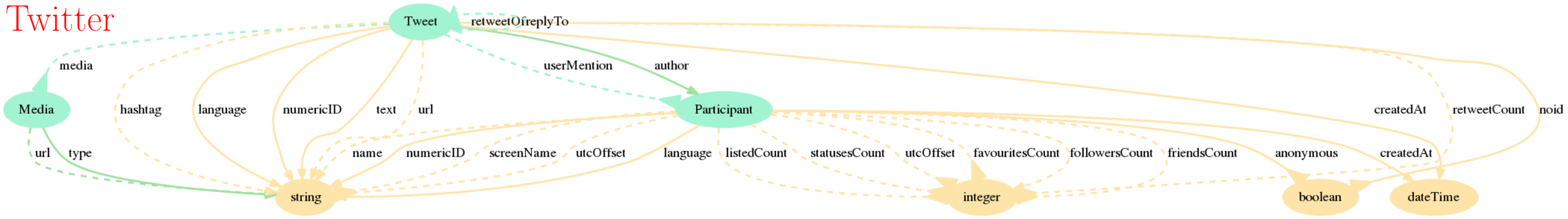
Each Facebook snapshot is yield by either an user, from which the friends constitute a friendship network, or a group, which participants can yield friendship and interaction networks and posts information with text and some metadata. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article [1].

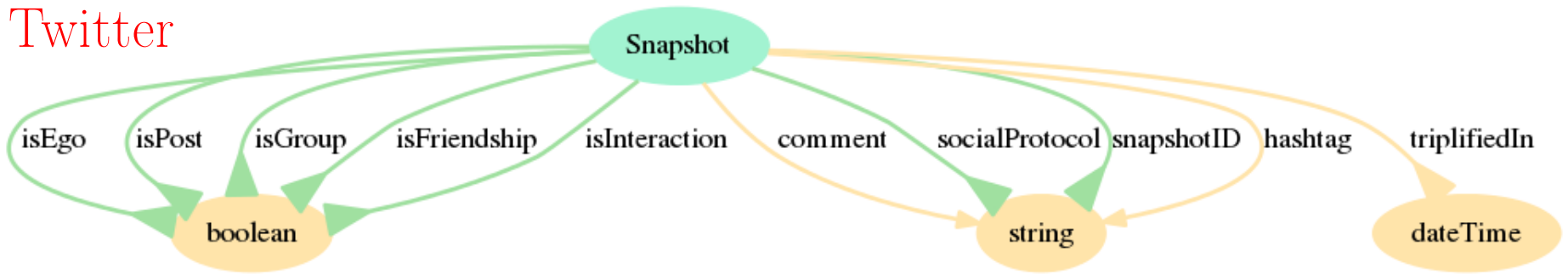

Facebook
Interaction
Friendship
interactionTo
interactionFrom
member
weight
Post
Participant
numberOfComments
numberOfLikes
createdAt
text
postType
postID
numberOfPosts
locale
sex
name
numericID
integer
dateTime
string


Facebook
Snapshot
isFriendship
friendshipsAnonymized
isInteraction
isGroup
interactionsAnonymized
isEgo
isPost
numericID
snapshotID
acquiredThrough
comment
socialProtocol
name
triplifiedIn
dateObtained
boolean
string
dateTime
date

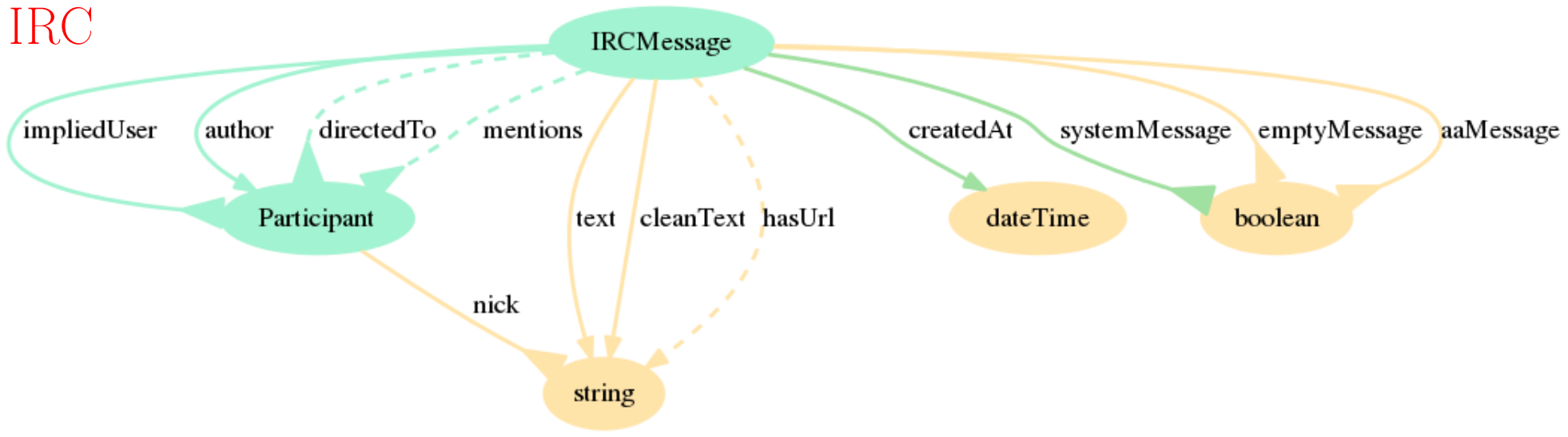
3. Twitter data

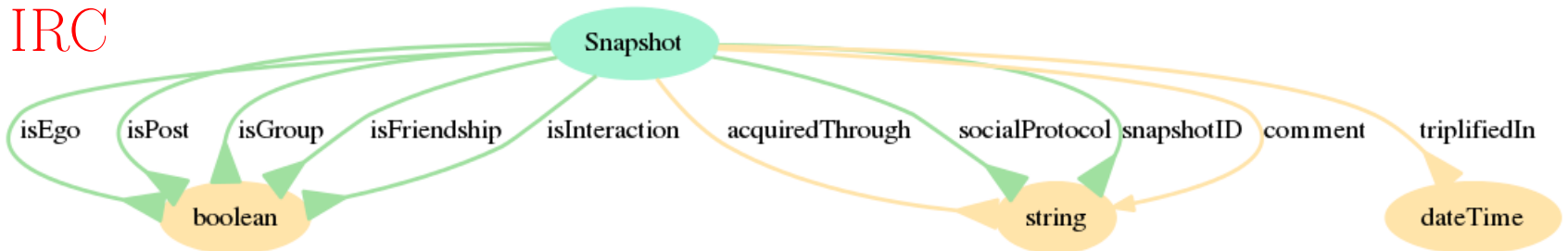
Each Twitter snapshot is yield by a hashtag. Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article [1].


Twitter
Tweet
retweetOfreplyTo
media
userMention
author
Media
hashtag
language
numericID
text
url
Participant
createdAt
retweetCount
noid
url
type
name
numericID
screenName
utcOffset
language
listedCount
statusesCount
utcOffset
favouritesCount
followersCount
friendsCount
anonymous
createdAt
string
integer
boolean
dateTime

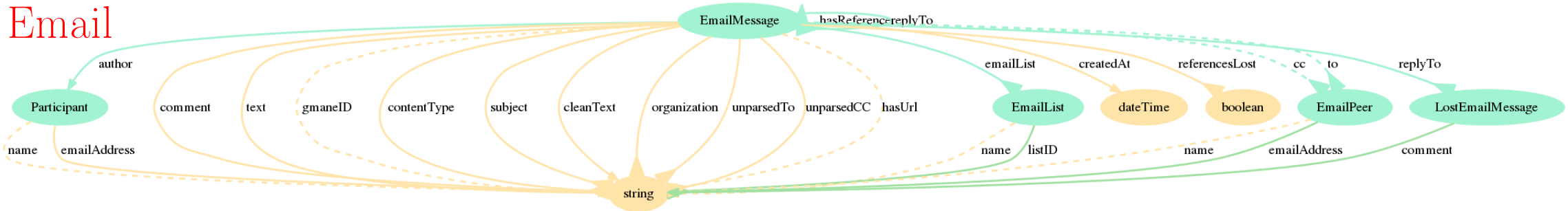

Twitter
Snapshot
isEgo
isPost
isGroup
isFriendship
isInteraction
comment
socialProtocol
snapshotID
hashtag
triplifiedIn
boolean
string
dateTime

4. IRC data

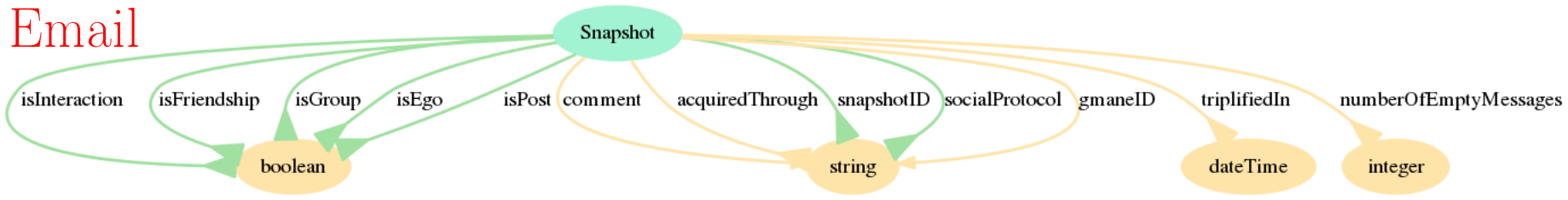

IRC
IRCMessage
impliedUser
author
directedTo
mentions
createdAt
systemMessage
emptyMessage
aaMessage
Participant
text
cleanText
hasUrl
dateTime
boolean
nick
string


IRC
Snapshot
isEgo
isPost
isGroup
isFriendship
isInteraction
acquiredThrough
socialProtocol
snapshotID
comment
triplifiedIn
boolean
string
dateTime

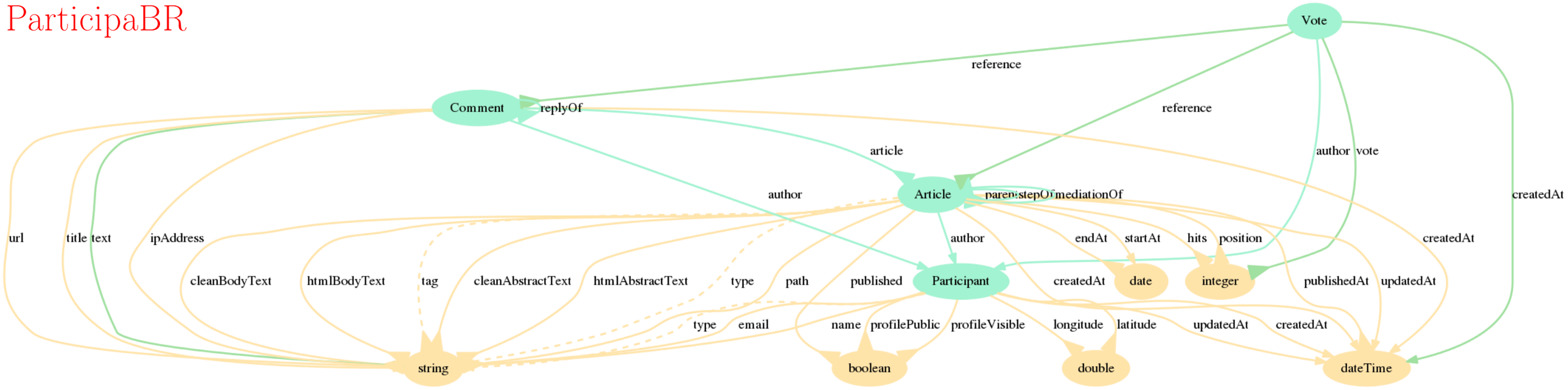
5. Email data

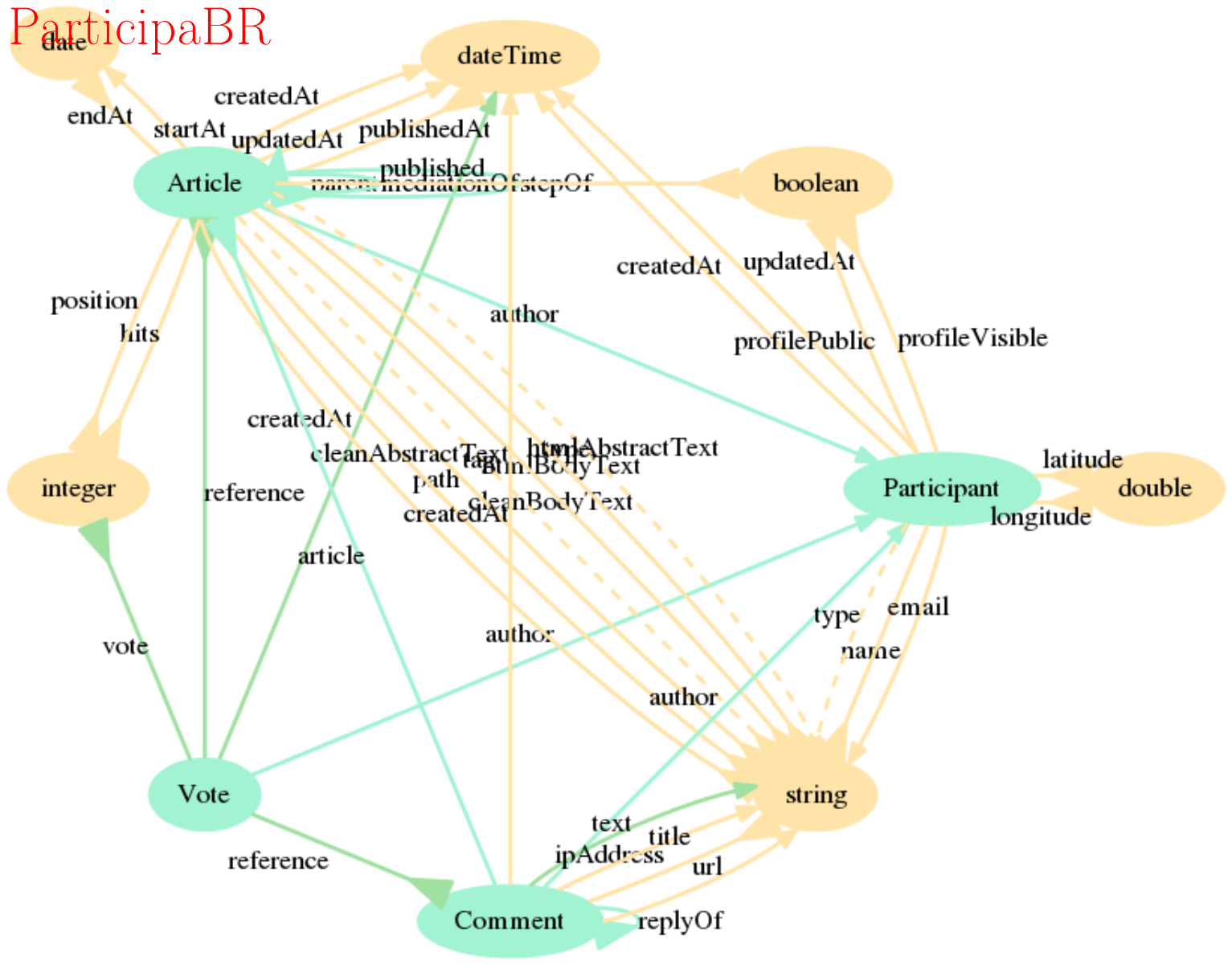

Email
EmailMessage
hasReference
replyTo
author
Participant
comment
text
gmaneID
contentType
subject
cleanText
organization
unparsedTo
unparsedCC
hasUrl
emailList
createdAt
referencesLost
cc
to
replyTo
EmailList
dateTime
boolean
EmailPeer
LostEmailMessage
name
emailAddress
name
listID
name
emailAddress
comment
string

Email

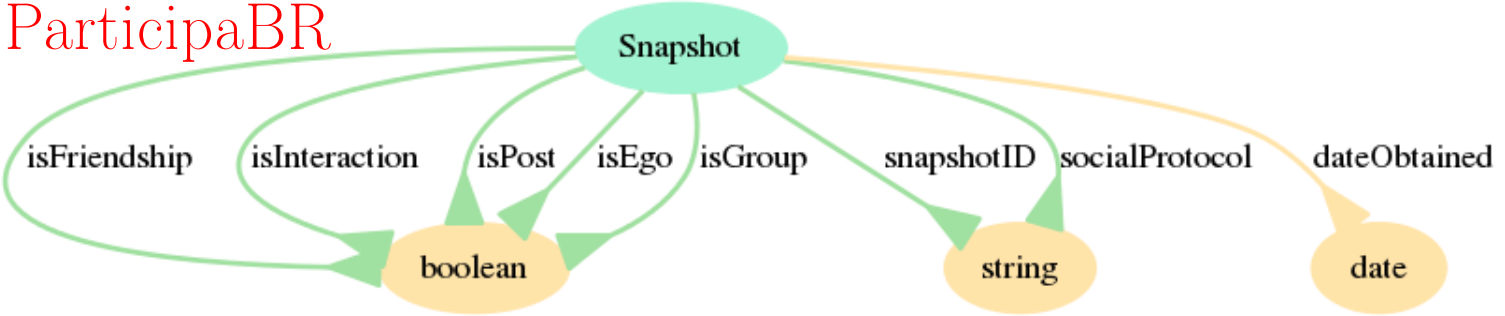

Snapshot
isInteraction
isFriendship
isGroup
isEgo
isPost
comment
acquiredThrough
snapshotID
socialProtocol
gmaneID
triplifiedIn
numberOfEmptyMessages
boolean
string
dateTime
integer

6. ParticipaBR data

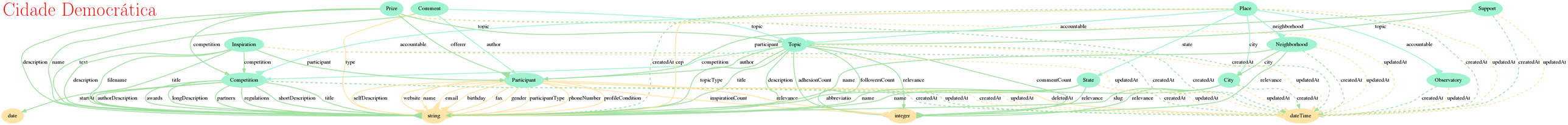

ParticipaBR
Vote
reference
Comment
replyOf
reference
article
author
vote
createdAt
author
Article
parentStepOf
mediationOf
url
title
text
ipAddress
author
endAt
startAt
hits
position
createdAt
cleanBodyText
htmlBodyText
tag
cleanAbstractText
htmlAbstractText
type
path
published
Participant
createdAt
date
integer
publishedAt
updatedAt
type
email
name
profilePublic
profileVisible
longitude
latitude
updatedAt
createdAt
string
boolean
double
dateTime


ParticipaBR
date
dateTime
createdAt
endAt
startAt
updatedAt
publishedAt
published
Article
boolean
createdAt
updatedAt
position
hits
author
profilePublic
profileVisible
createdAt
cleanAbstractText
htmlAbstractText
integer
reference
path
cleanBodyText
createdAt
Participant
latitude
double
longitude
article
type
email
vote
author
name
author
Vote
string
text
title
ipAddress
url
reference
Comment
replyOf

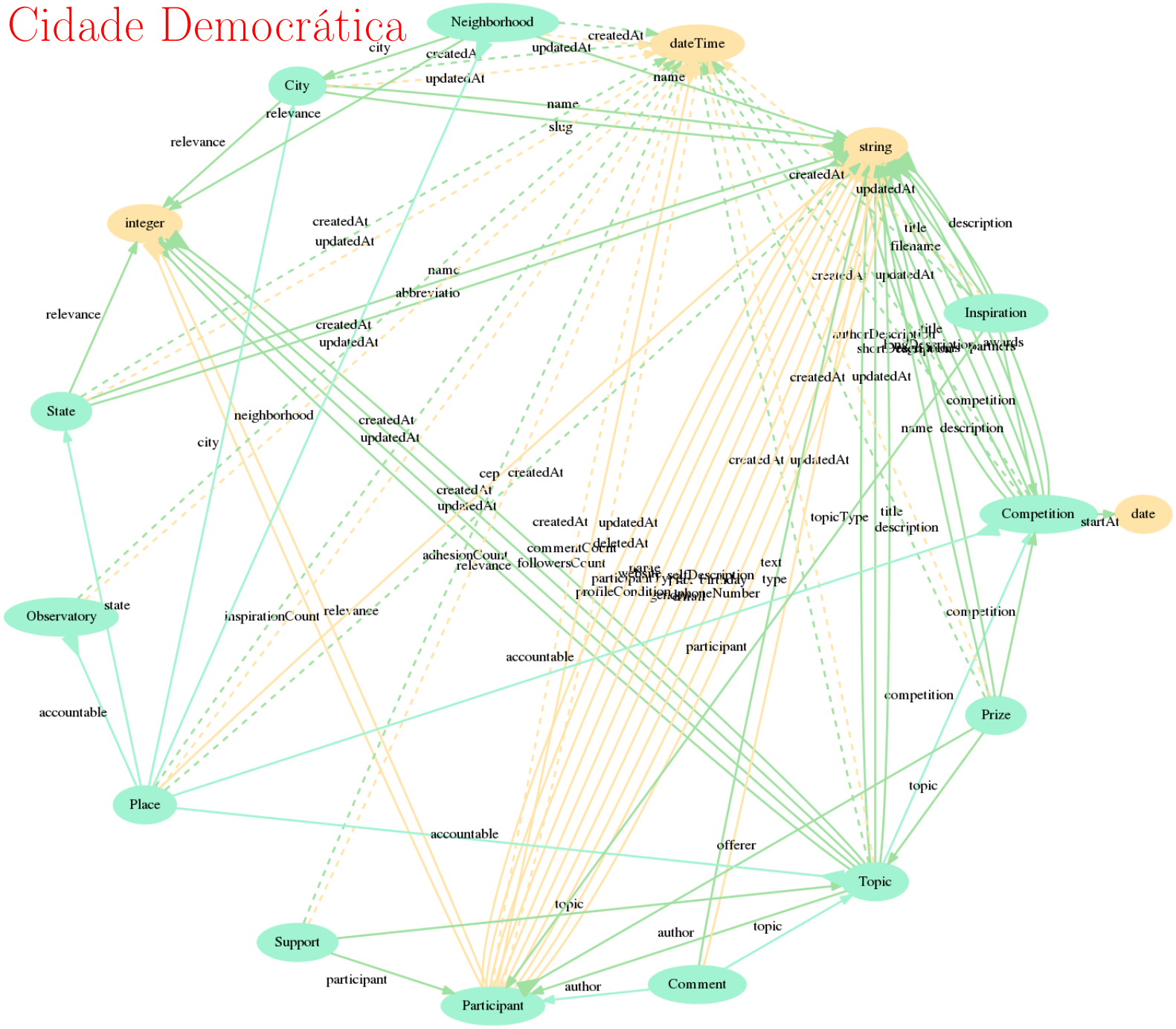
ParticipaBR



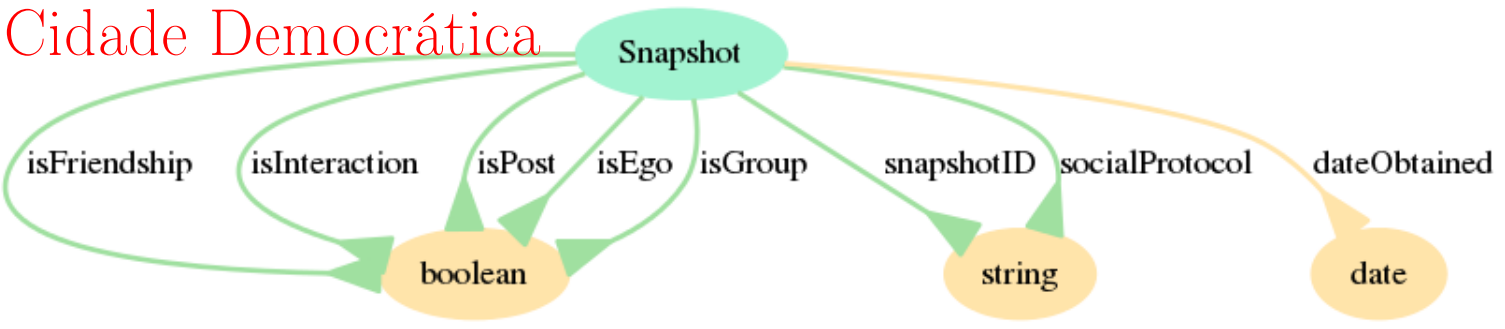
7. Cidade Democrática data


Cidade Democrática
Prize
Comment
Place
Support
Inspiration
Topic
Neighborhood
Competition
Participant
State
City
Observatory
date
string
integer
dateTime
description
name
text
topic
competition
accountable
offerer
author
participant
state
city
neighborhood
filename
title
type
createdAt
cep
topicType
adhesionCount
followersCount
relevance
commentCount
updatedAt
startAt
authorDescription
awards
longDescription
partners
regulations
shortDescription
selfDescription
website
email
birthday
fax
gender
participantType
phoneNumber
profileCondition
inspirationCount
abbreviatio
deletedAt
slug

Cidade Democrática

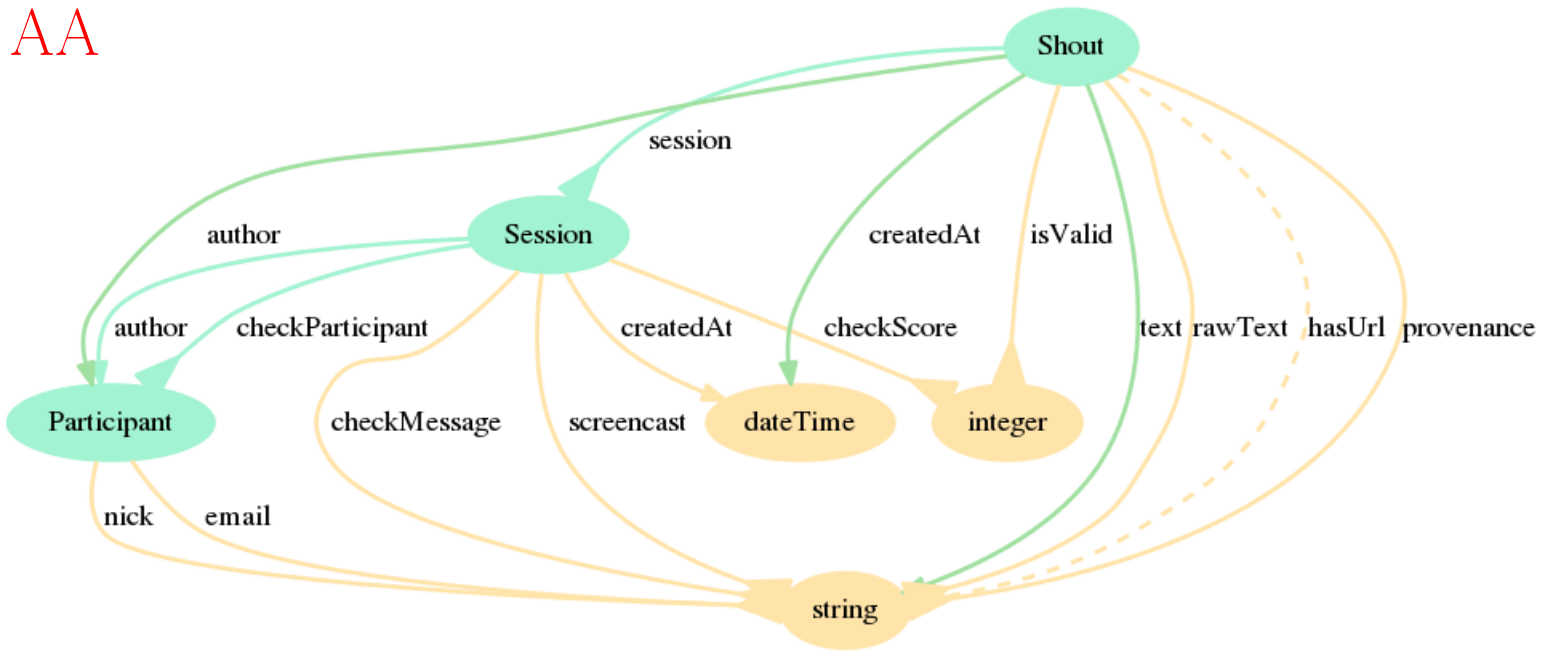


Cidade Democrática

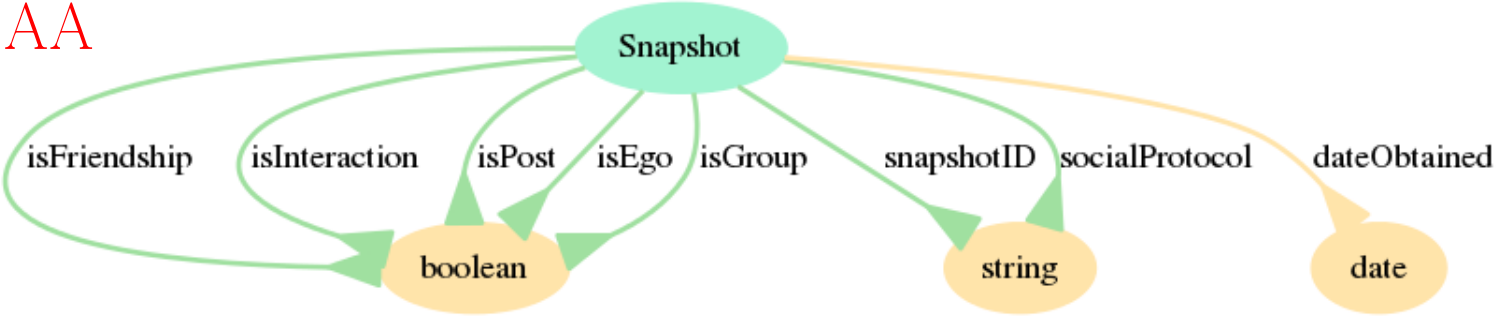


8. AA data

AA


Shout
session
author
Session
createdAt
isValid
author
checkParticipant
createdAt
checkScore
text
rawText
hasUrl
provenance
Participant
checkMessage
screencast
dateTime
integer
nick
email
string

AA


Snapshot
isFriendship
isInteraction
isPost
isEgo
isGroup
snapshotID
socialProtocol
dateObtained
boolean
string
date

9. Snapshot references

10. Elementary counting in snapshots

References

[1] O. N. d. O. J. Renato Fabbri, Linked open social data for scientific benchmarking, https://github.com/ttm/linkedOpenSocialData/raw/master/paper.pdf (2016).