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

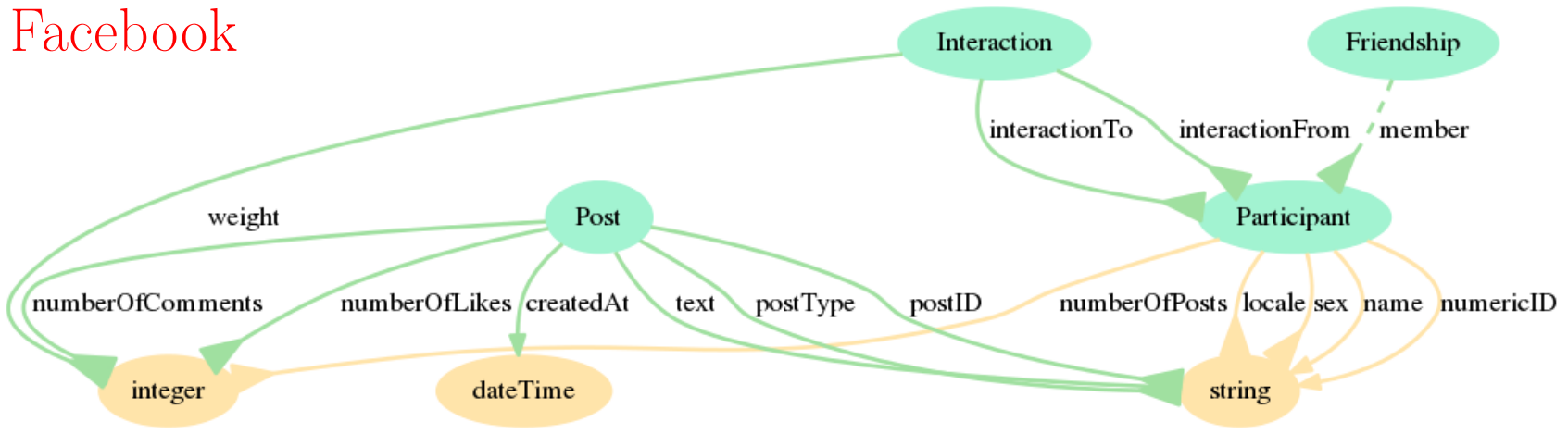
In this document we provide diagrams for the provenances in the LOSD: Facebook, Twitter, IRC, Email, ParticipaBR, Cidade Democrática and AA. Each provenance diagram was broken in two, one presents the relations among main classes (blue nodes) and data types (orange nodes), the other presents metadata 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

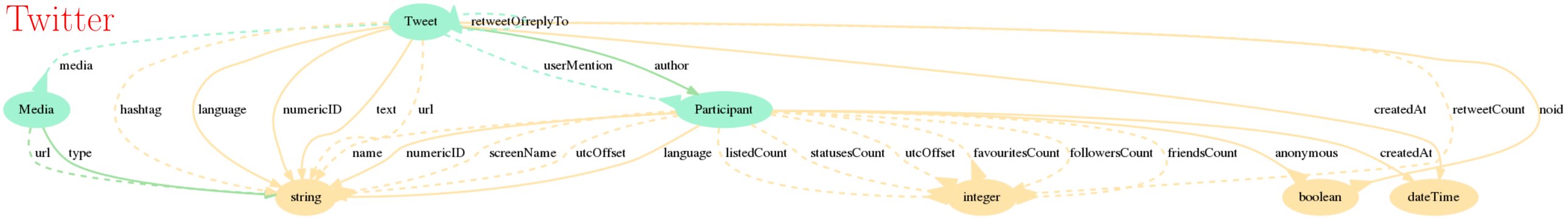
Each Facebook snapshot is yield by either an user, from which the friends constitute a friendship network, or a group, which participants can yield friendship and interaction networks and posts information with text and some metadata. Further information is found on the following diagrams, the tables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

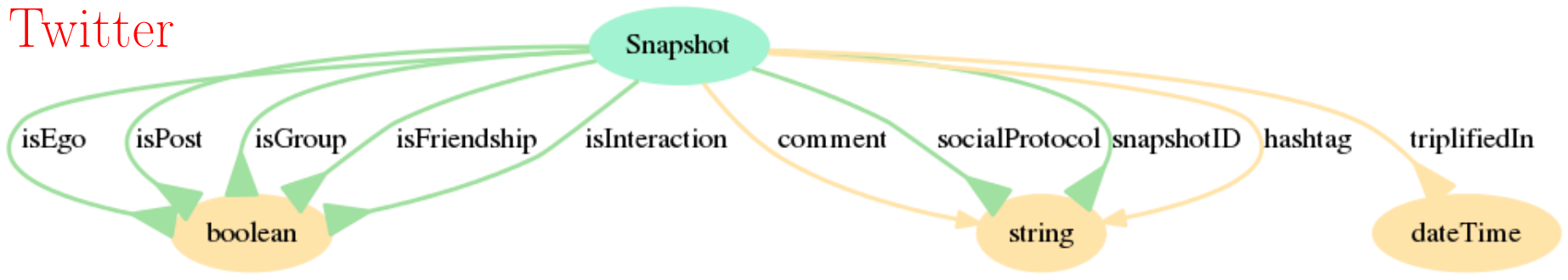

Facebook
Interaction
Friendship
interactionTo
interactionFrom
member
weight
Post
Participant
numberOfComments
numberOfLikes
createdAt
text
postType
postID
numberOfPosts
locale
sex
name
numericID
integer
dateTime
string


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

3. Twitter data

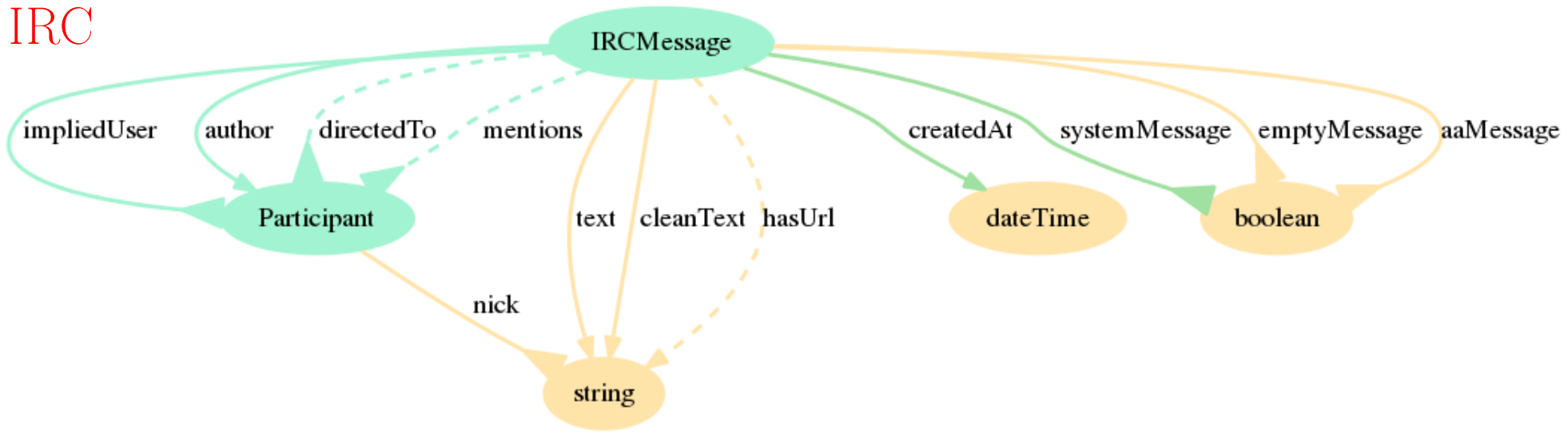
Each Twitter snapshot is yield by a hashtag. Retweets (`po:retweetOf` are usually considered to yield the interactions between users. Users are identified through authors `po:numericID` (global as given by Twitter API) or . The database present also `po:replyTo` and `po:userMention` which might also be useful in understanding the networking. 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].

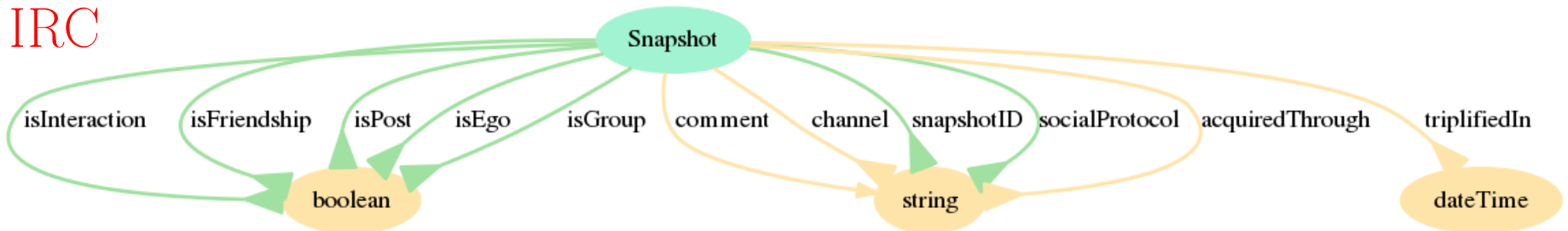

Twitter
Tweet
retweetOfreplyTo
media
userMention
author
Media
hashtag
language
numericID
text
url
Participant
createdAt
retweetCount
noid
url
type
name
numericID
screenName
utcOffset
language
listedCount
statusesCount
utcOffset
favouritesCount
followersCount
friendsCount
anonymous
createdAt
string
integer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


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

4. IRC data

Each IRC snapshot is yield by an IRC channel. IRC messages are either server messages (e.g. join and exit channel) marked with `po:systemMessage true` and having an `po:impliedUser user_uri`, or user messages, which yield interactions through `po:directedTo` and `po:mentions` properties. Text messages without the user names are delivered through the `po:cleanText` property. 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].

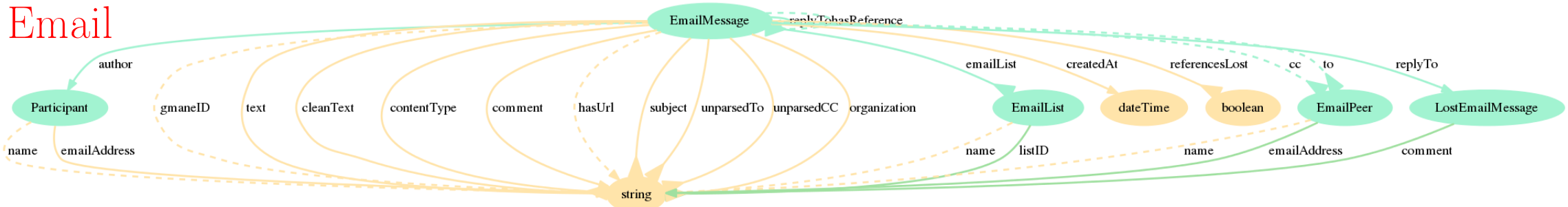

IRC
IRCMessage
impliedUser
author
directedTo
mentions
createdAt
systemMessage
emptyMessage
aaMessage
Participant
text
cleanText
hasUrl
dateTime
boolean
nick
string


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

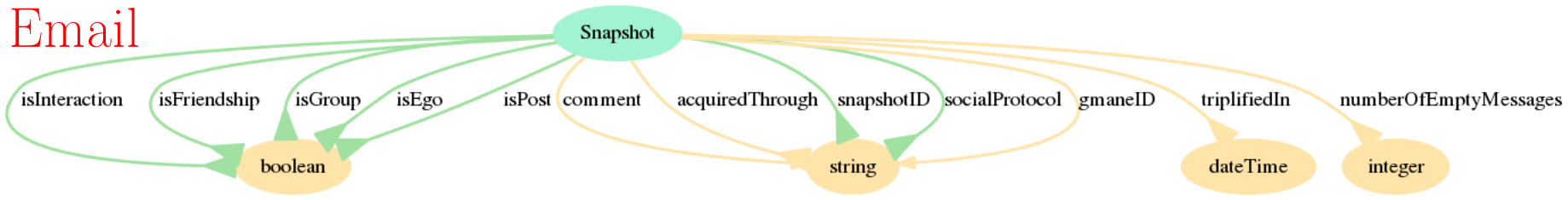
5. Email data

Each IRC snapshot is yield by an Email list. Interactions is yield through `po:replyTo` relations although `po:to` and `po:cc` can also be considered. the email body is given by `po:text` relations while `po:cleanText` has the text with lines removed where they are trivially from previous messages or computer code. 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].

Email

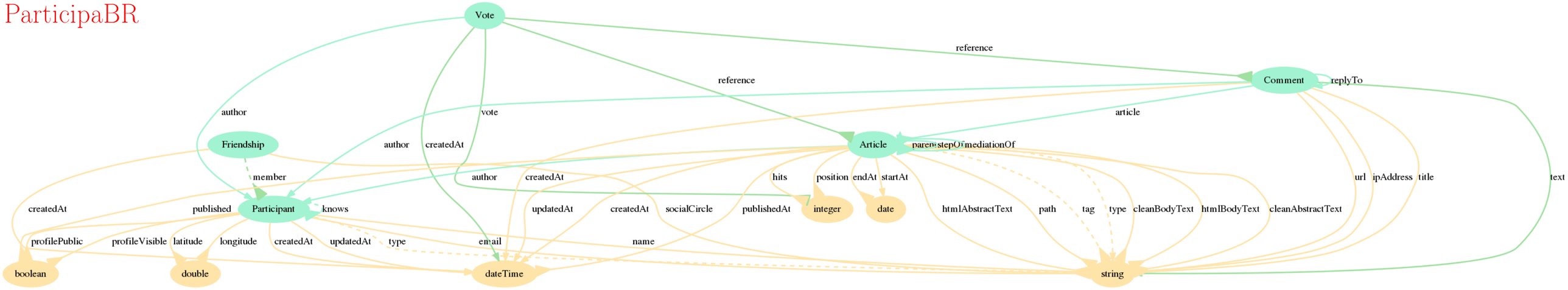

EmailMessage
replyToInReference
author
Participant
gmaneID
text
cleanText
contentType
comment
hasUrl
subject
unparsedTo
unparsedCC
organization
emailList
createdAt
referencesLost
cc
to
replyTo
EmailList
dateTime
boolean
EmailPeer
LostEmailMessage
name
emailAddress
name
listID
name
emailAddress
comment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

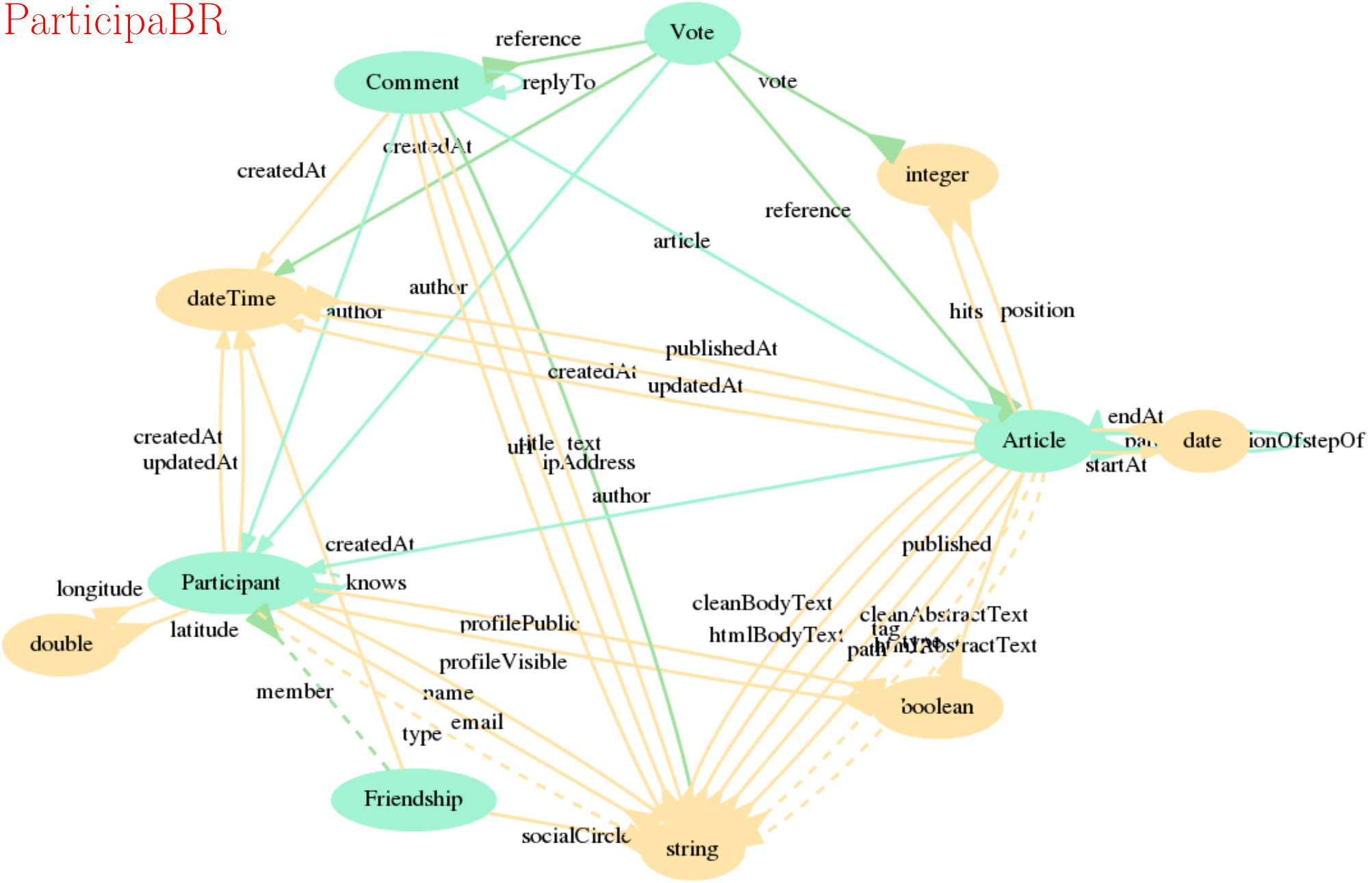
Email


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

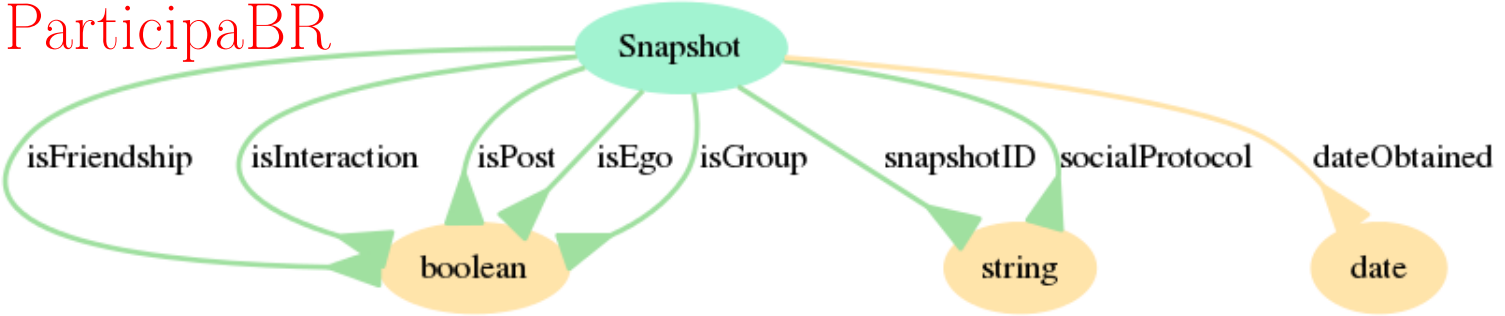
6. ParticipaBR data

The ParticipaBR snapshot is yield by a data dump donated by the system administrators of the federal portal of social participation ParticipaBR. Articles can have parent articles (`po:parent`), be step of a collection of articles (`po:stepOf`) and be a mediation of other articles (`po:mediationOf`). Interactions are yield by comments which are `po:replyTo` other comments or which are made directly to an article. This snapshot holds also friendship structures. The language used is mainly Brazilian Portuguese, but English and Spanish are also incident. Due to the higher complexity of the diagram, an additional figure is given rendered with another layout algorithm 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].


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

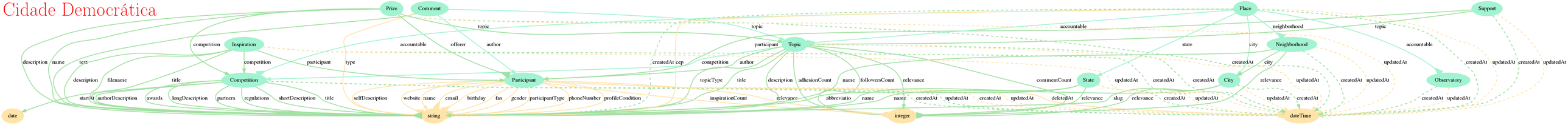

ParticipaBR
Vote
reference
Comment
replyTo
vote
createdAt
createdAt
integer
reference
article
dateTime
author
author
hits
position
publishedAt
createdAt
updatedAt
endAt
createdAt
updatedAt
title
text
ipAddress
Article
date
startAt
author
createdAt
published
longitude
Participant
knows
cleanBodyText
cleanAbstractText
double
latitude
profilePublic
htmlBodyText
profileVisible
member
name
email
type
boolean
Friendship
string

ParticipaBR

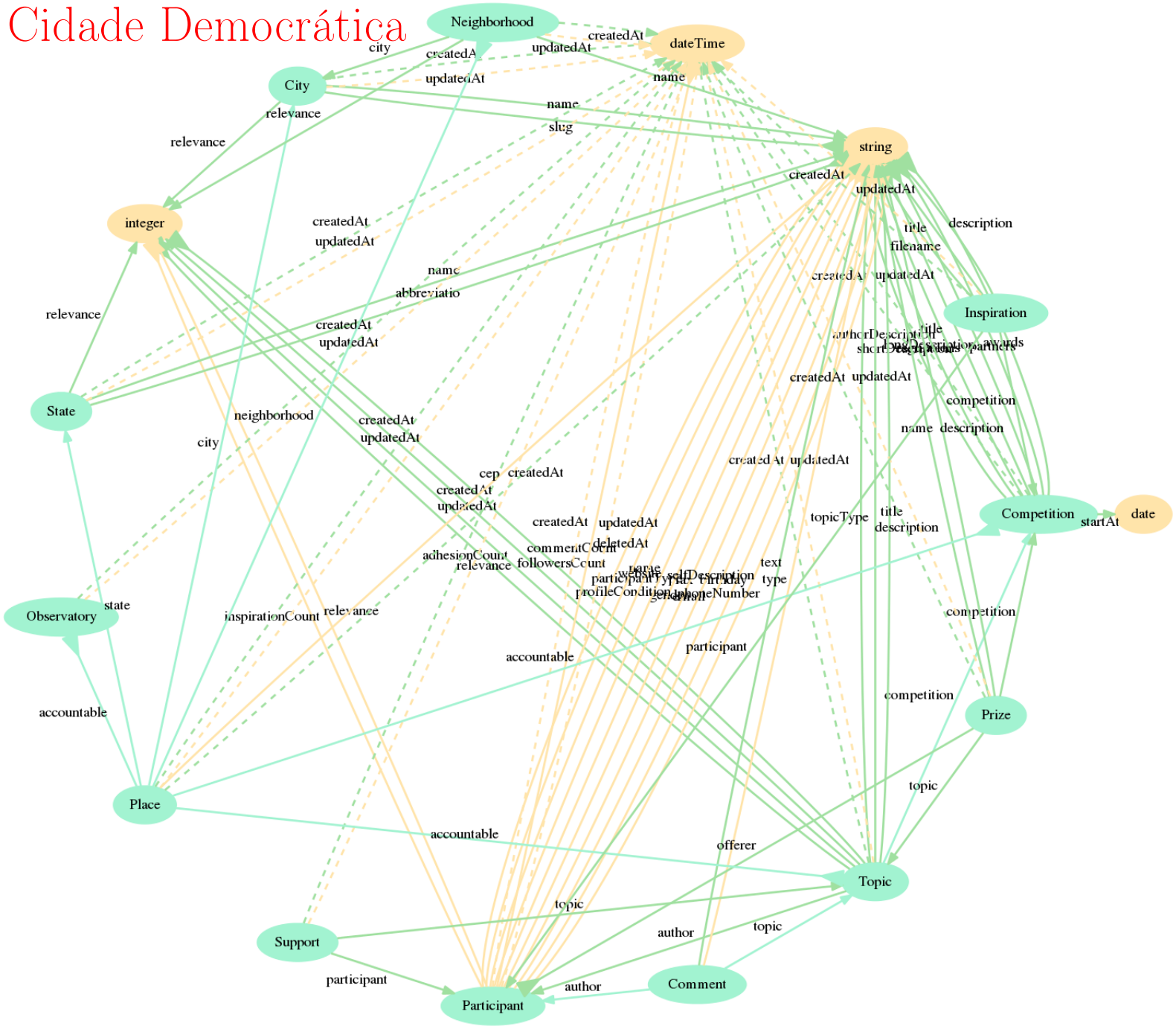


7. Cidade Democrática data

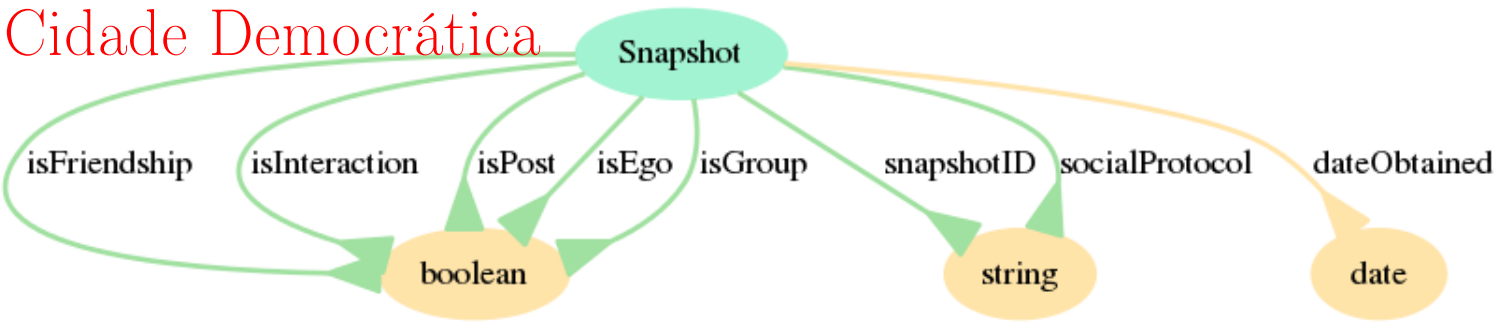
The Cidade Democrática snapshot is yield by a data dump donated by the system administrators of the civil society social participation portal Cidade Democrática. This snapshot holds a complex structure of both Topics/Inspirations/Observatories/Supports/Competitions/Prizes and of State/City/Neighborhood/Place. The language used is mainly Brazilian Portuguese. Due to the higher complexity of the diagram, an additional figure is given rendered with another layout algorithm 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].


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



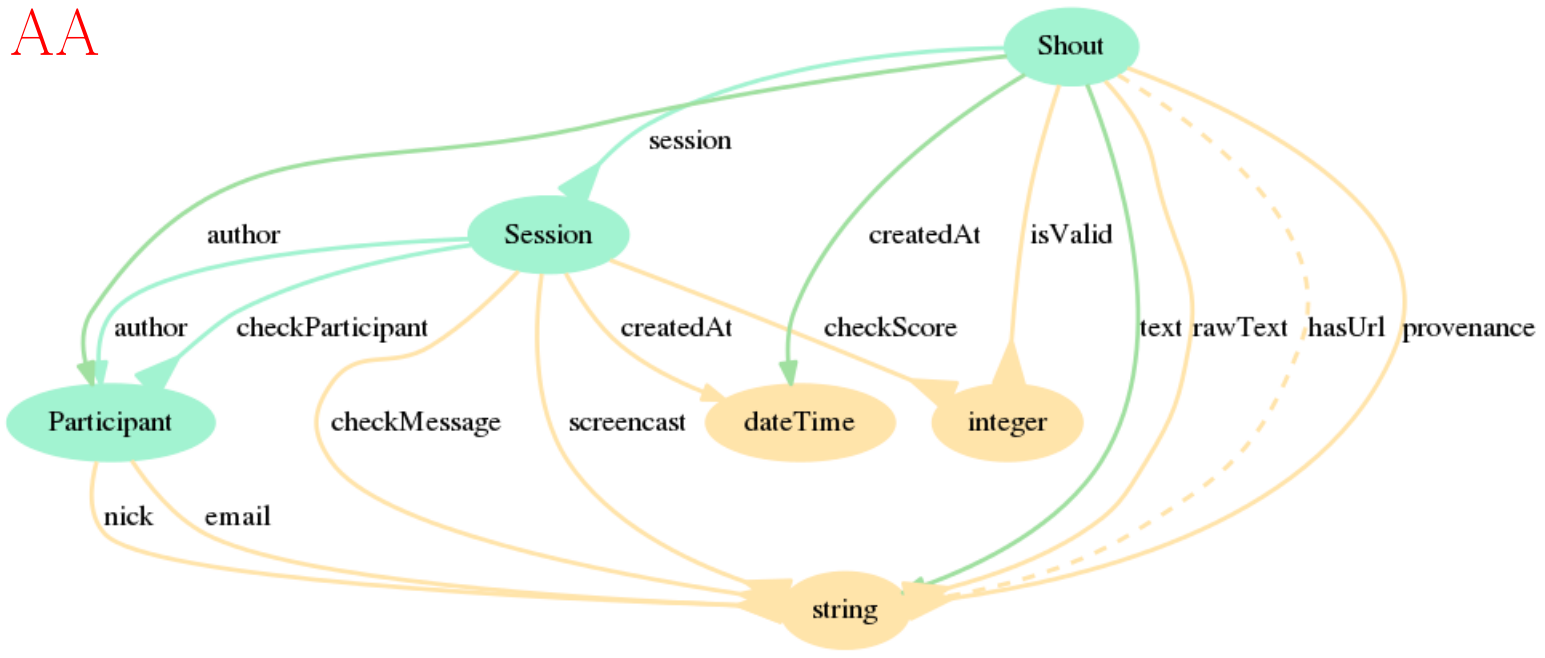
Cidade Democrática



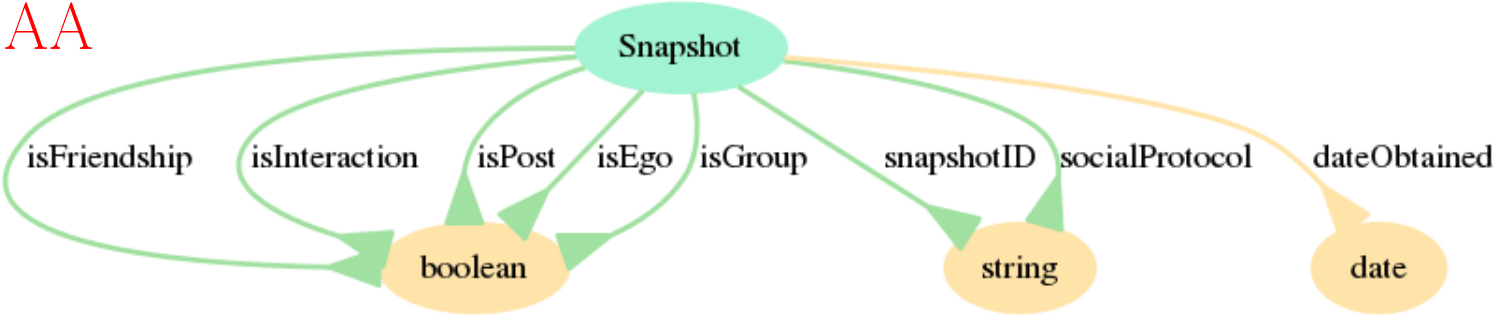
8. AA data

The AA (Algorithmic Autoregulation) snapshots are yield by a data dump donated by the system administrators and by a mined IRC log. The system pursue simplicity and most of data consists of detached shouts with `po:text` and `po:author`. 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].

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

All the Facebook snapshots are either the result of individuals who downloaded their data (and donated to the first author) or data downloaded from groups. In the first case, it is senseless to present references and the. In the second case, we present the group name and a link to a post in the group where data and figures were delivered to the group.

Table 1: Different Twitter snapshots are yield by different hashtags. In this table we present each snapshot with the respective hashtag and a reference to the subject.

snapshot hashtag	**observation**	**reference**
#arenaNETmundial	a Brazilian discussion hub about free culture, democracy and the internet	http://www.participa.br/netmundial
#art	tweets with the generic hashtag #art	https://en.wikipedia.org/wiki/Art
#ChennaiFloods	heavy rainfall generated by the annual northeast monsoon in November–December 2015	https://en.wikipedia.org/wiki/2015_South_Indian_floods
#dilma	the 36th President of Brazil	https://en.wikipedia.org/wiki/Dilma_Rousseff
#ForaDilma	2015-16 anti-government protests in Brazil	https://en.wikipedia.org/wiki/2015-16_protests_in_Brazil
#ForaCunha	2015-16 anti-corruption protests in Brazil	https://en.wikipedia.org/wiki/2015-16_protests_in_Brazil
#fuck	tweets with the generic hashtag #fuck	https://en.wikipedia.org/wiki/Fuck
#game	tweets with the generic hashtag #game	https://en.wikipedia.org/wiki/Game
#god	tweets with the generic hashtag #god	https://en.wikipedia.org/wiki/God
#MAMA2015	the grand 2015 Mnet Asian Music Awards	https://en.wikipedia.org/wiki/2015_Mnet_Asian_Music_Awards
#music	tweets with the generic hashtag #music	https://en.wikipedia.org/wiki/Music
#obama	the 44th President of the United States	https://en.wikipedia.org/wiki/Barack_Obama
#python	the Python programming language	https://en.wikipedia.org/wiki/Python_(programming_language)
#QuartaSemRacismoClubeSDV	an anti-racism netweaving	https://twitter.com/hashtag/quartasemracismoclubesdv
#science	tweets with the generic hashtag #science	https://en.wikipedia.org/wiki/Science
#SnapDetremura	reference for Snapchat about a celebrated person	https://twitter.com/detremura

Table 2: Different IRC snapshots are yield by different channels. In this table we present each snapshot with the respective channel and a reference to the subject.

snapshot channel	**observation**	**reference**
#foradoeixo	a Brazilian network of culture related collectives	https://pt.wikipedia.org/wiki/Fora_do_Eixo
#hackerspace-cps	a hackerspace in Campinas, Brazil	https://lhc.net.br/wiki/P%C3%A1gina_principal
#hackerspaces-br	Brazilian hackerspaces channel	https://garoa.net.br/wiki/Hackerspaces_Brasileiros
#labmacambira	Brazilian channel for the lab-Macambira collective	http://labmacambira.sourceforge.net/

Table 3: Different Email snapshots are yield by different email lists. In this table we present each snapshot with the respective list and a reference to the subject.

Gmane ID	**observation**	**reference**
gmane.linux.audio.users	the Linux Audio Users	http://linuxaudio.org
gmane.politics.organizations.metareciclagem	a network about technology and social transformation	https://metareciclagem.github.io
gmane.linux.audio.devel	the Linux Audio Developers	http://lists.linuxaudio.org/listinfo/linux-audio-dev
gmane.comp.gcc.libstdc++.devel	the C++ standard library	https://gcc.gnu.org/libstdc++/

Table 4: References for the snapshots of the detached instances ParticipaBR, Cidade Democrática and AA.

social protocol	**observations**	**reference**
ParticipaBR	a Brazilian federal portal of social participation	http://www.participa.br/
Cidade Demorática	a Brazilian civil society portal of social participation	http://www.cidadedemocratica.org.br/
AA	the Algorithmic Autoregulation software development methodology	[2]

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).

[2] R. Fabbri, R. Fabbri, V. Vieira, D. Penalva, D. Shiga, M. Mendonça, A. Negrão, L. Zambianchi, G. S. Thumé, The algorithmic autoregulation software development methodology/a metodologia de desenvolvimento de software autorregulação algorítmica, Revista Electronica de Sistemas de Informaçao 13 (2) (2014) 1.