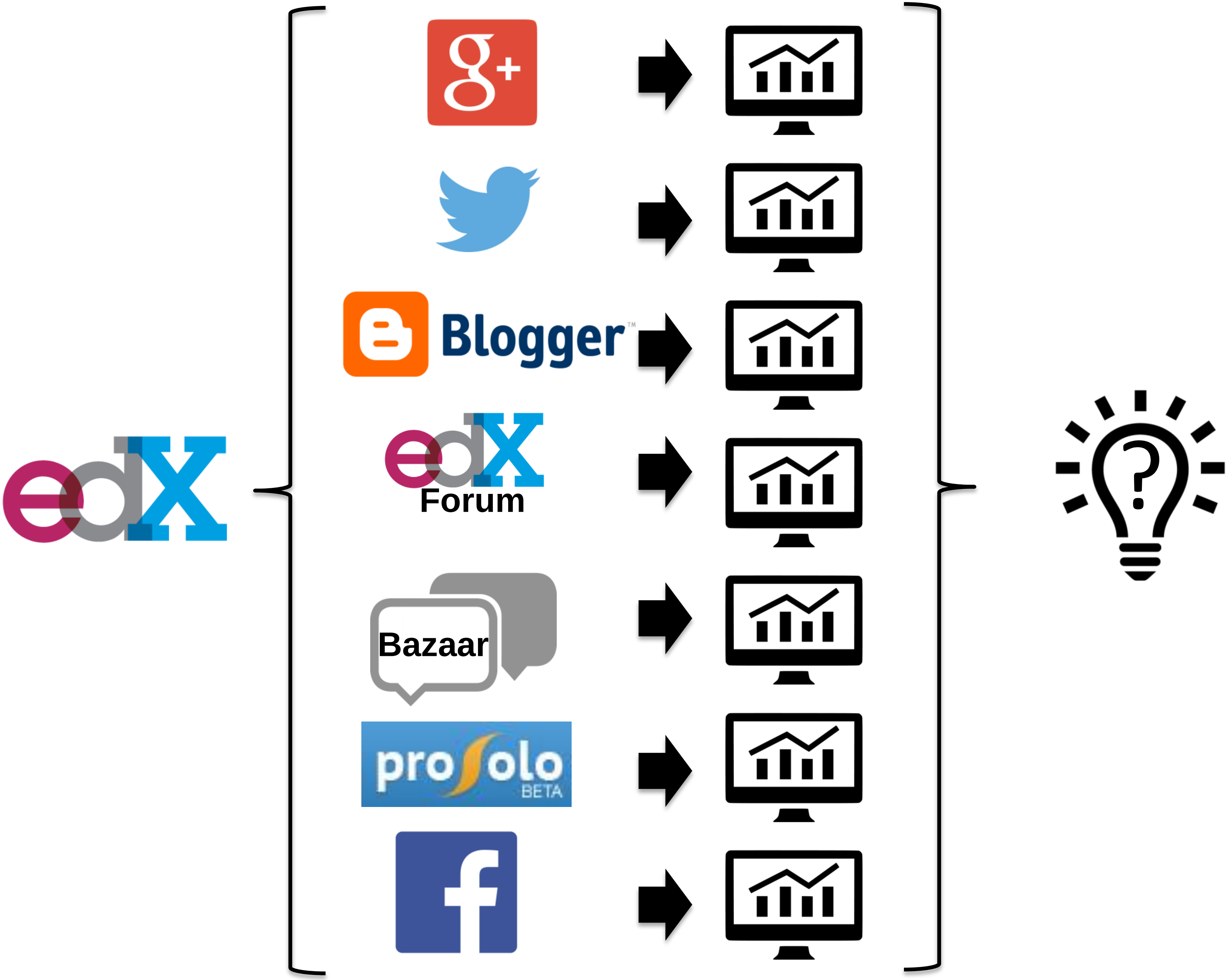
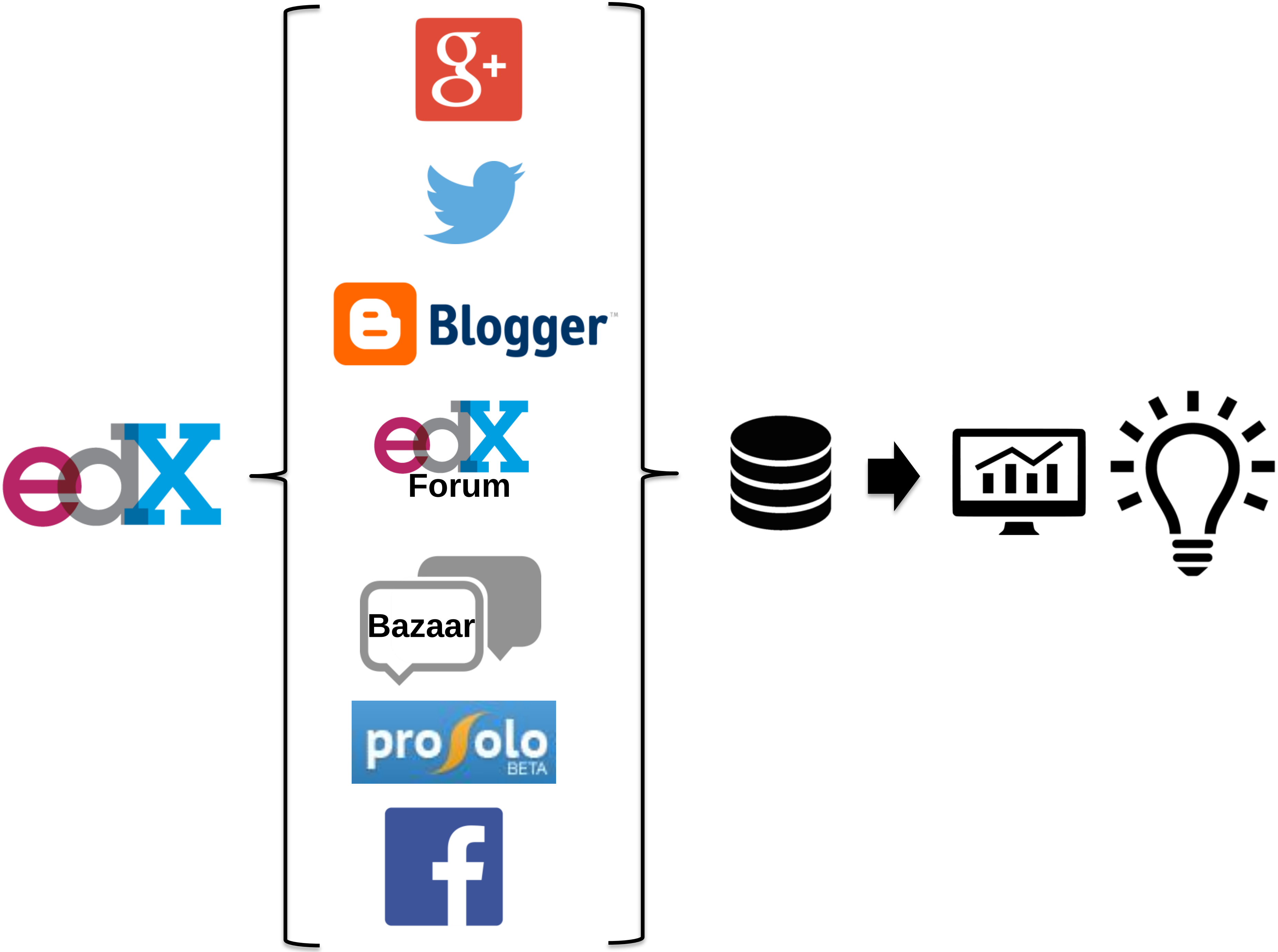
School of Computer Science
Language Technologies Institute

DiscourseDB

Facilitating Cross-Platform Analyses of Educational Discourse


edX
Blogger
edX
Forum
Bazaar
prosolo
BETA



edX
g+
Blogger™
edX
Forum
Bazaar
prosolo
BETA
f

DiscourseDB

D
A
T
A

Fusion >

- inference over multiple data sources

Abstraction >

- uniform representation
- simplification helps to form hypotheses
- retention of complexity avoids oversimplification

Contextualization >

- conversational context is key to understanding
- represent multimodal context and synchronize with discussion

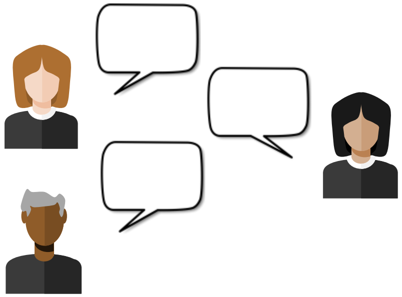
DiscourseDB

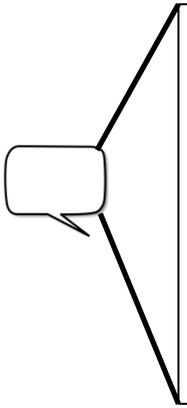
Macro Structure

- network structure of discourse
- integrates content from multiple sources
- keep track of changes over time

Micro Structure

- internal organization of contributions
- linguistic annotation
- content analysis
- multiple layers
- multimodal





Try, try and try again. Here is another selection of fields which captures some areas of CS and which are in the lead paragraph. Unfortunately the text may seem a bit long. Anyone who has any suggestions for a shorter text or objections to this?

Discussion in Context

- Course assignments with a discussion component require an integrated analysis of discussion and context
- Easy in traditional courses
- More loosely structured courses (cMOOCs) more challenging
  - Discussion spread out over platforms
  - Fuzzy boundaries between learning tasks

- **Discussion context is an integral component of DiscourseDB and will enable analyses in weakly structured learning environments.**

Multi-Layer Representation of Discourse

- Discourse is multi-dimensional – different research questions require different views
- Explicit vs. implicit (inferred) discourse structure
- Access and clickstream data to represent “silent” user interactions
  - Reading posts or profiles
  - Following users/threads/pages

Data Source Management

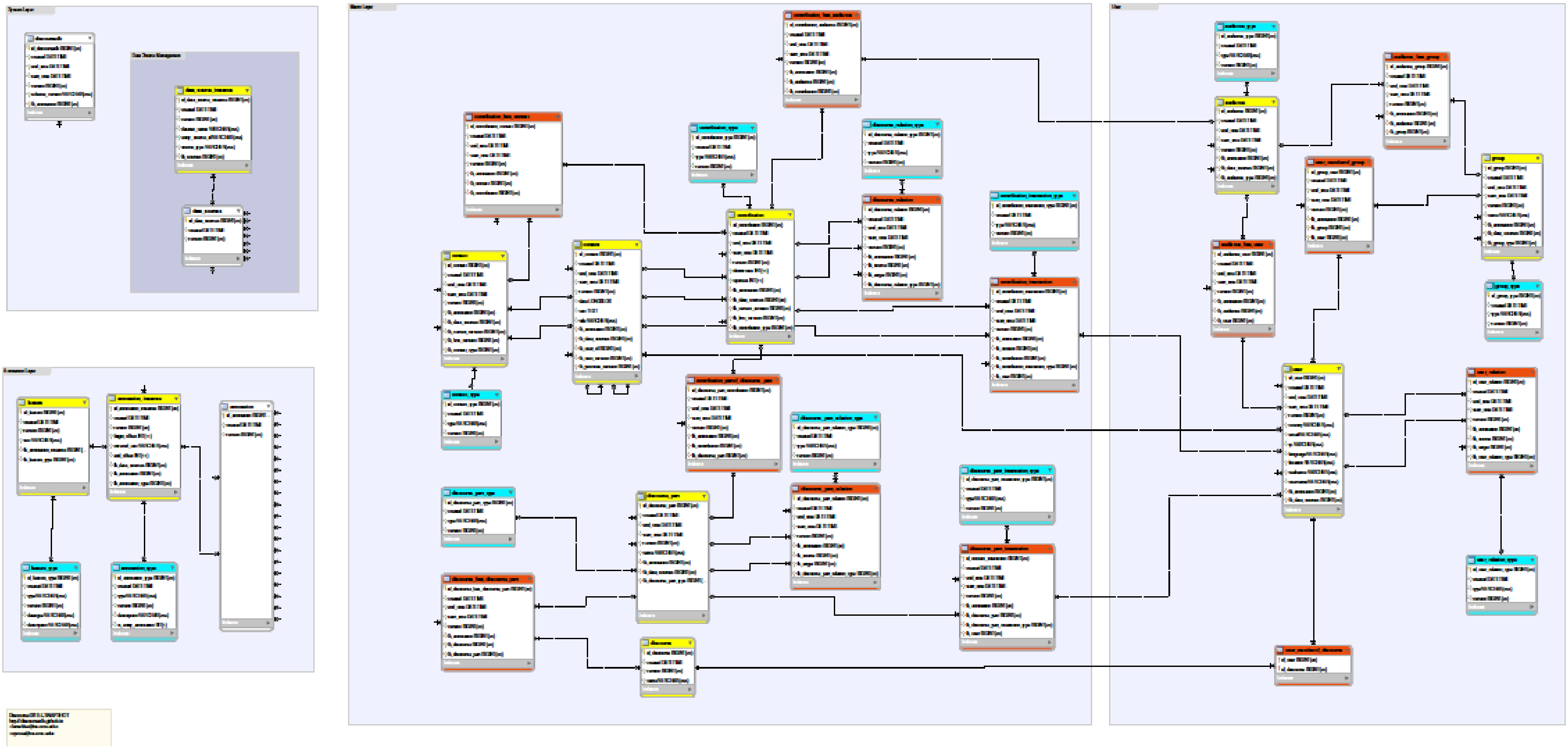
- DiscourseDB unifies discourse data
- BUT: we want to keep track of where each data point came from

➔Necessary at import time in order to make sense of relational data

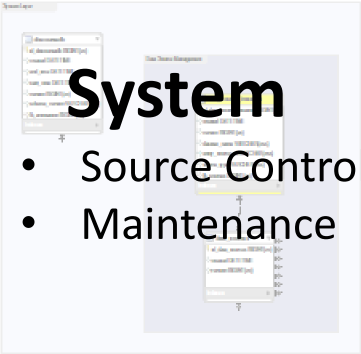
➔Helpful later to recover details that are not explicitly represented in DiscourseDB

Source-controlled **entities in DiscourseDB keep track of where they came from and how the entity can be found in the source.**

SCHEMA

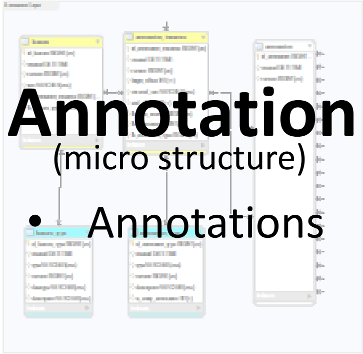


DiscourseDB Layers



System

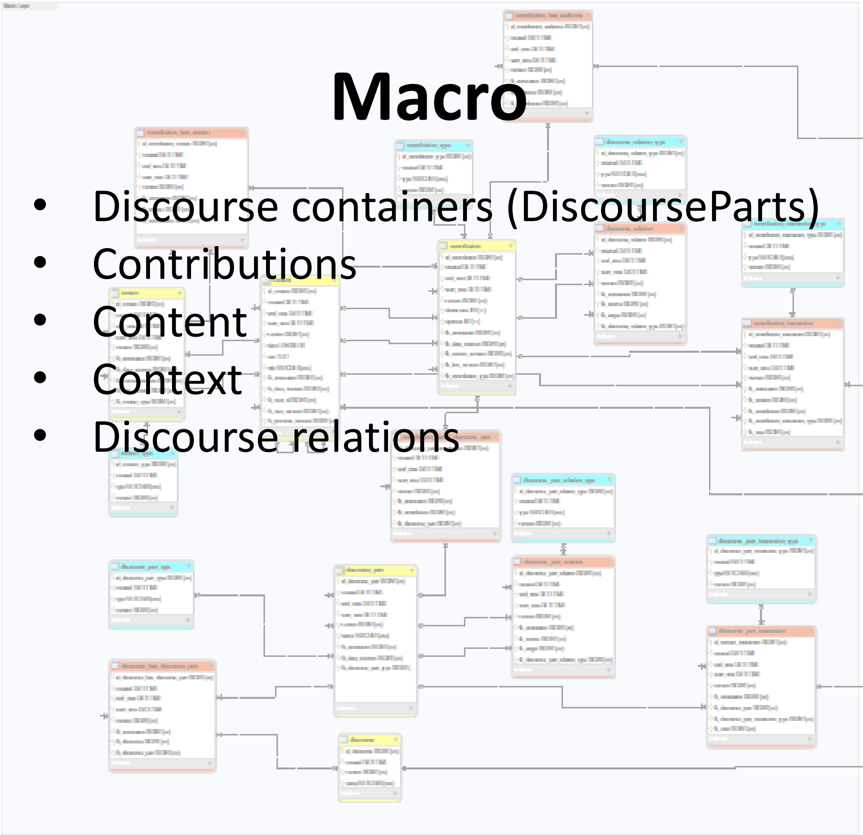
- Source Control
- Maintenance



Annotation

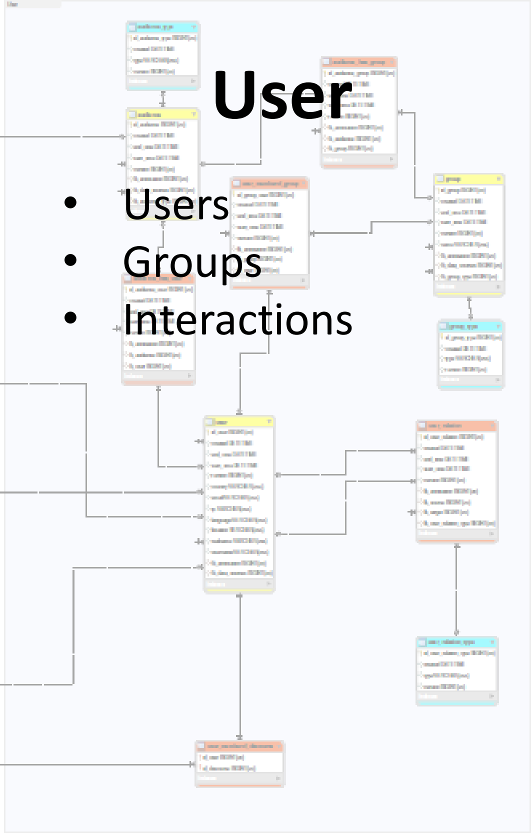
(micro structure)

- Annotations



Macro

- Discourse containers (DiscourseParts)
- Contributions
- Content
- Context
- Discourse relations



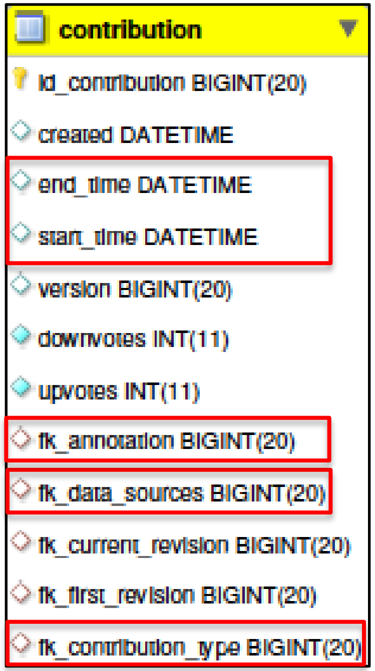
User

- Users
- Groups
- Interactions

DiscourseDB Entity Properties

aka “entity capabilities”

- Timed entities
  - – start date
  - – end date
- Annotatable entities
  - – set of annotations
- Source controlled entities
  - – set of data sources
- Type entities
  - – tied to a non-type entities



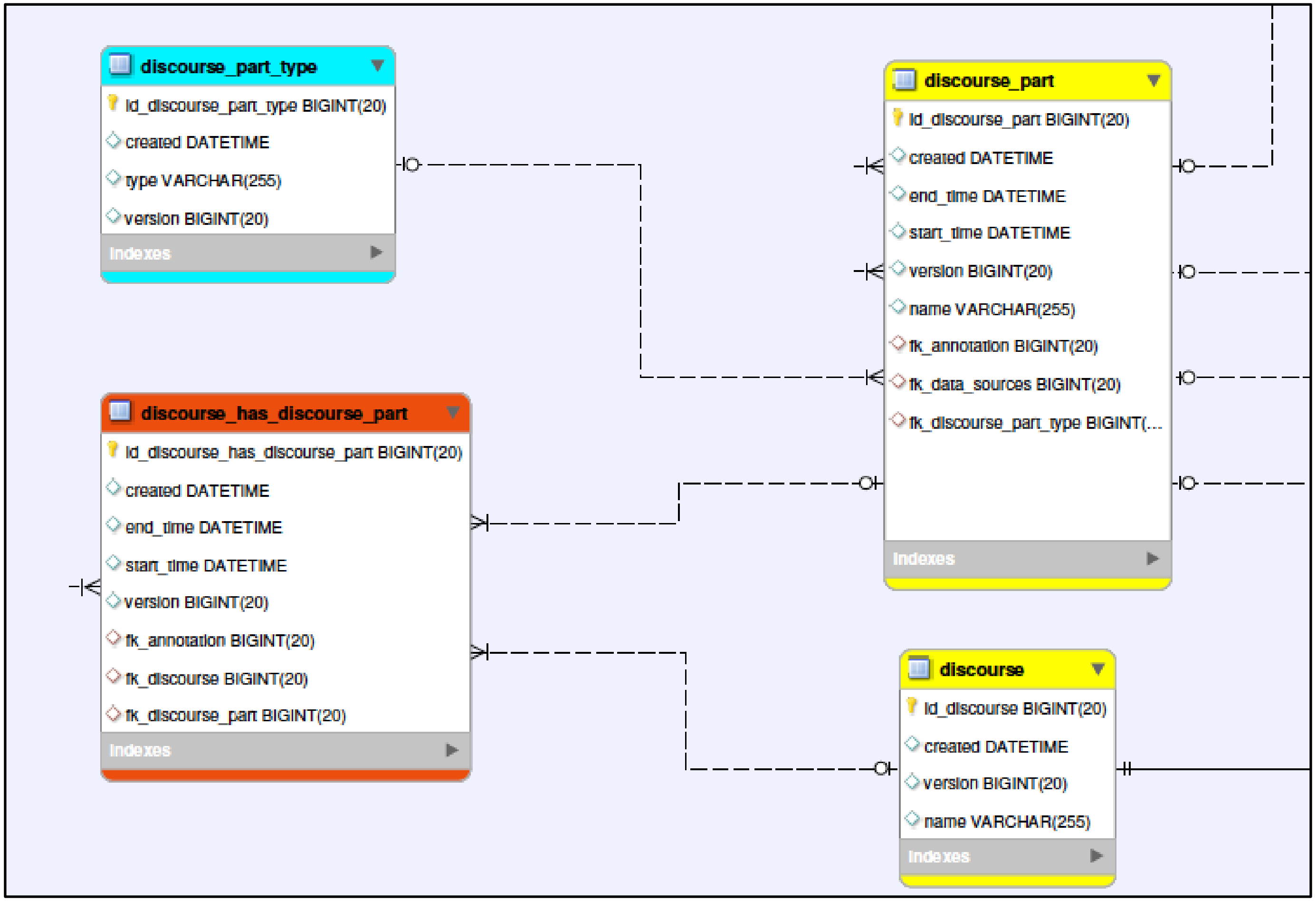
contribution

id_contribution BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
downvotes INT(11)
upvotes INT(11)
fk_annotation BIGINT(20)
fk_data_sources BIGINT(20)
fk_current_revision BIGINT(20)
fk_first_revision BIGINT(20)
fk_contribution_type BIGINT(20)

Central Entities

- Discourse
- DiscoursePart
- Contribution
- Context
- Content
- User

Discourse and DiscoursePart



discourse_part_type
- id_discourse_part_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)
- Indexes

discourse_part
- id_discourse_part BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- name VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_discourse_part_type BIGINT(...
- Indexes

discourse_has_discourse_part
- id_discourse_has_discourse_part BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- fk_annotation BIGINT(20)
- fk_discourse BIGINT(20)
- fk_discourse_part BIGINT(20)
- Indexes

discourse
- id_discourse BIGINT(20)
- created DATETIME
- version BIGINT(20)
- name VARCHAR(255)
- Indexes

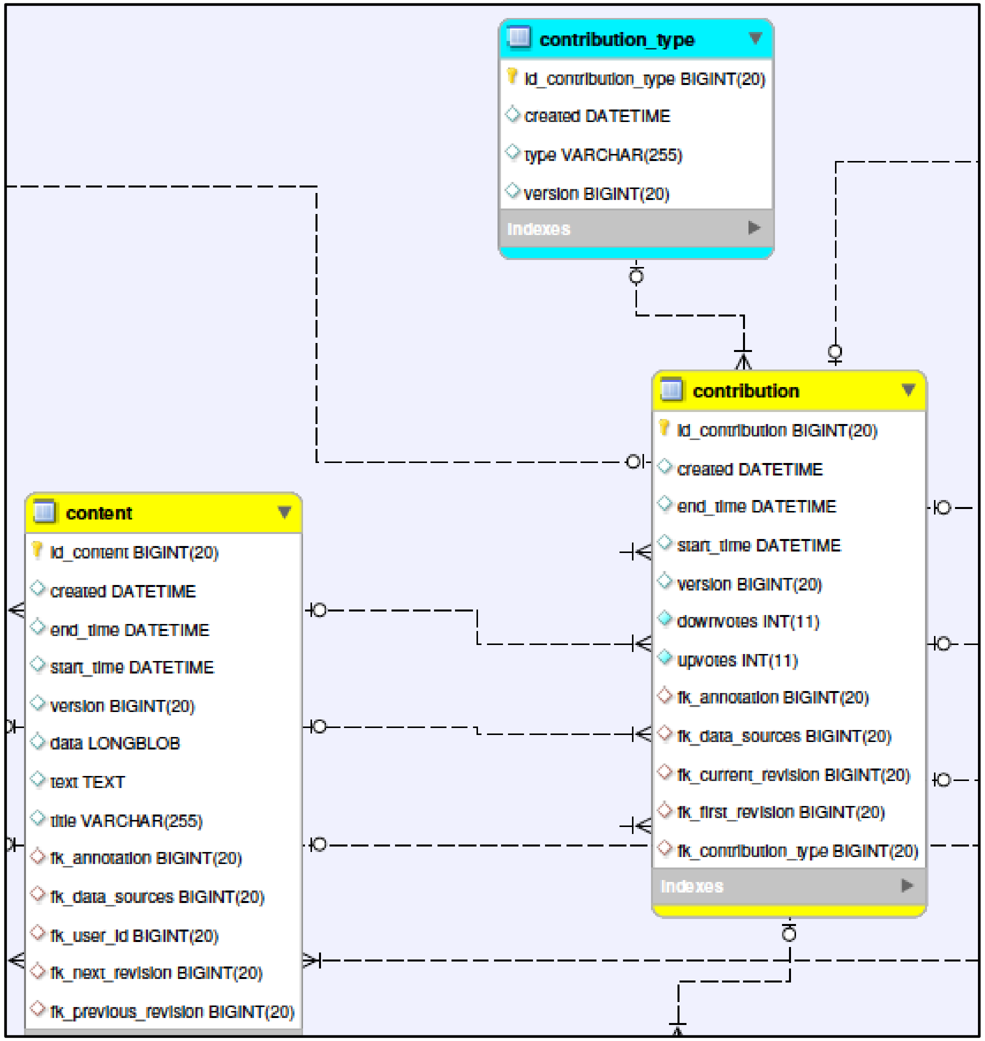
DISCOURSE

- COURSE INSTANCE
- WIKIPEDIA INSTANCE
- EXPERIMENT

DISCOURSE PART

- FORUM
- SUBFORUM
- THREAD
- TALK PAGE
- BLOG

Contribution/Content



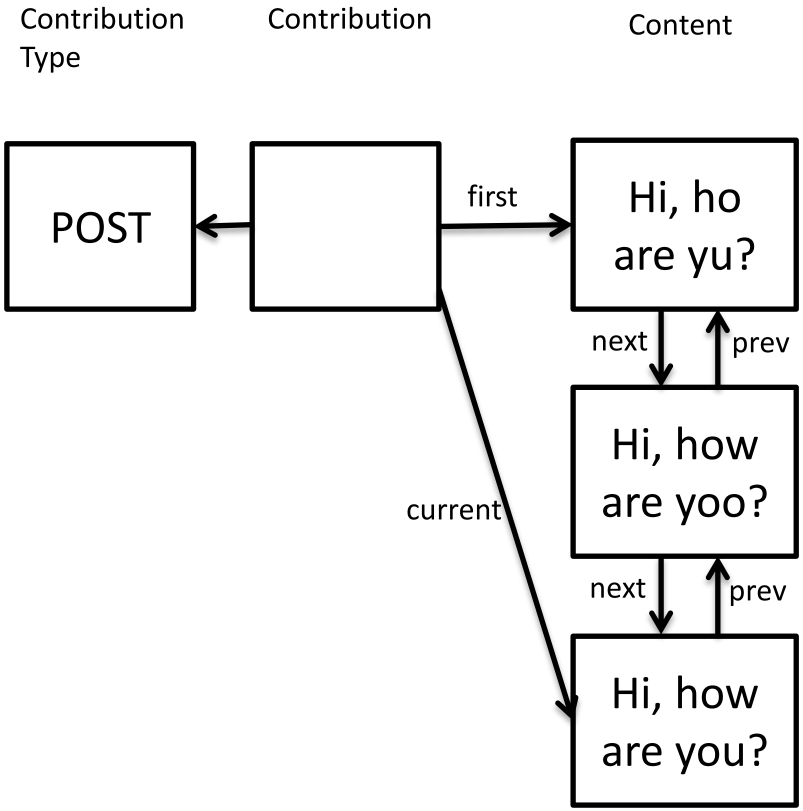
contribution_type
- id_contribution_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)

Indexes

contribution
- id_contribution BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- downvotes INT(11)
- upvotes INT(11)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_current_revision BIGINT(20)
- fk_first_revision BIGINT(20)
- fk_contribution_type BIGINT(20)

Indexes

content
- id_content BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- data LONGBLOB
- text TEXT
- title VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_user_id BIGINT(20)
- fk_next_revision BIGINT(20)
- fk_previous_revision BIGINT(20)



Contribution Type

Contribution

Content

POST

first

Hi, ho are yu?

next

prev

Hi, how are yoo?

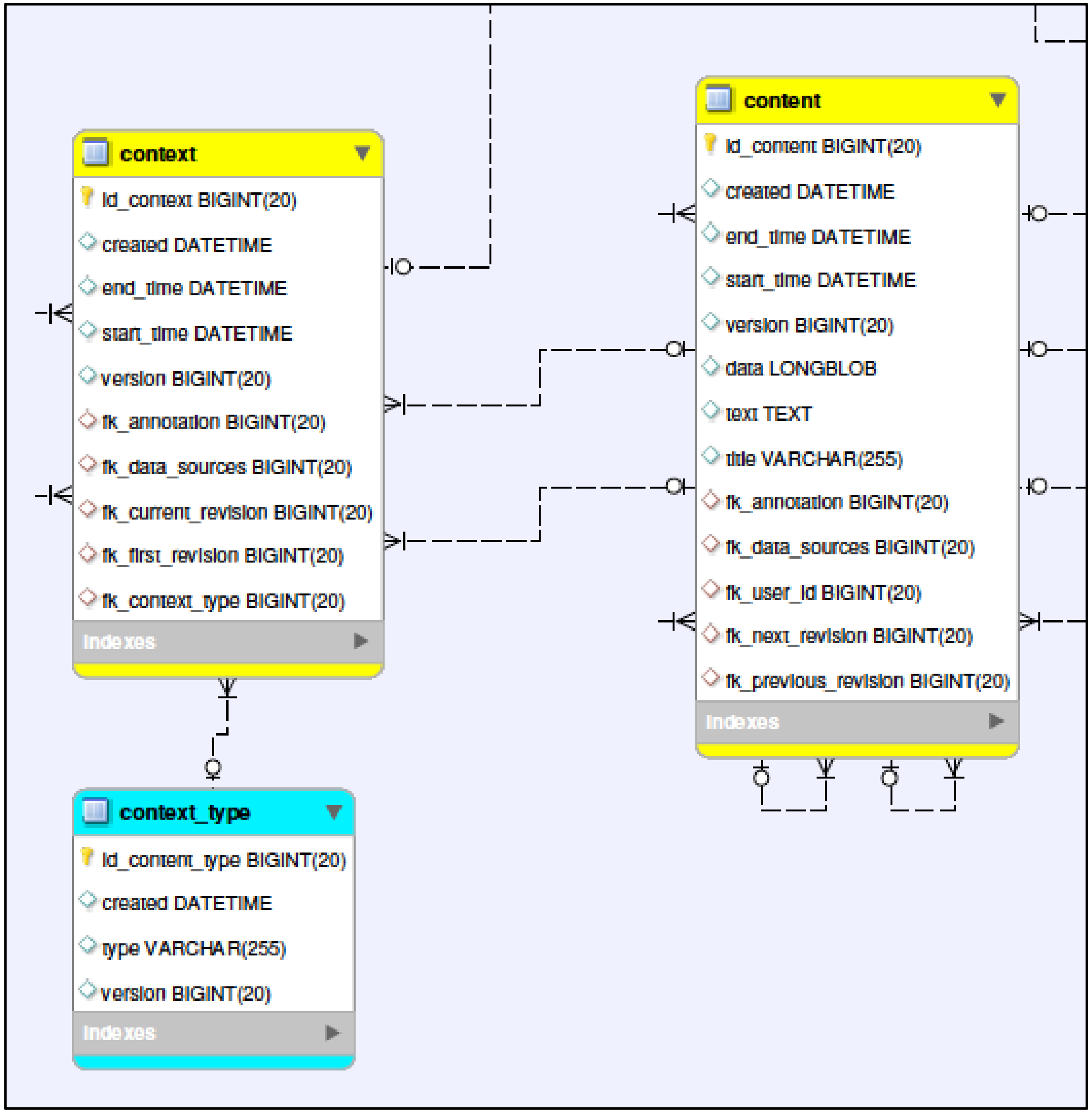
current

next

prev

Hi, how are you?

Context



context
- id_context BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_current_revision BIGINT(20)
- fk_first_revision BIGINT(20)
- fk_context_type BIGINT(20)

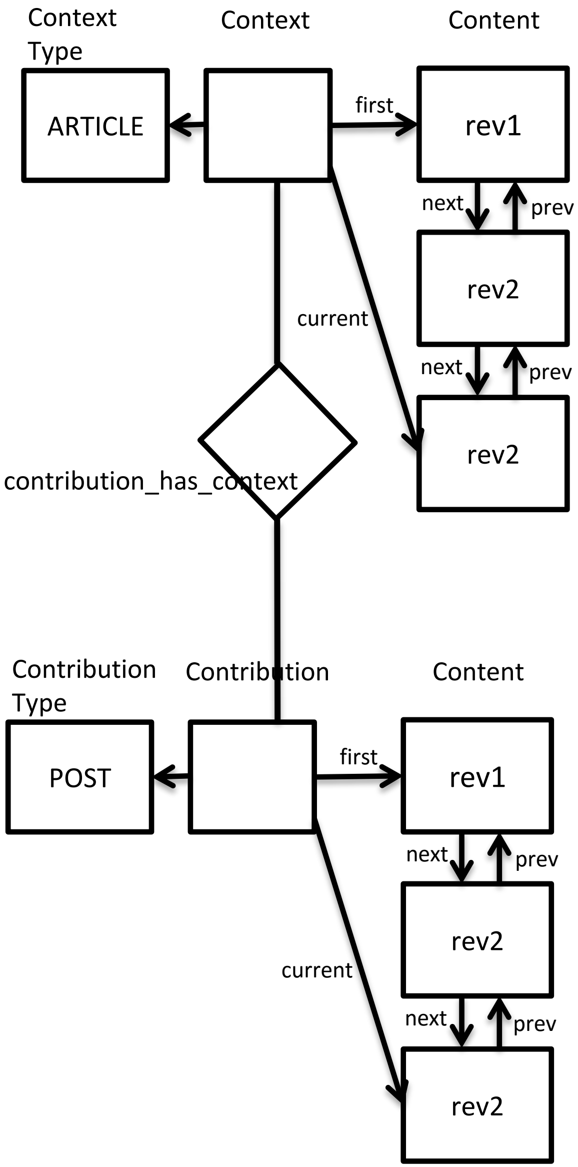
Indexes

content
- id_content BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- data LONGBLOB
- text TEXT
- title VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_user_id BIGINT(20)
- fk_next_revision BIGINT(20)
- fk_previous_revision BIGINT(20)

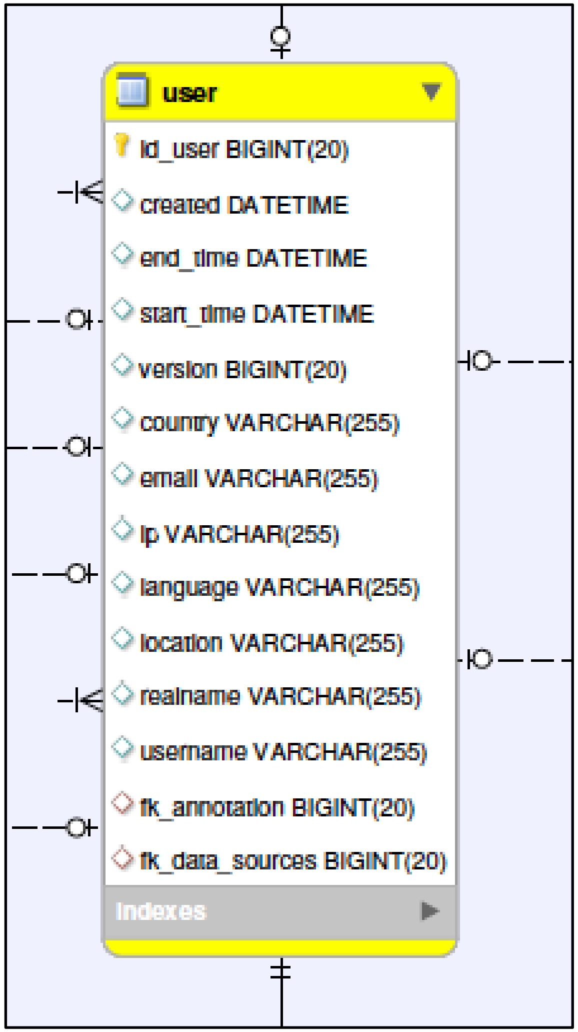
Indexes

context_type
- id_content_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)

Indexes

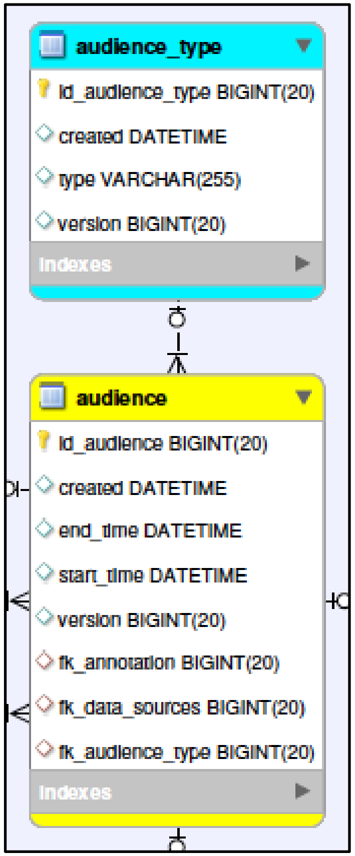


User



user

- id_user BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- country VARCHAR(255)
- email VARCHAR(255)
- ip VARCHAR(255)
- language VARCHAR(255)
- location VARCHAR(255)
- realname VARCHAR(255)
- username VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)

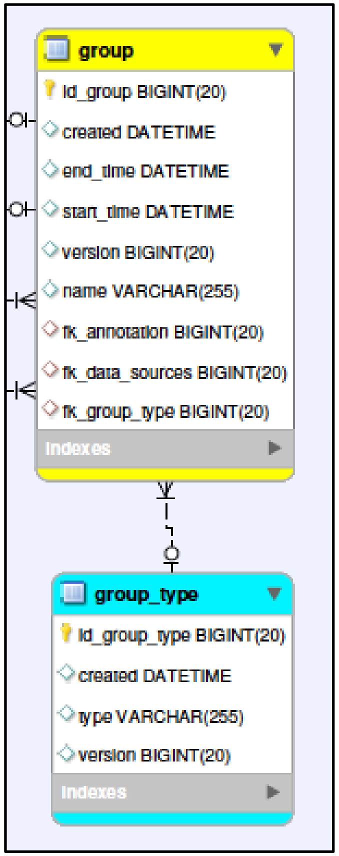


audience_type

- id_audience_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)

audience

- id_audience BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_audience_type BIGINT(20)



group

- id_group BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- name VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_group_type BIGINT(20)

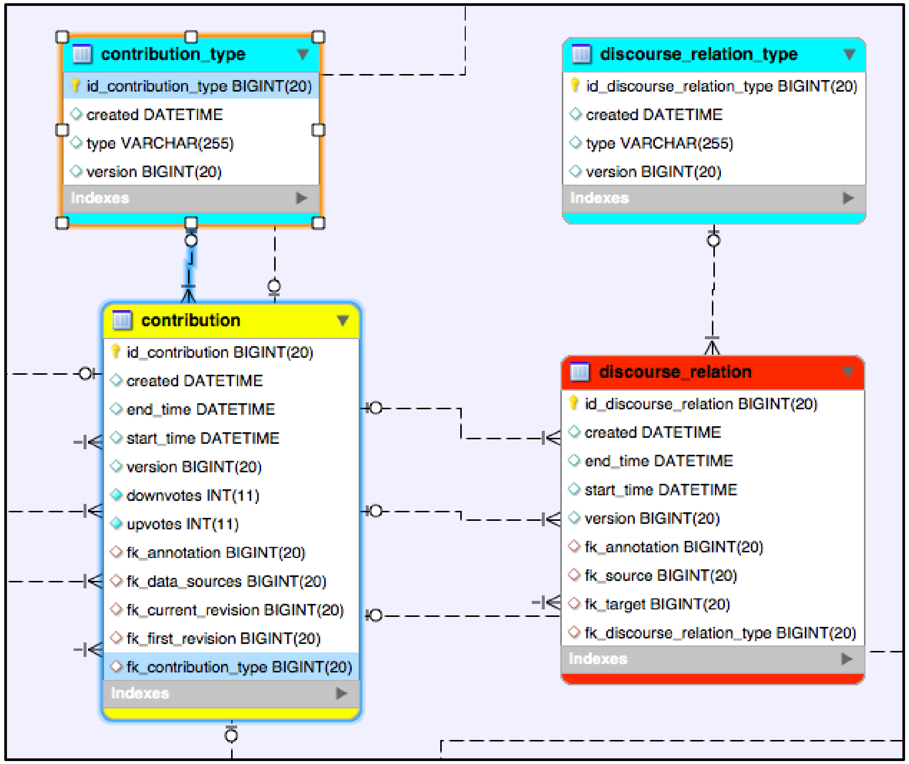
group_type

- id_group_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)

Central Relations

- DiscourseRelation
- DiscoursePartRelation
- DiscoursePartInteraction
- ContributionInteraction

Discourse(Part)Relation

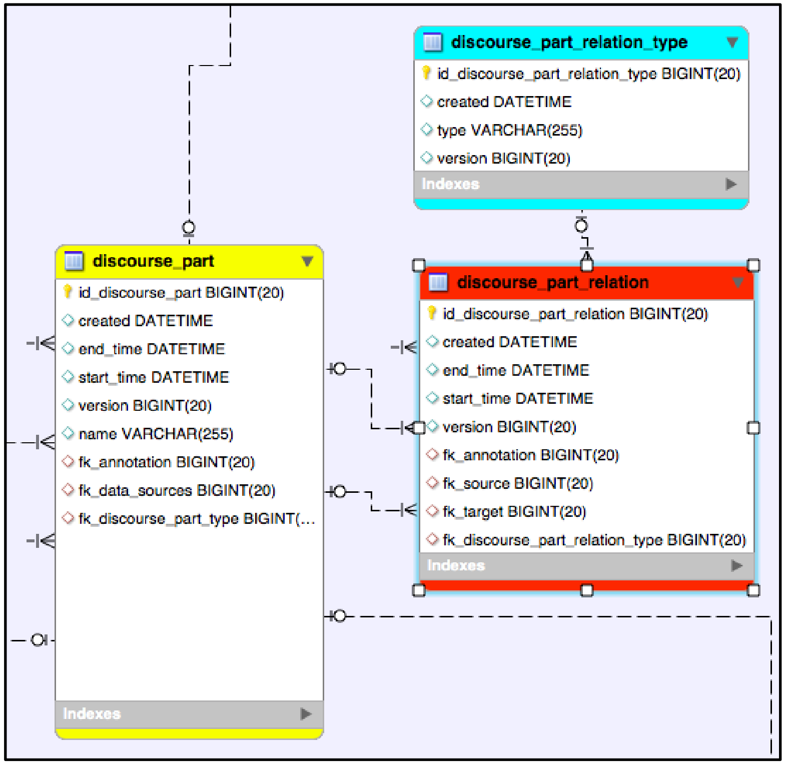


contribution_type
id_contribution_type BIGINT(20)
created DATETIME
type VARCHAR(255)
version BIGINT(20)
Indexes

discourse_relation_type
id_discourse_relation_type BIGINT(20)
created DATETIME
type VARCHAR(255)
version BIGINT(20)
Indexes

contribution
id_contribution BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
downvotes INT(11)
upvotes INT(11)
fk_annotation BIGINT(20)
fk_data_sources BIGINT(20)
fk_current_revision BIGINT(20)
fk_first_revision BIGINT(20)
fk_contribution_type BIGINT(20)
Indexes

discourse_relation
id_discourse_relation BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
fk_annotation BIGINT(20)
fk_source BIGINT(20)
fk_target BIGINT(20)
fk_discourse_relation_type BIGINT(20)
Indexes



discourse_part_relation_type
id_discourse_part_relation_type BIGINT(20)
created DATETIME
type VARCHAR(255)
version BIGINT(20)
Indexes

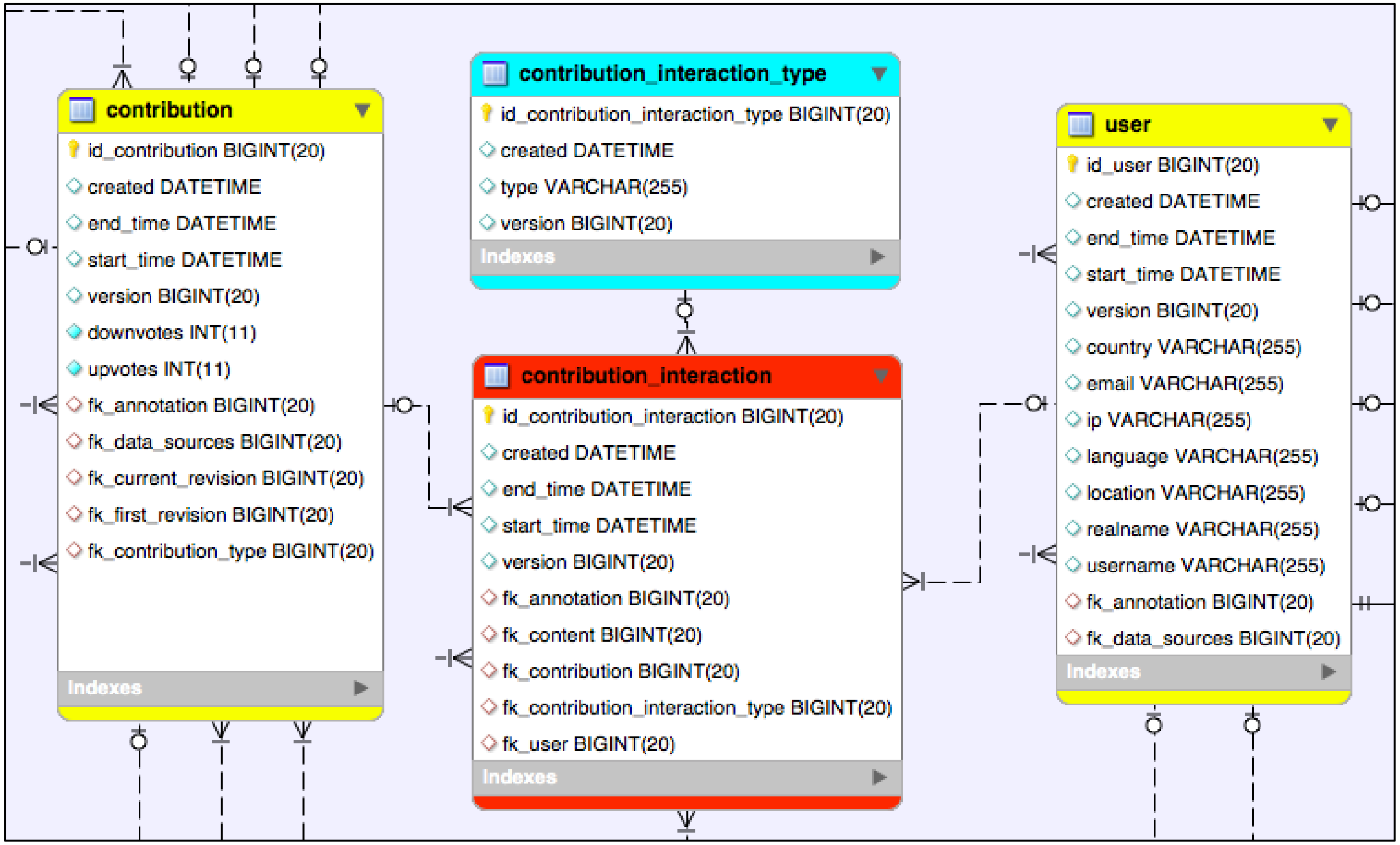
discourse_part
id_discourse_part BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
name VARCHAR(255)
fk_annotation BIGINT(20)
fk_data_sources BIGINT(20)
fk_discourse_part_type BIGINT(…
Indexes

discourse_part_relation
id_discourse_part_relation BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
fk_annotation BIGINT(20)
fk_source BIGINT(20)
fk_target BIGINT(20)
fk_discourse_part_relation_type BIGINT(20)
Indexes

- Reply
- Descendant
- Comment
- Reshare

- TalkPageDiscussion
- SubForum

Interactions



contribution

- id_contribution BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- downvotes INT(11)
- upvotes INT(11)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)
- fk_current_revision BIGINT(20)
- fk_first_revision BIGINT(20)
- fk_contribution_type BIGINT(20)

Indexes

contribution_interaction_type

- id_contribution_interaction_type BIGINT(20)
- created DATETIME
- type VARCHAR(255)
- version BIGINT(20)

Indexes

contribution_interaction

- id_contribution_interaction BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- fk_annotation BIGINT(20)
- fk_content BIGINT(20)
- fk_contribution BIGINT(20)
- fk_contribution_interaction_type BIGINT(20)
- fk_user BIGINT(20)

Indexes

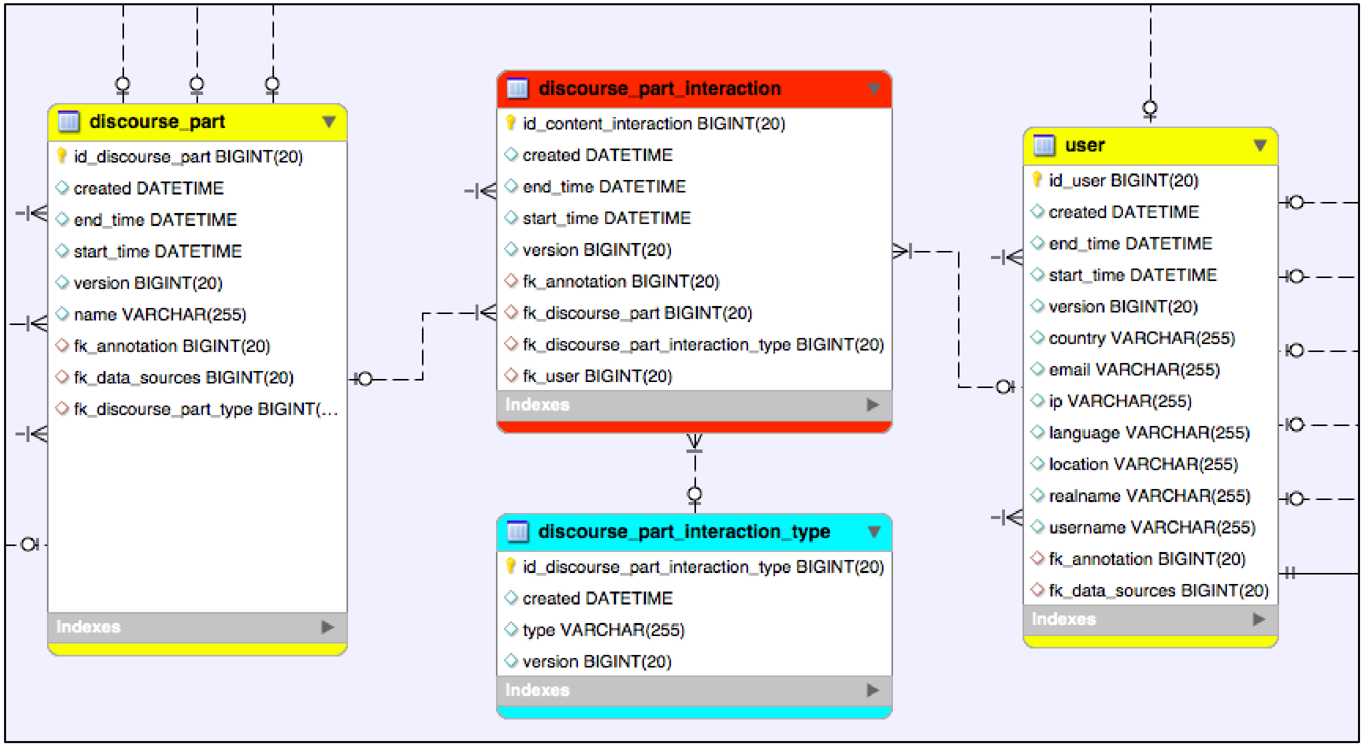
user

- id_user BIGINT(20)
- created DATETIME
- end_time DATETIME
- start_time DATETIME
- version BIGINT(20)
- country VARCHAR(255)
- email VARCHAR(255)
- ip VARCHAR(255)
- language VARCHAR(255)
- location VARCHAR(255)
- realname VARCHAR(255)
- username VARCHAR(255)
- fk_annotation BIGINT(20)
- fk_data_sources BIGINT(20)

Indexes

- Read
- Follow
- Delete
- Revert
- Share

Interactions



discourse_part
id_discourse_part BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
name VARCHAR(255)
fk_annotation BIGINT(20)
fk_data_sources BIGINT(20)
fk_discourse_part_type BIGINT(…
Indexes

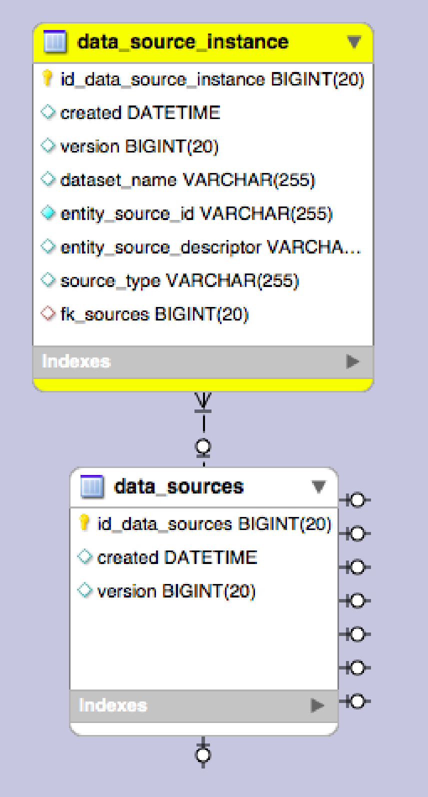
discourse_part_interaction
id_content_interaction BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
fk_annotation BIGINT(20)
fk_discourse_part BIGINT(20)
fk_discourse_part_interaction_type BIGINT(20)
fk_user BIGINT(20)
Indexes

discourse_part_interaction_type
id_discourse_part_interaction_type BIGINT(20)
created DATETIME
type VARCHAR(255)
version BIGINT(20)
Indexes

user
id_user BIGINT(20)
created DATETIME
end_time DATETIME
start_time DATETIME
version BIGINT(20)
country VARCHAR(255)
email VARCHAR(255)
ip VARCHAR(255)
language VARCHAR(255)
location VARCHAR(255)
realname VARCHAR(255)
username VARCHAR(255)
fk_annotation BIGINT(20)
fk_data_sources BIGINT(20)
Indexes

- Join
- Ready
- Leave

Data Source Management



data_source_instance

- id_data_source_instance BIGINT(20)
- created DATETIME
- version BIGINT(20)
- dataset_name VARCHAR(255)
- entity_source_id VARCHAR(255)
- entity_source_descriptor VARCHA...
- source_type VARCHAR(255)
- fk_sources BIGINT(20)

Indexes

data_sources

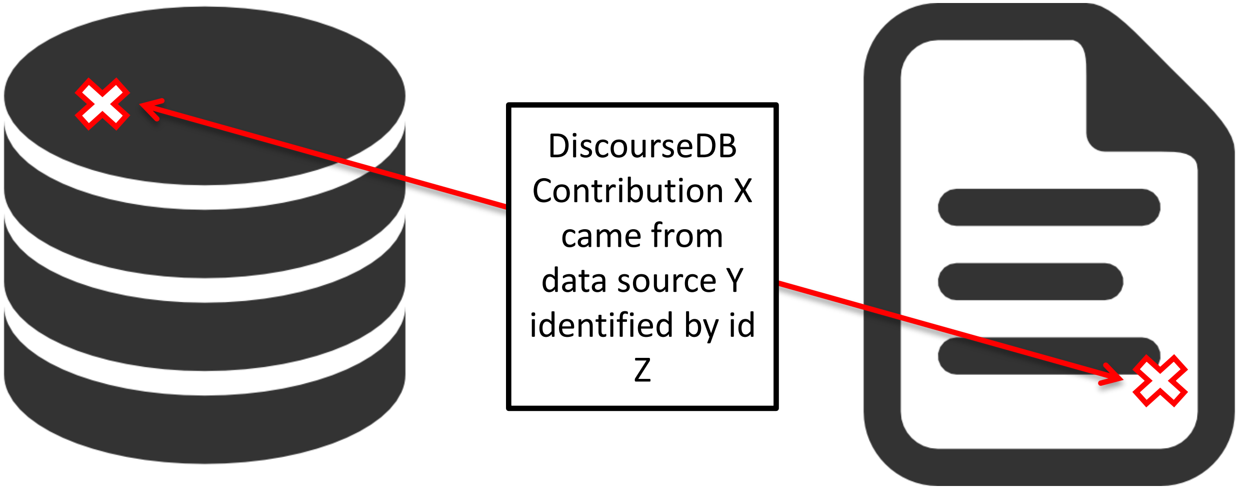
- id_data_sources BIGINT(20)
- created DATETIME
- version BIGINT(20)

Indexes

We want to keep track of where imported data came from

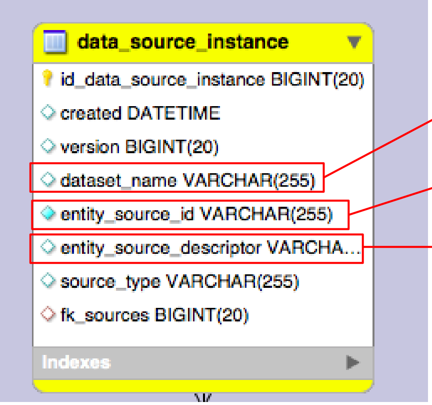
➔Necessary at import time in order to make sense of relational data

➔Helpful later to recover details that are not explicitly represented in DiscourseDB



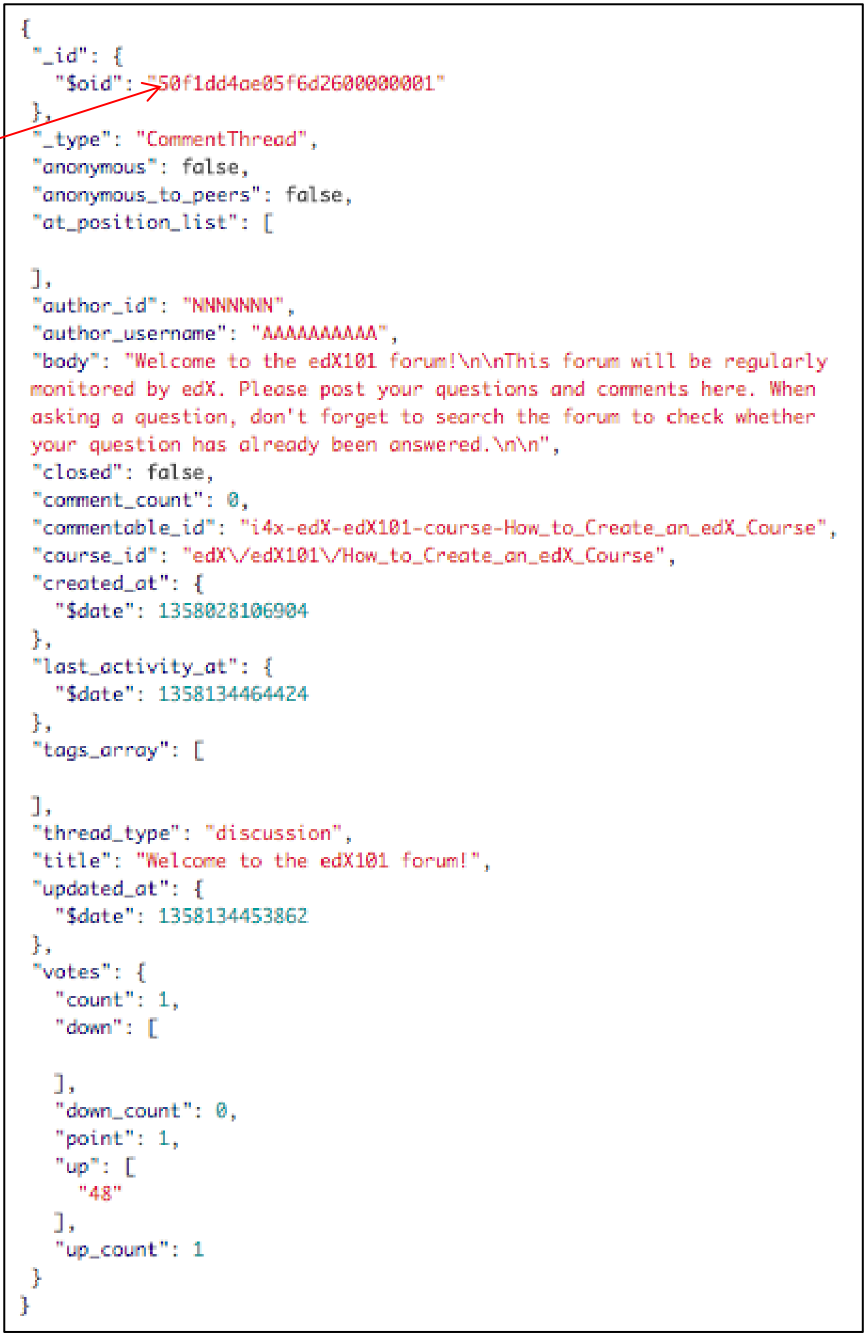
DiscourseDB Contribution X came from data source Y identified by id Z

<edx_test_forum.json>



data_source_instance
id_data_source_instance BIGINT(20)
created DATETIME
version BIGINT(20)
dataset_name VARCHAR(255)
entity_source_id VARCHAR(255)
entity_source_descriptor VARCHA...
source_type VARCHAR(255)
fk_sources BIGINT(20)

_id.$oid



```
{
  "_id": {
    "$oid": "50f1dd4ae05f6d2600000001"
  },
  "_type": "CommentThread",
  "anonymous": false,
  "anonymous_to_peers": false,
  "at_position_list": [

  ],
  "author_id": "NNNNNNN",
  "author_username": "AAAAAAAAAA",
  "body": "Welcome to the edX101 forum!\n\nThis forum will be regularly
monitored by edX. Please post your questions and comments here. When
asking a question, don't forget to search the forum to check whether
your question has already been answered.\n\n",
  "closed": false,
  "comment_count": 0,
  "commentable_id": "i4x-edX-edX101-course-How_to_Create_an_edX_Course",
  "course_id": "edX\/edX101\/How_to_Create_an_edX_Course",
  "created_at": {
    "$date": 1358028106904
  },
  "last_activity_at": {
    "$date": 1358134464424
  },
  "tags_array": [

  ],
  "thread_type": "discussion",
  "title": "Welcome to the edX101 forum!",
  "updated_at": {
    "$date": 1358134453862
  },
  "votes": {
    "count": 1,
    "down": [

    ],
    "down_count": 0,
    "point": 1,
    "up": [
      "48"
    ],
    "up_count": 1
  }
}
```

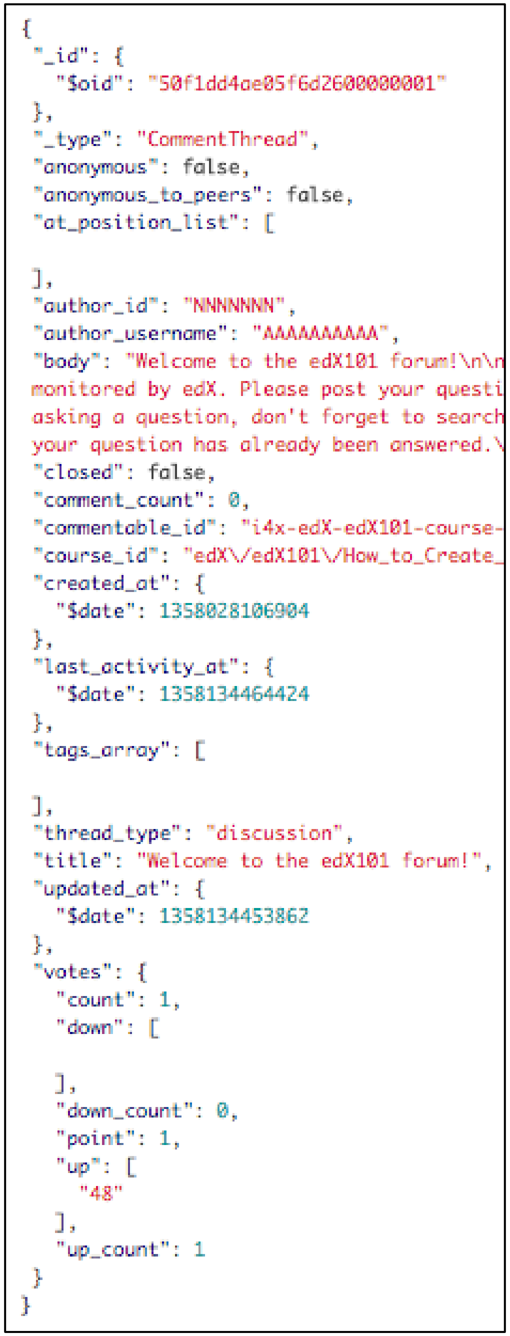
Questions?

Use Case: EdX Forum

- A course has a single forum
- A **CommentThread** represents the first level of interaction: a post that opens a new thread, often a student question of some sort.
- A **Comment** represents both the second and third levels of interaction: a response made directly to the conversation started by a CommentThread is a Comment. Any further contributions made to a specific response are also in Comment objects.

https://edx.readthedocs.org/en/latest/internal_data_formats/discussion_data.html

comment thread



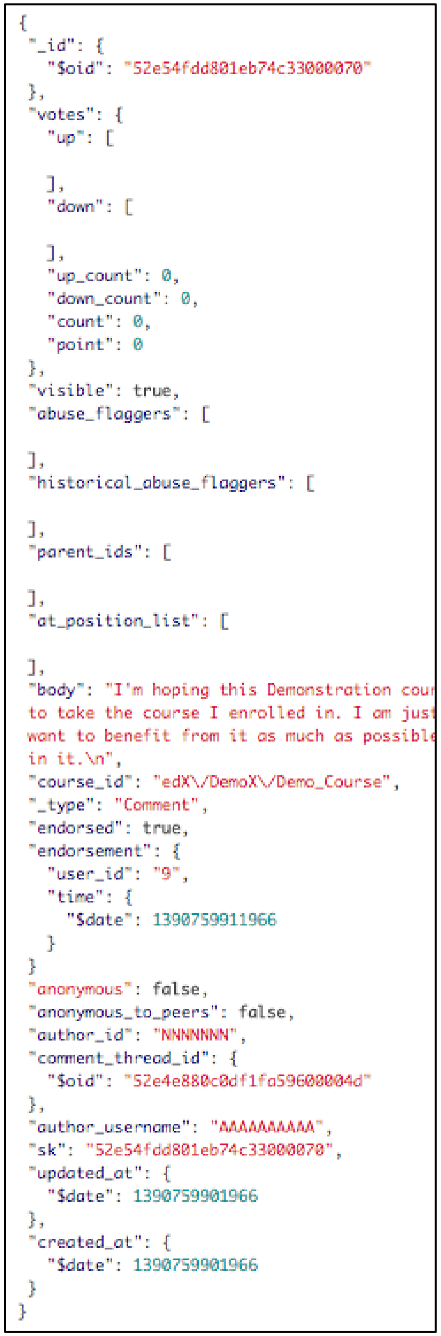
```
{
 "_id": {
   "$oid": "50f1dd4ae05f6d2600000001"
 },
 "_type": "CommentThread",
 "anonymous": false,
 "anonymous_to_peers": false,
 "at_position_list": [

 ],
 "author_id": "NNNNNNN",
 "author_username": "AAAAAAAAAA",
 "body": "Welcome to the edX101 forum!\n\n
monitored by edX. Please post your questi
asking a question, don't forget to search
your question has already been answered.\
 "closed": false,
 "comment_count": 0,
 "commentable_id": "i4x-edX-edX101-course-
 "course_id": "edX\/edX101\/How_to_Create_
 "created_at": {
   "$date": 1358028106904
 },
 "last_activity_at": {
   "$date": 1358134464424
 },
 "tags_array": [

 ],
 "thread_type": "discussion",
 "title": "Welcome to the edX101 forum!",
 "updated_at": {
   "$date": 1358134453862
 },
 "votes": {
   "count": 1,
   "down": [

   ],
   "down_count": 0,
   "point": 1,
   "up": [
     "48"
   ],
   "up_count": 1
 }
}
```

comment



```
{
 "_id": {
   "$oid": "52e54fdd801eb74c33000070"
 },
 "votes": {
   "up": [

   ],
   "down": [

   ],
   "up_count": 0,
   "down_count": 0,
   "count": 0,
   "point": 0
 },
 "visible": true,
 "abuse_flaggers": [

 ],
 "historical_abuse_flaggers": [

 ],
 "parent_ids": [

 ],
 "at_position_list": [

 ],
 "body": "I'm hoping this Demonstration cour
 to take the course I enrolled in. I am just
 want to benefit from it as much as possible
 in it.\n",
 "course_id": "edX\/DemoX\/Demo_Course",
 "_type": "Comment",
 "endorsed": true,
 "endorsement": {
   "user_id": "9",
   "time": {
     "$date": 1390759911966
   }
 }
 "anonymous": false,
 "anonymous_to_peers": false,
 "author_id": "NNNNNNN",
 "comment_thread_id": {
   "$oid": "52e4e880c0df1fa59600004d"
 },
 "author_username": "AAAAAAAAAA",
 "sk": "52e54fdd801eb74c33000070",
 "updated_at": {
   "$date": 1390759901966
 },
 "created_at": {
   "$date": 1390759901966
 }
}
```

Use Case: EdX Forum

- A course has a single forum
- A **Com**
  represe
  interac
  a new t
  questic
- A **Com**
  the sec
  interac
  directly
  started
  is a Co
  contrib
  specifi
  Comm

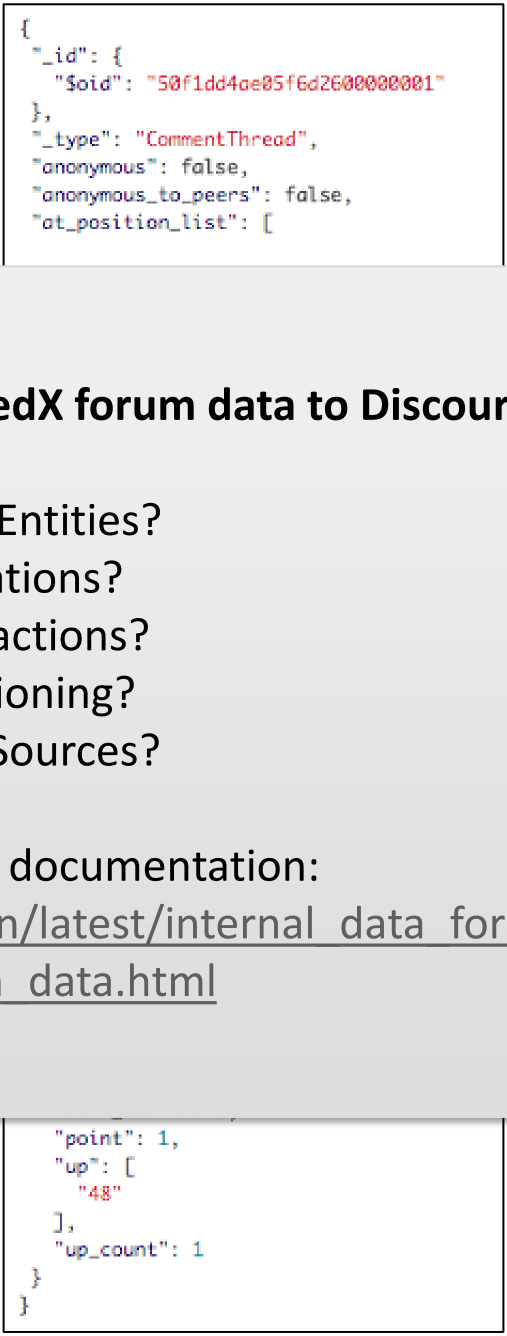
How would we translate the edX forum data to DiscourseDB?

Main Entities?
Relations?
Interactions?
Versioning?
Data Sources?

Forum format documentation:
https://edx.readthedocs.org/en/latest/internal_data_formats/discussion_data.html

https://edx.readthedocs.org/en/latest/internal_data_formats/discussion_data.html

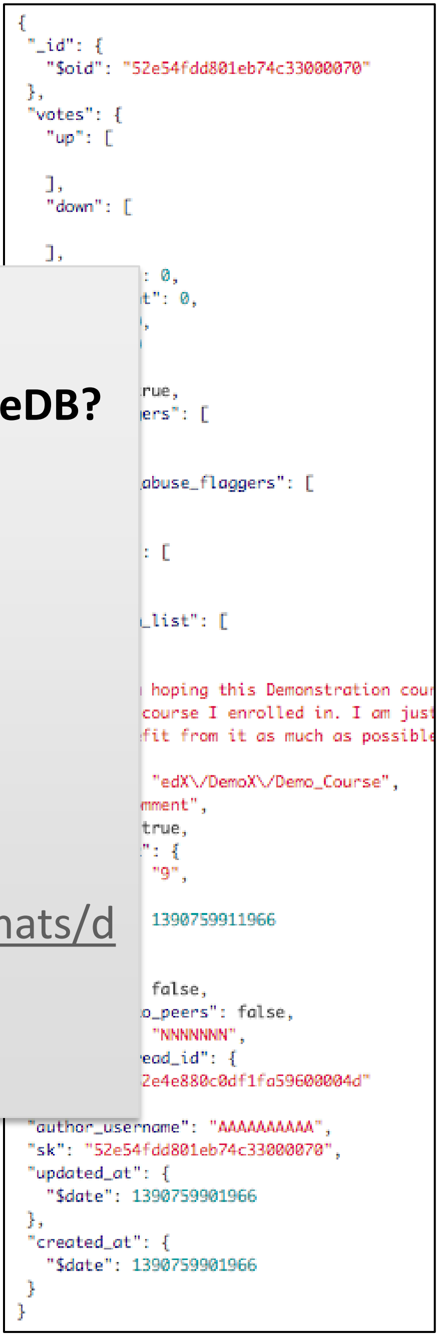
comment thread



```
{
 "_id": {
   "$oid": "50f1dd4ae05f6d2600000001"
 },
 "_type": "CommentThread",
 "anonymous": false,
 "anonymous_to_peers": false,
 "at_position_list": [

   "point": 1,
   "up": [
     "48"
   ],
   "up_count": 1
  }
}
```

comment



```
{
 "_id": {
   "$oid": "52e54fdd801eb74c33000070"
 },
 "votes": {
   "up": [

   ],
   "down": [

   ],
              : 0,
              t": 0,
              ,

              rue,
              ers": [

              abuse_flaggers": [

              : [

              _list": [

              hoping this Demonstration cour
              course I enrolled in. I am just
              fit from it as much as possible
               "edX\/DemoX\/Demo_Course",
              mment",
              true,
              ": {
               "9",

               1390759911966

               false,
              o_peers": false,
               "NNNNNNN",
              ead_id": {
              2e4e880c0df1fa59600004d"

 "author_username": "AAAAAAAAAA",
 "sk": "52e54fdd801eb74c33000070",
 "updated_at": {
   "$date": 1390759901966
 },
 "created_at": {
   "$date": 1390759901966
 }
}
```