

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

1	General guidance	2
2	Facebook data	2
3	Twitter data	5
4	IRC data	8
5	Email data	11
6	ParticipaBR data	14
7	Cidade Democrática data	18
8	AA data	22
9	Snapshot references	25
10	Elementary counting in snapshots	25

*Corresponding author

Email addresses: fabbri@usp.br (Renato Fabbri), chu@ifsc.usp.br (Osvaldo Novais de Oliveira Junior)

¹URL: <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:

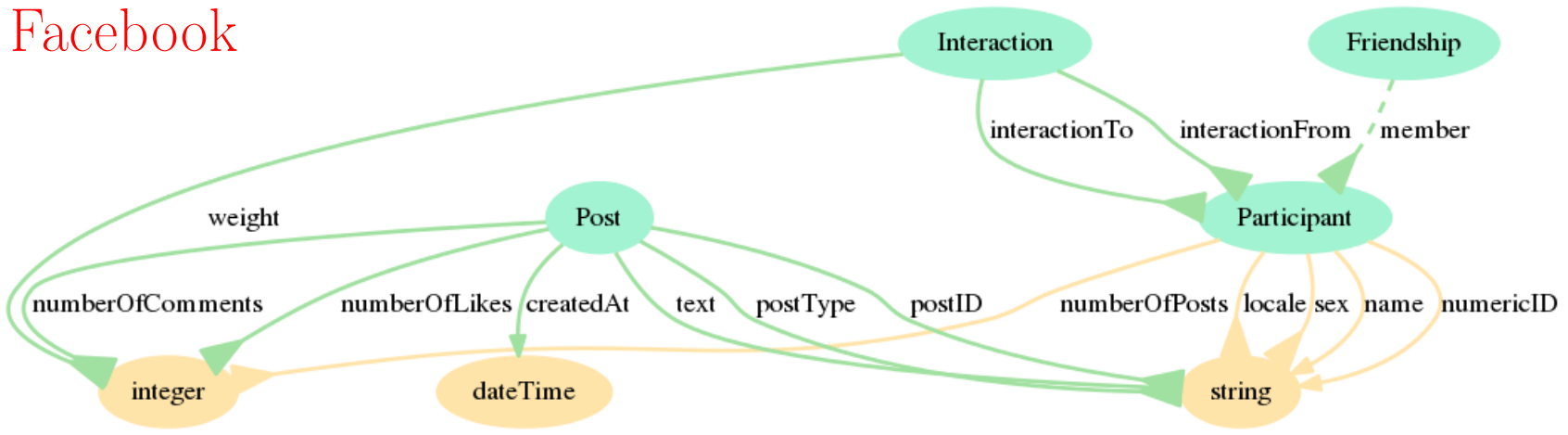
- green color if representing an OWL existential class restriction (all individuals from the class present at least one triple with the property as predicate);
- inverted nip if representing an OWL universal class restriction (all individuals presenting triples with the property as predicate are from 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.

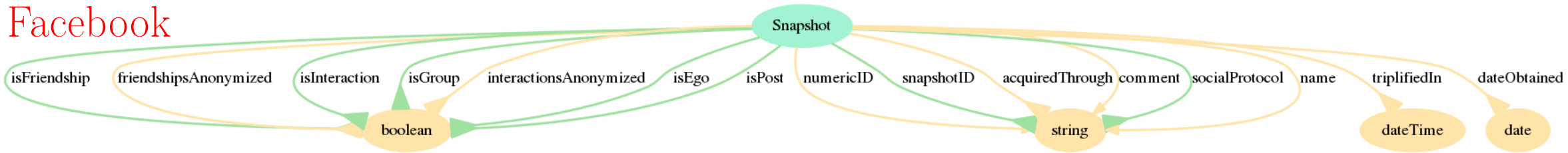
2. Facebook data

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



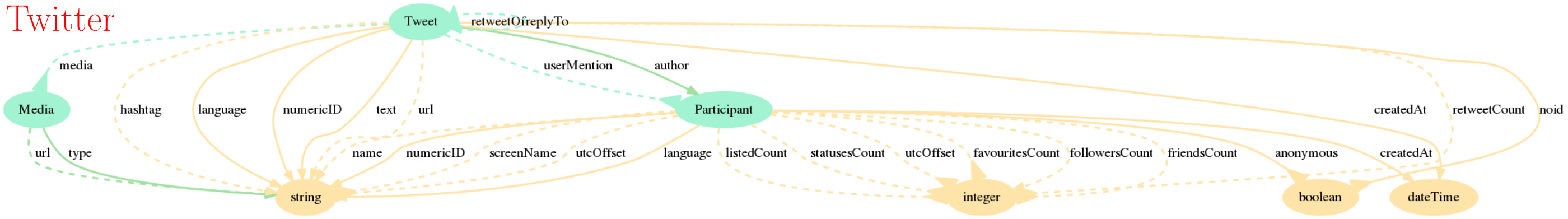
Facebook



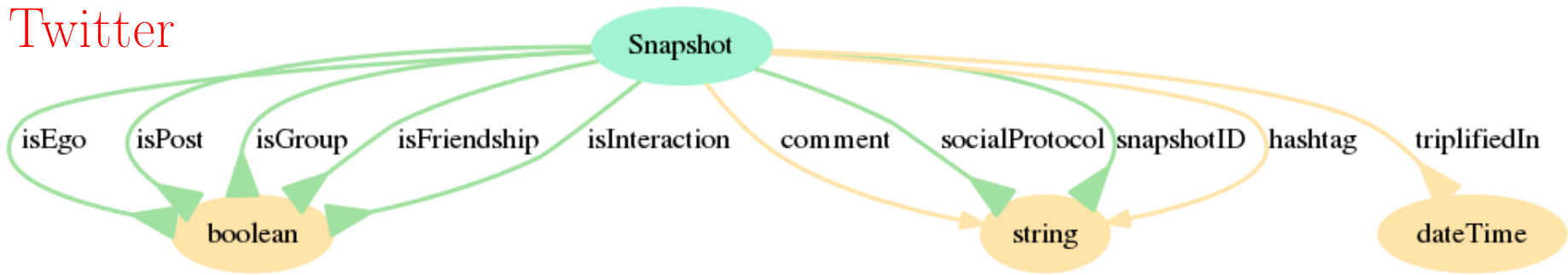
3. Twitter data

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



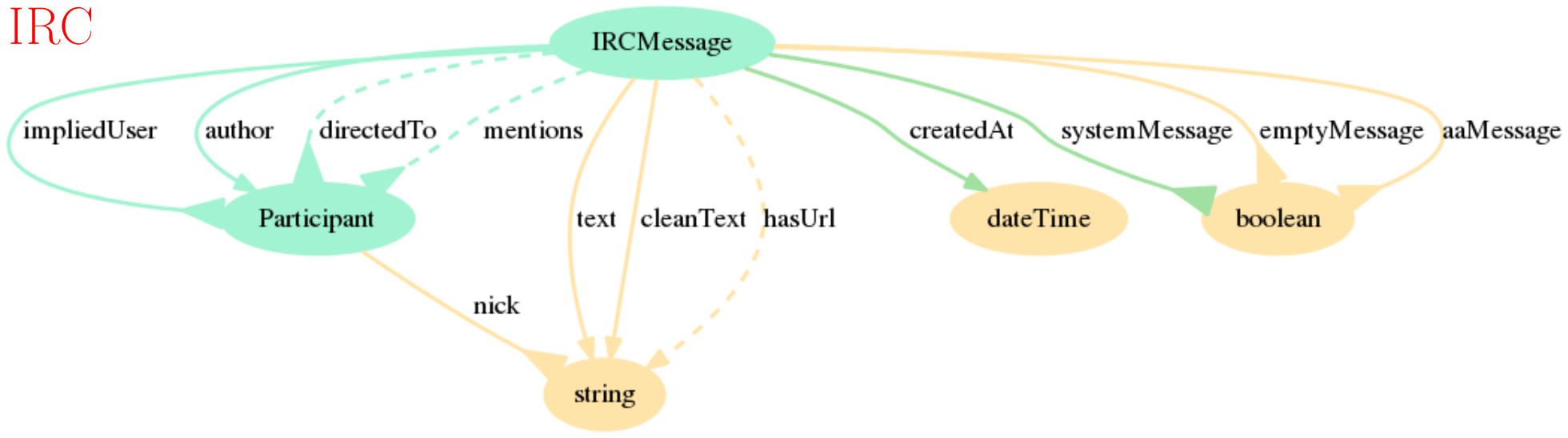
Twitter



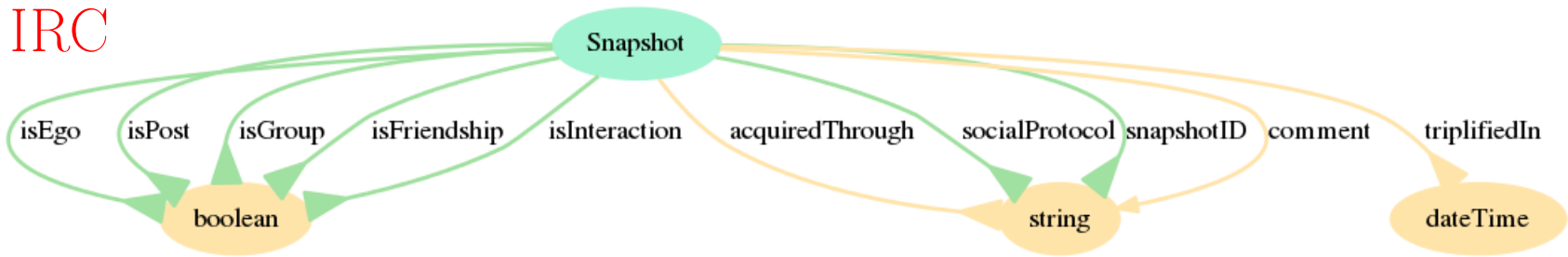
4. IRC data

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



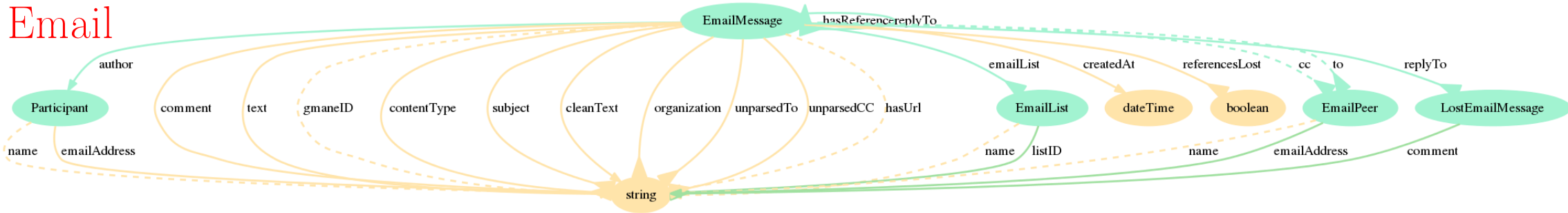
IRC

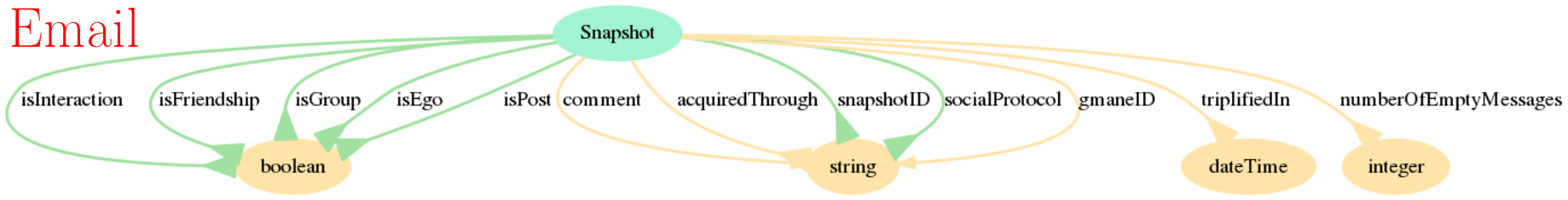


5. Email data

Each IRC snapshot is yield by an Email list. Interactions might be considered through `po:replyTo` relations Further information is found on the following diagrams, the tables on the end of this document or in the main document of this article [\[1\]](#).

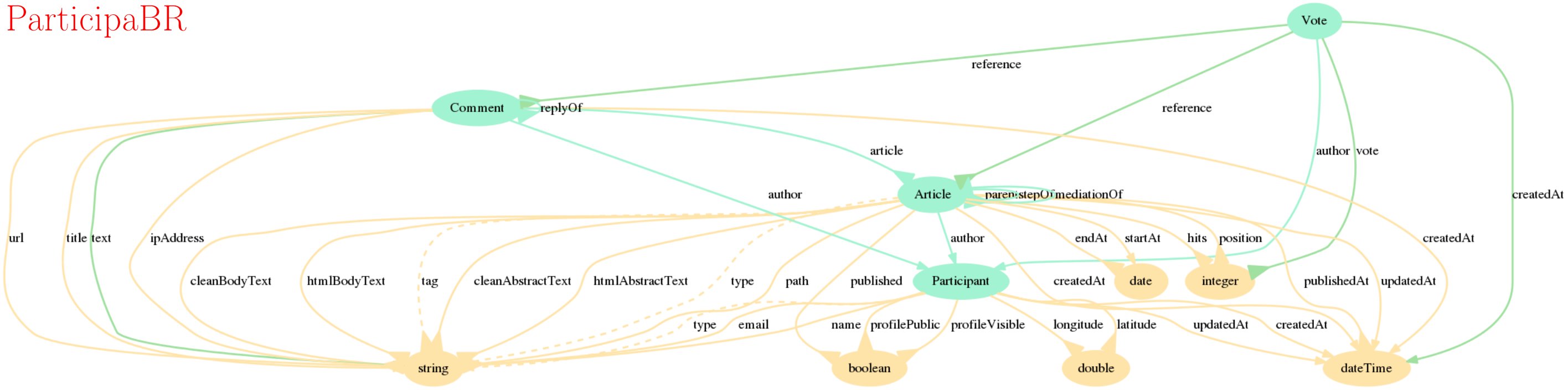
Email



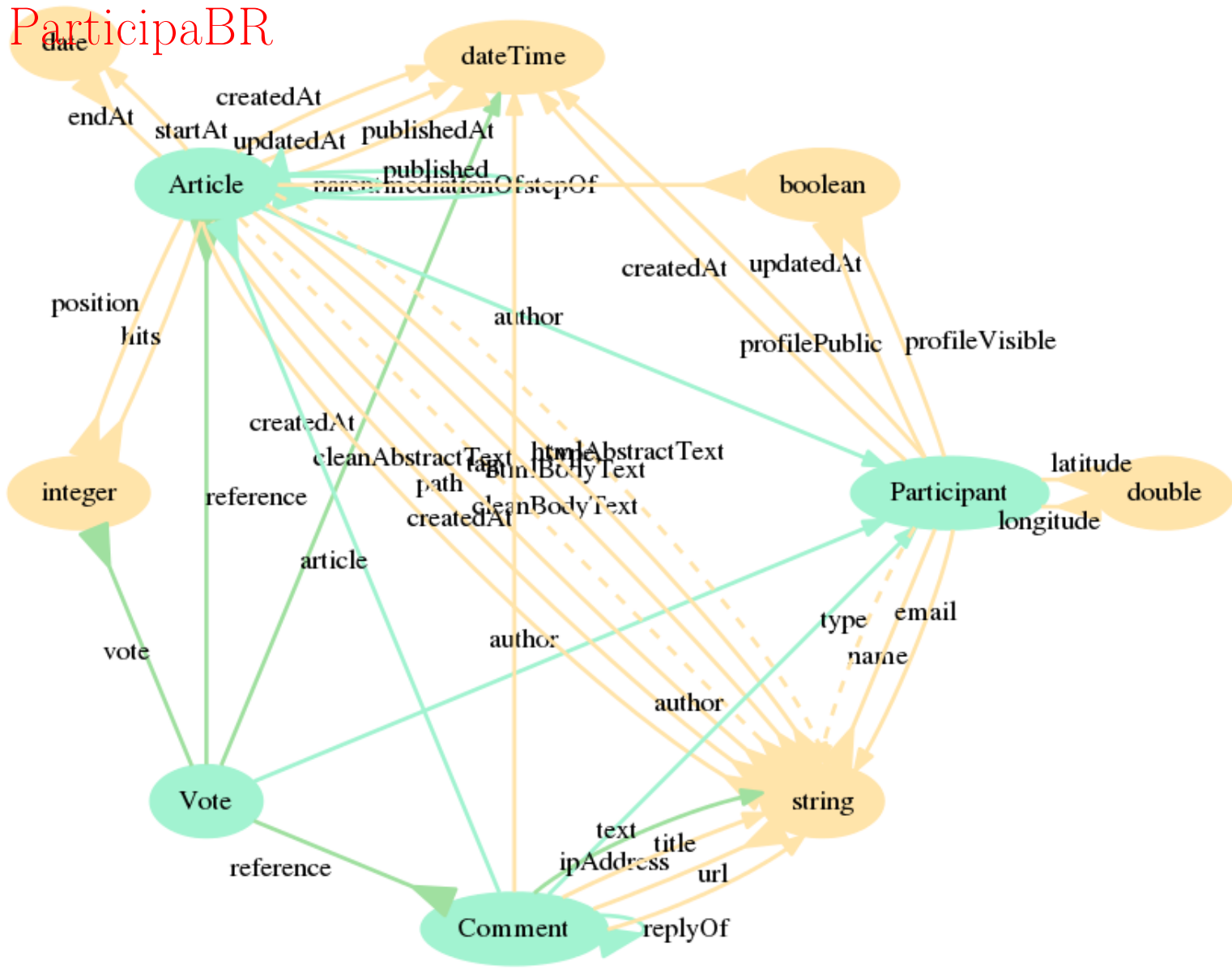


6. ParticipaBR data

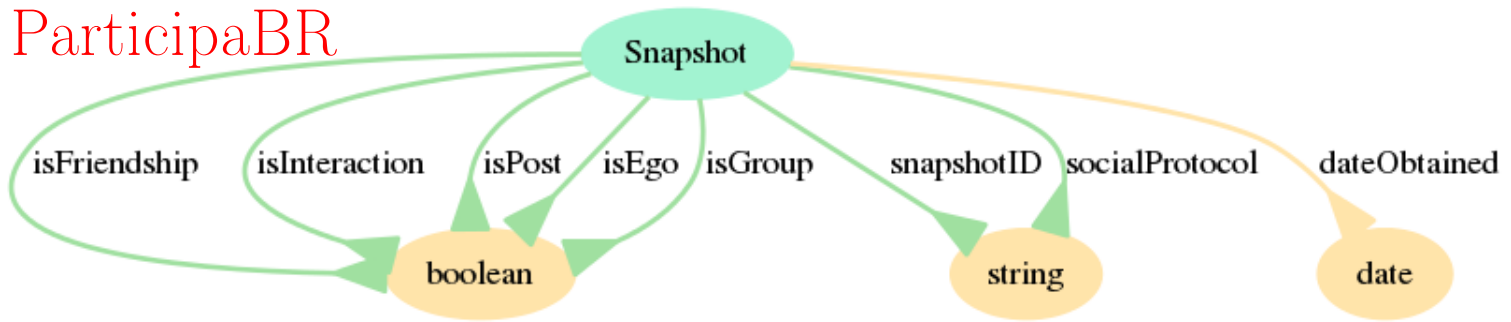
ParticipaBR



ParticipaBR

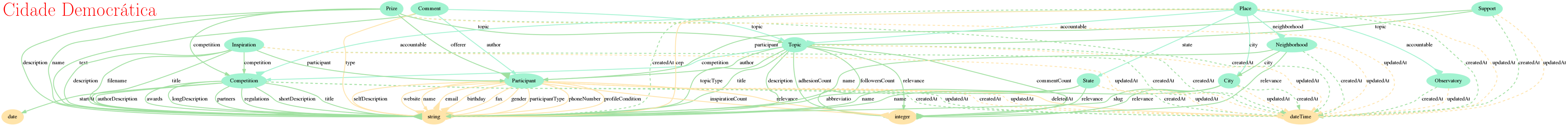


ParticipaBR

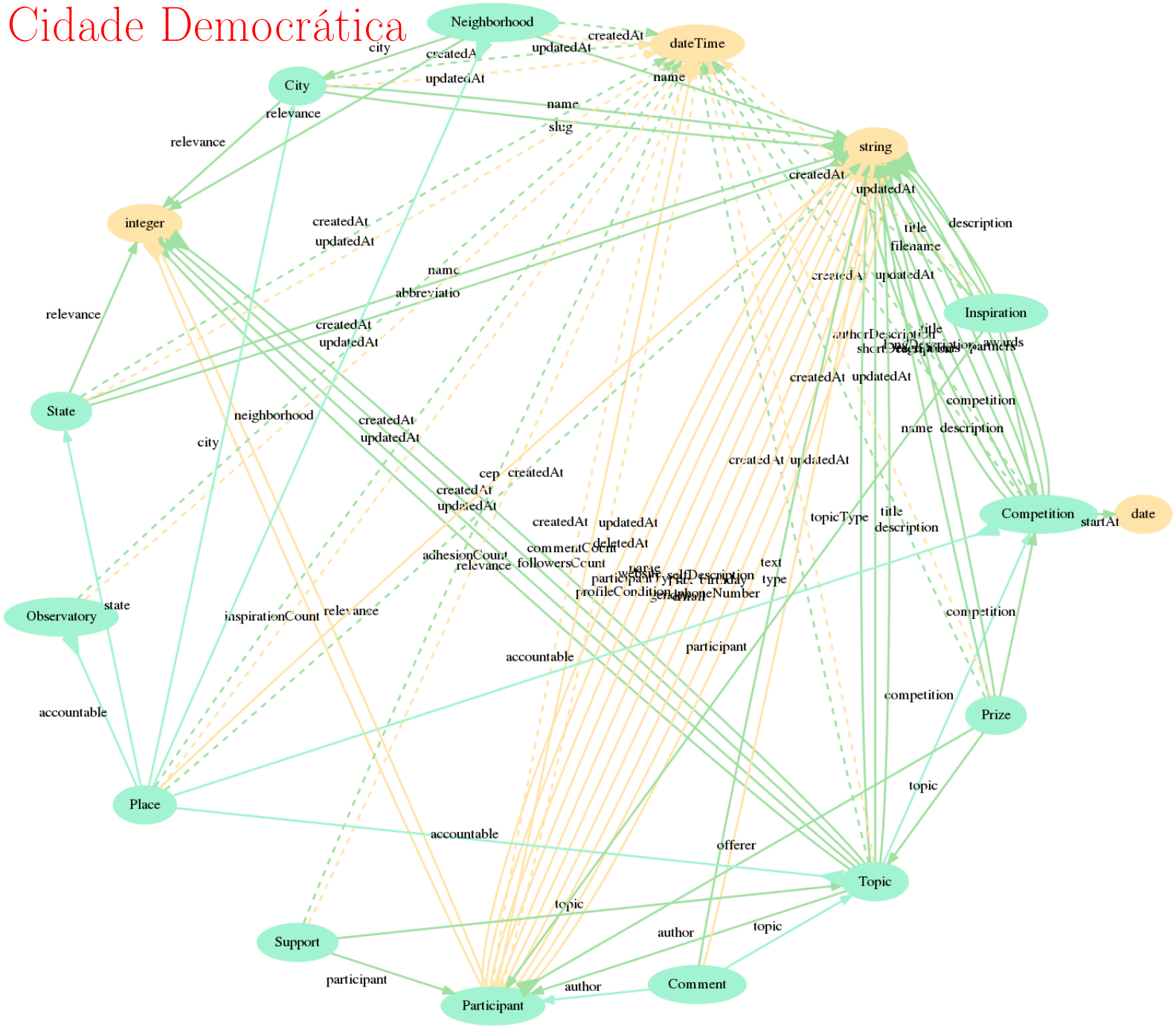


7. Cidade Democrática data

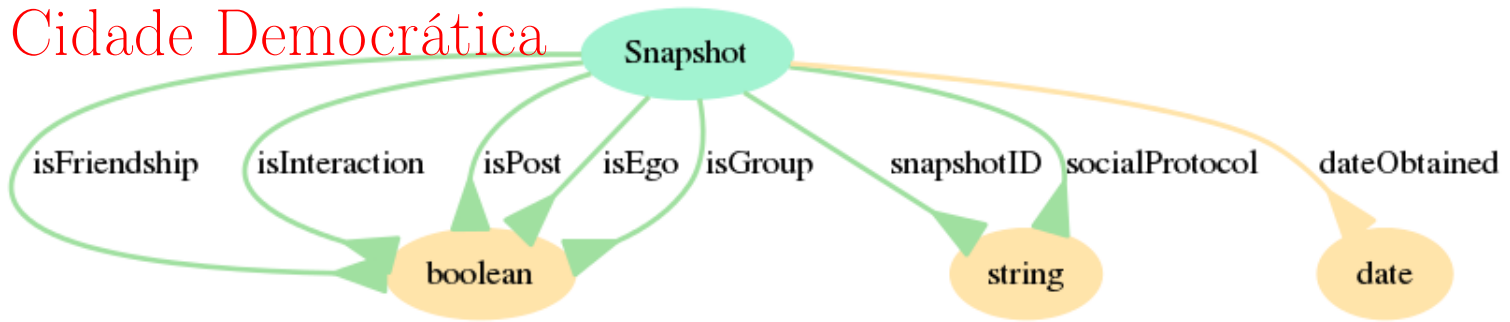
Cidade Democrática



Cidade Democrática

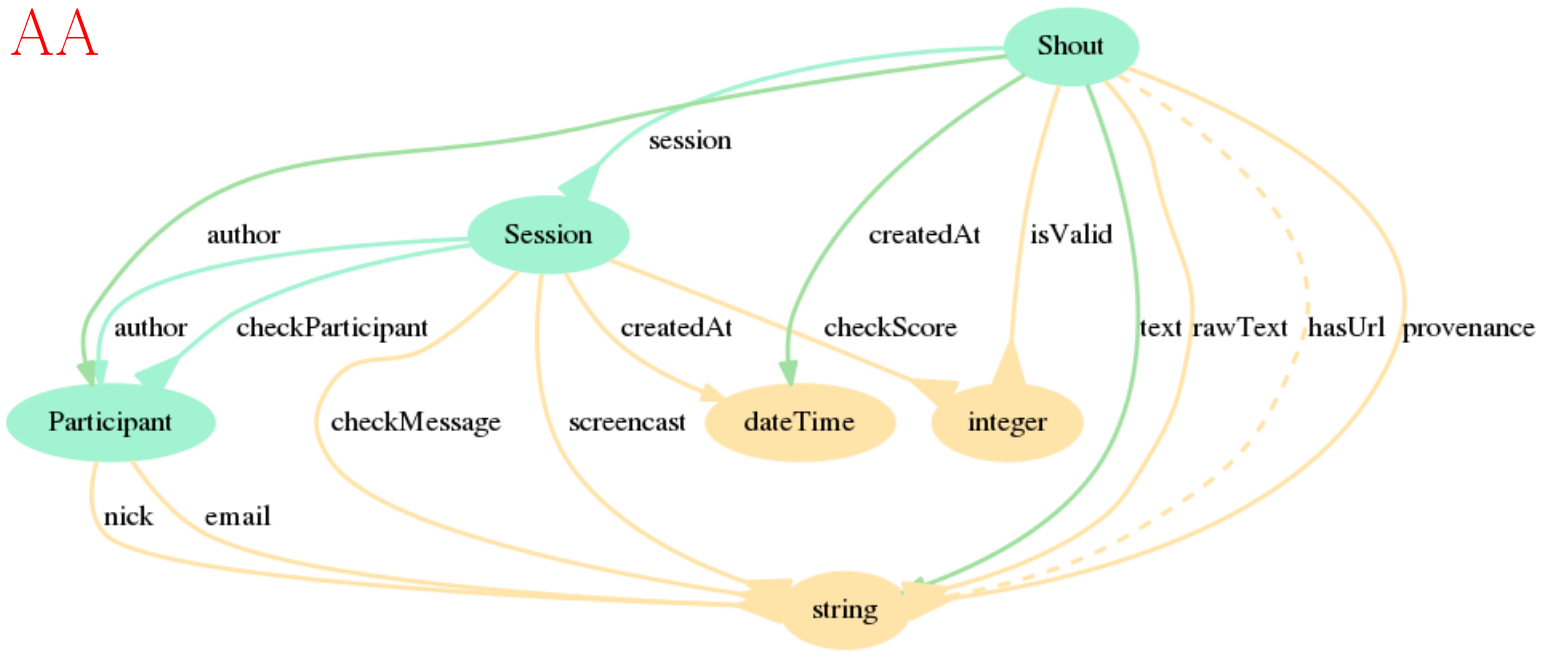


Cidade Democrática

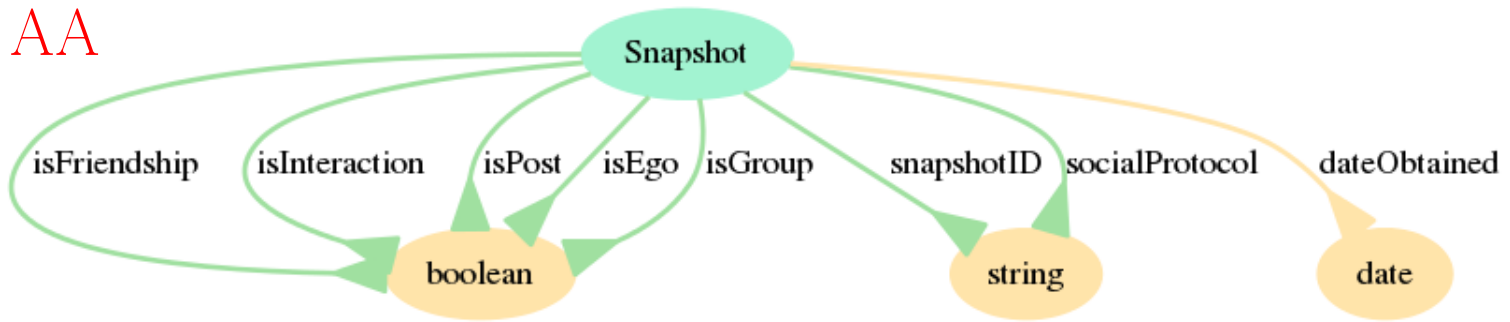


8. AA data

AA



AA



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).