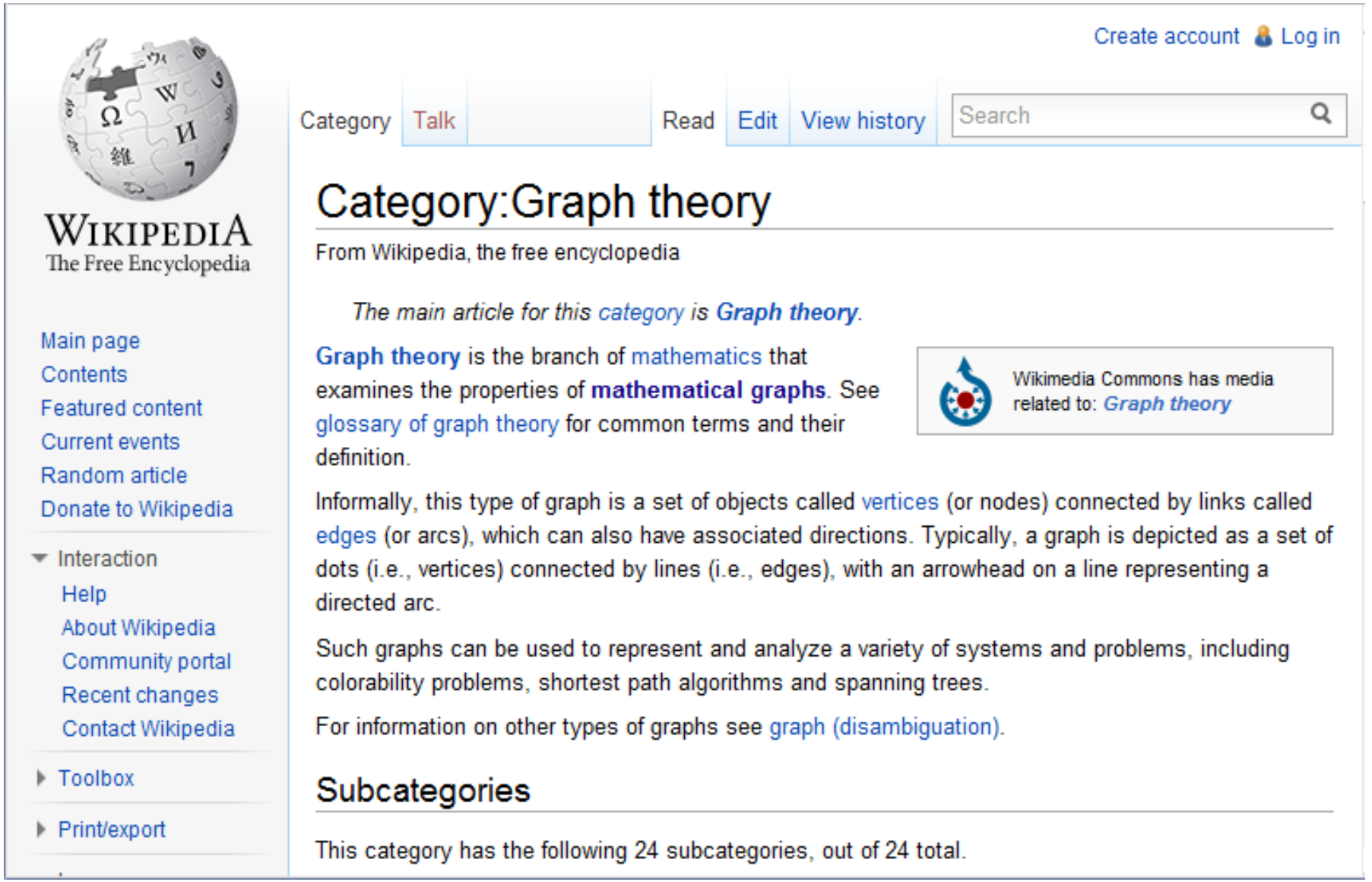
Исследование графа категорий английской версии Wikipedia

Александр Шкотин
ashkotin@acm.org

Государственный Геологический Музей РАН
Москва

Пример категорной страницы-1



Create account Log in

Category Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Category:Graph theory

From Wikipedia, the free encyclopedia

*The main article for this category is **Graph theory**.*

Graph theory is the branch of mathematics that examines the properties of **mathematical graphs**. See glossary of graph theory for common terms and their definition.

Wikimedia Commons has media related to: *Graph theory*

Informally, this type of graph is a set of objects called vertices (or nodes) connected by links called edges (or arcs), which can also have associated directions. Typically, a graph is depicted as a set of dots (i.e., vertices) connected by lines (i.e., edges), with an arrowhead on a line representing a directed arc.

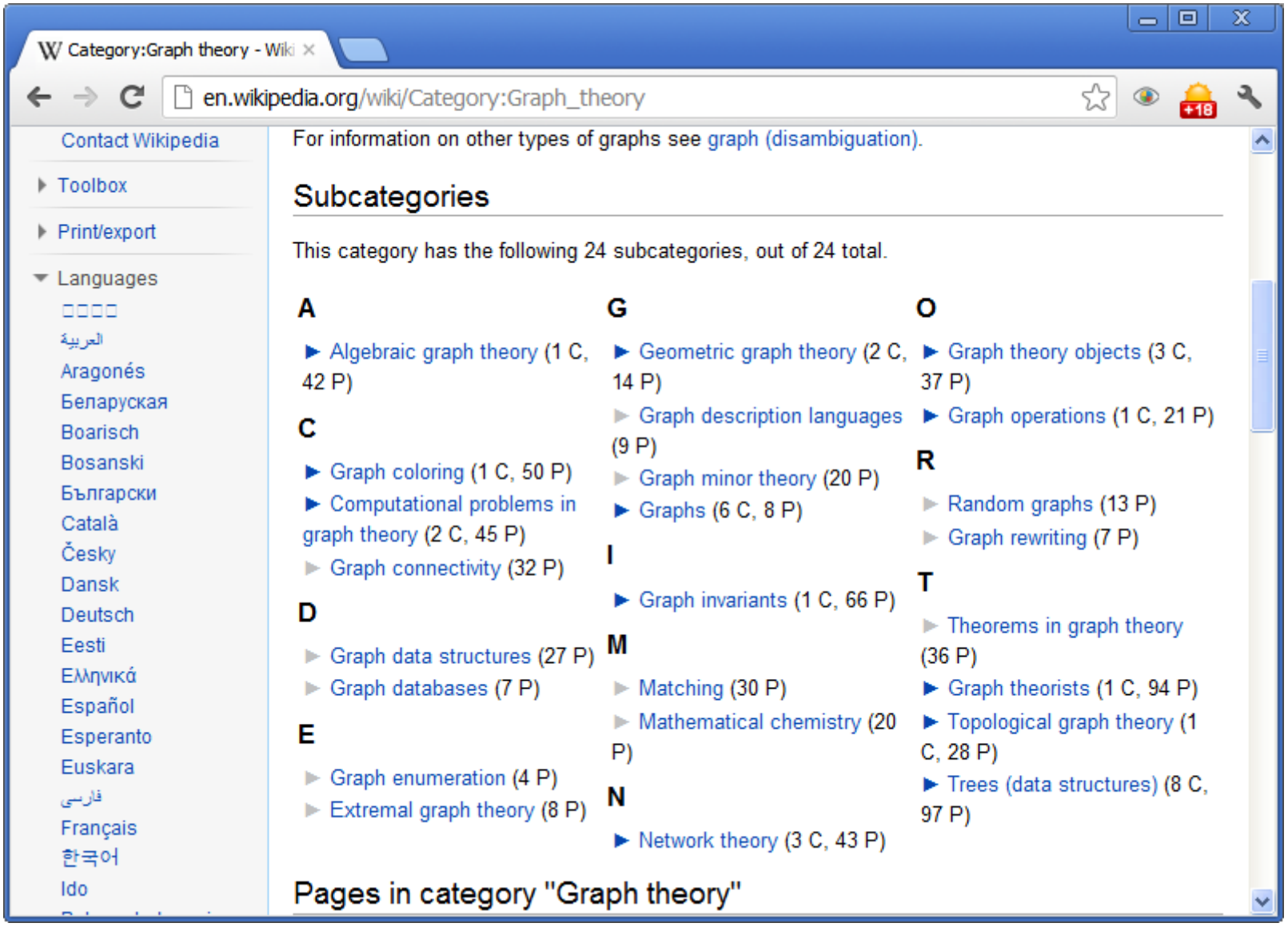
Such graphs can be used to represent and analyze a variety of systems and problems, including colorability problems, shortest path algorithms and spanning trees.

For information on other types of graphs see graph (disambiguation).

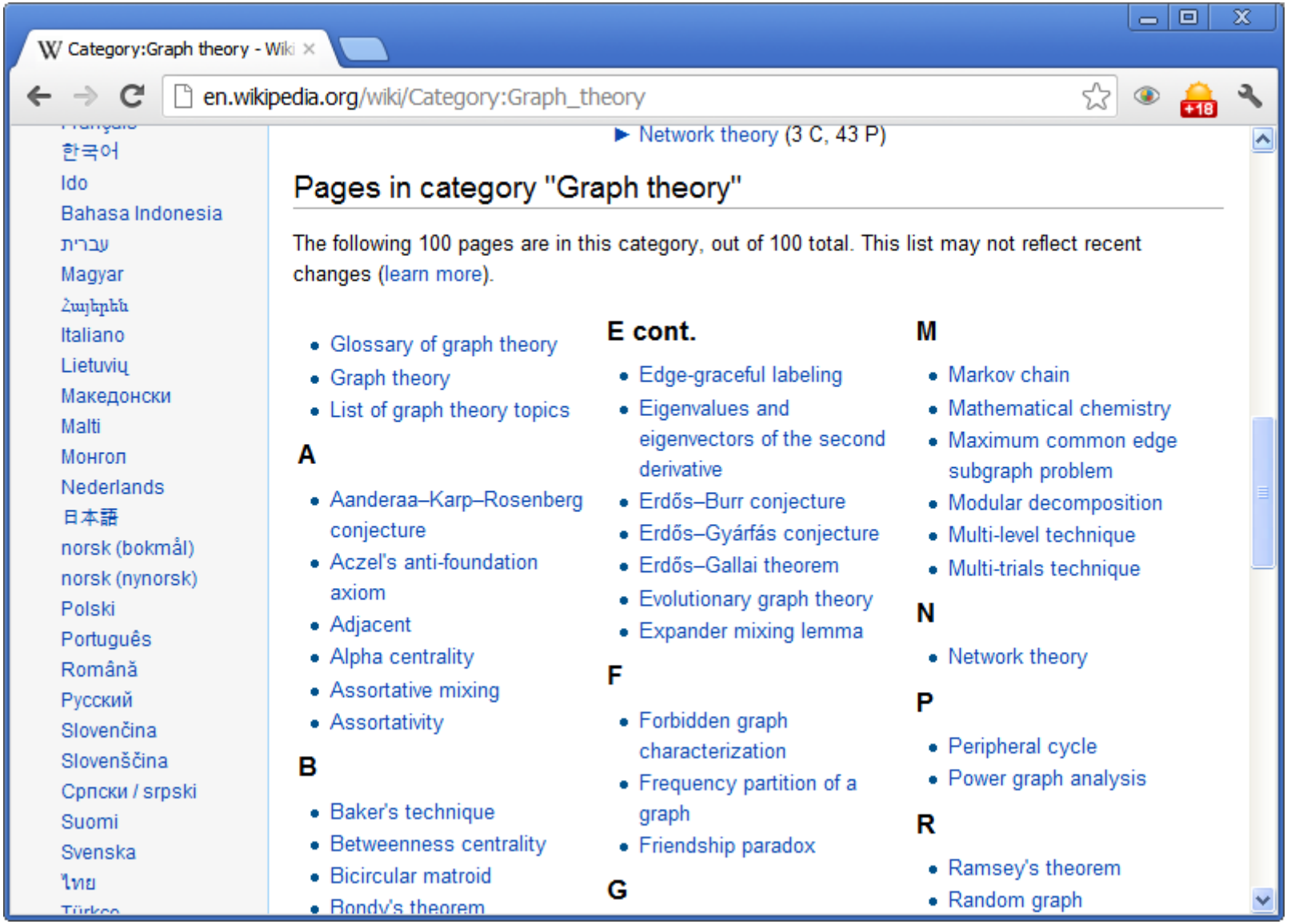
Subcategories

This category has the following 24 subcategories, out of 24 total.

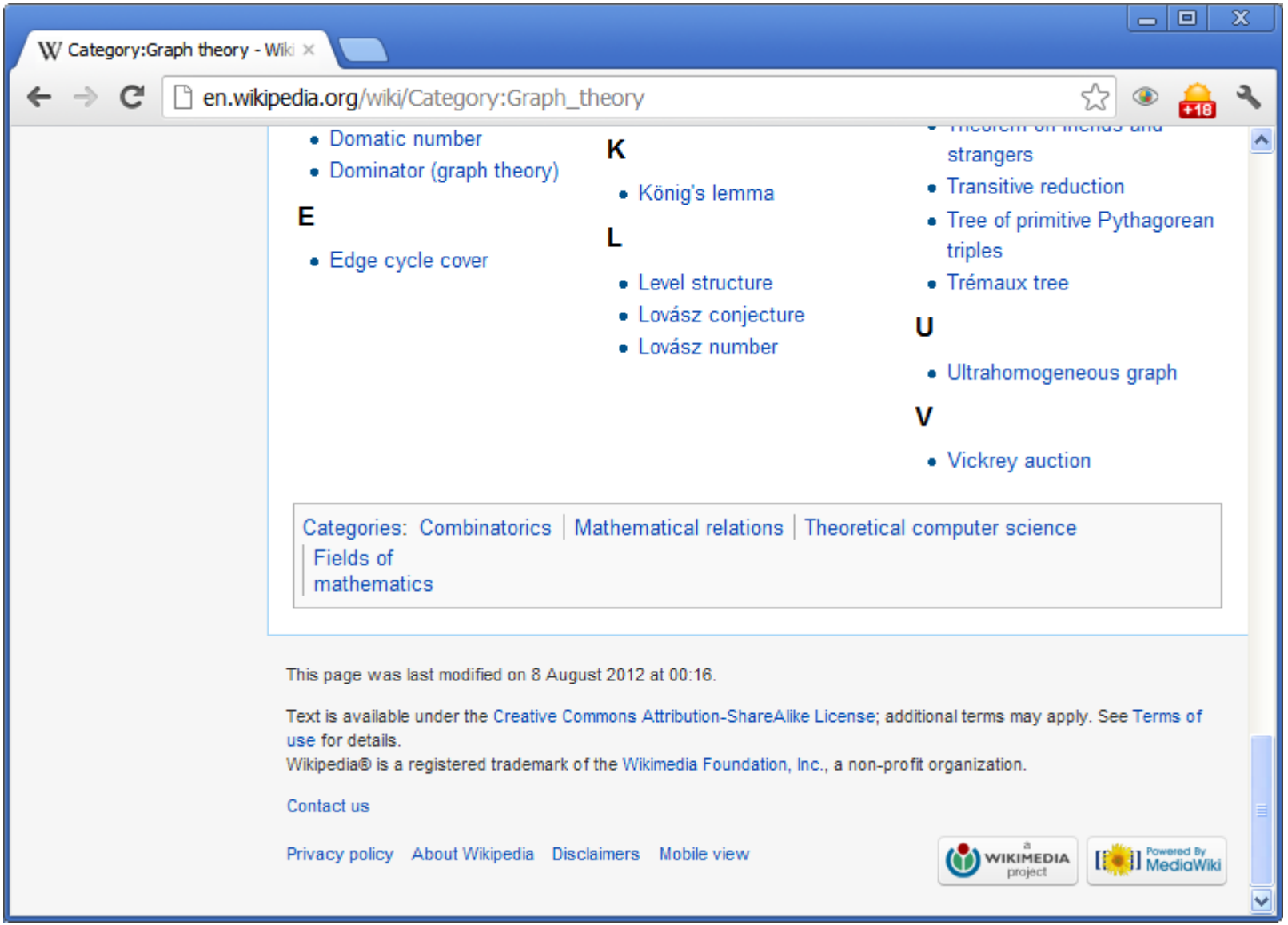
Пример категорной страницы-2



Пример категорной страницы-3



Пример категорной страницы-4



Граф категорий Википедии (ГКВ)

ГКВ есть орграф каждый узел которого взаимно-однозначно соответствует категорной странице и помечен её номером.

Дуга из узла N1 в узел N2 идёт тогда и только тогда, когда страница с номером N1 есть под-категория страницы с номером N2.

Исходные данные («дамп») [Антон Коршунов, ИСП РАН]:

- N1 N2 flag
- N title

flag: 0 - N1 - простая страница, 1 - N1 - категорная.

ГКВ. Характеристики

Количество узлов - 593796 (изолированных - 26272).

Количество стрелок - 1221133. Петель нет.

Количество связных компонент - 1987. Количество узлов в 10-и самых больших компонентах: 561636, 210, 36, 29, 27, 20, 19, 19, 17, 16.

Источников - 345597. Промежуточных узлов - 210160. Стоков - 11767. Это в основном «точки роста».

"Category:World War II" - 85 исходящих.

"Category:Albums by artist" - 12625 входящих.

Путь: Category:Anastacia songs → Category:Music, **154** узла.

Самый длинный - **294** узла.

ГКВ в целом

Есть ор-циклы [1], с.9 - чего быть не должно.

Ядро - объединение ор-циклов графа и ор-путей между циклами.

Мантия - дополнительная часть графа.

Связующие стрелки - между мантией и ядром. Часть идёт из ядра в мантию (591), а часть - из мантии в ядро (210514).

Ядро получается итеративным удалением источников и стоков.

Ядро ГКВ

Стрелок в ядре - 38538. Узлов - 13545.

Связные компоненты: одна большая - 13507 узлов. И ещё 19 пар узлов. Таким образом большинство связных компонент ГКВ ядер не имеет.

Чисто ядерная связная компонента ГКВ:

Category:Wikipedia sockpuppets of ShantanuSingh198

Category:Suspected Wikipedia sockpuppets of ShantanuSingh19

Сильно связные компоненты ядра-1

В ядре нас интересует зацикливание отношения под-категория - категория. Два подхода:

- общий - применить алгоритм поиска сильно связных компонент (ССК);
- частный - найти так называемые "линзы" - два узла ссылающиеся друг на друга (как под-категория - категория).

Математически цикл утверждает эквивалентность соответствующих терминов, т.е. синонимию, что в принципе возможно.

Чтобы получить состав сильно связных компонент ядра была использована программа Pajek [2].

Сильно связные компоненты ядра-2

Количество ССК - 457.

Узлов не входящих в ССК — 7646.

Гигантская ССК - 3967 узлов.

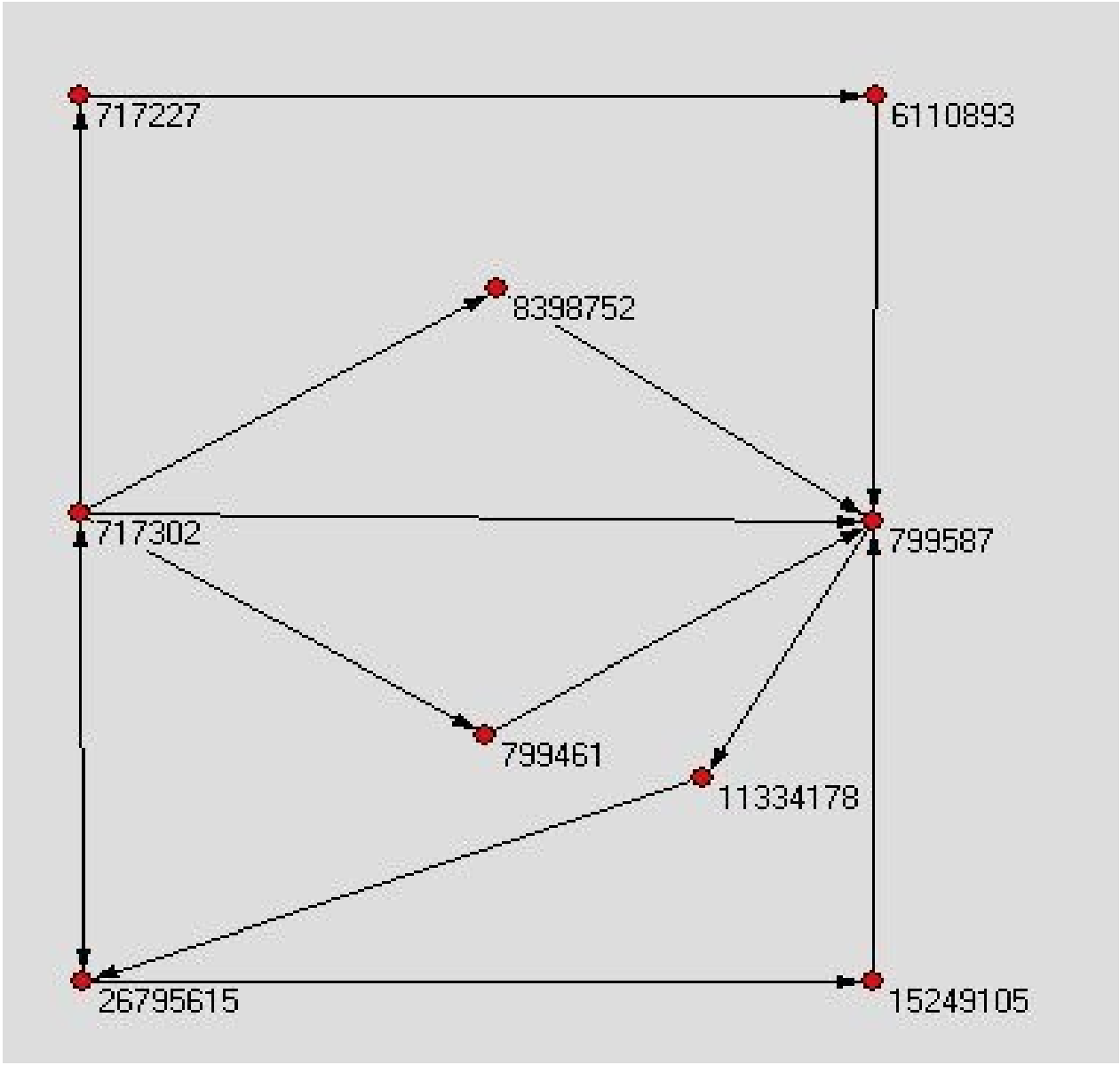
Количество узлов в следующих по величине ССК:

96, 95, 68, 52(7), 51(3), 18.

Линз - 1269.

Рассмотрим компоненту №41 у которой всего 9 узлов.

ССК 41. Рисунок графа компоненты



ССК 41. Заголовки узлов

717227	Category:Orthodox rabbis
717302	Category:Talmud rabbis
799461	Category:Mishnah
799587	Category:Talmud
6110893	Category:Talmudists
8398752	Category:Talmud people
11334178	Category:Rabbinic literature
15249105	Category:Talmud concepts and terminology
26795615	Category:Chazal

Мантия ГКВ отдельно

Изолировавшихся источников - 14421, стоков - 60.

Не изолированных стоков - 11707.

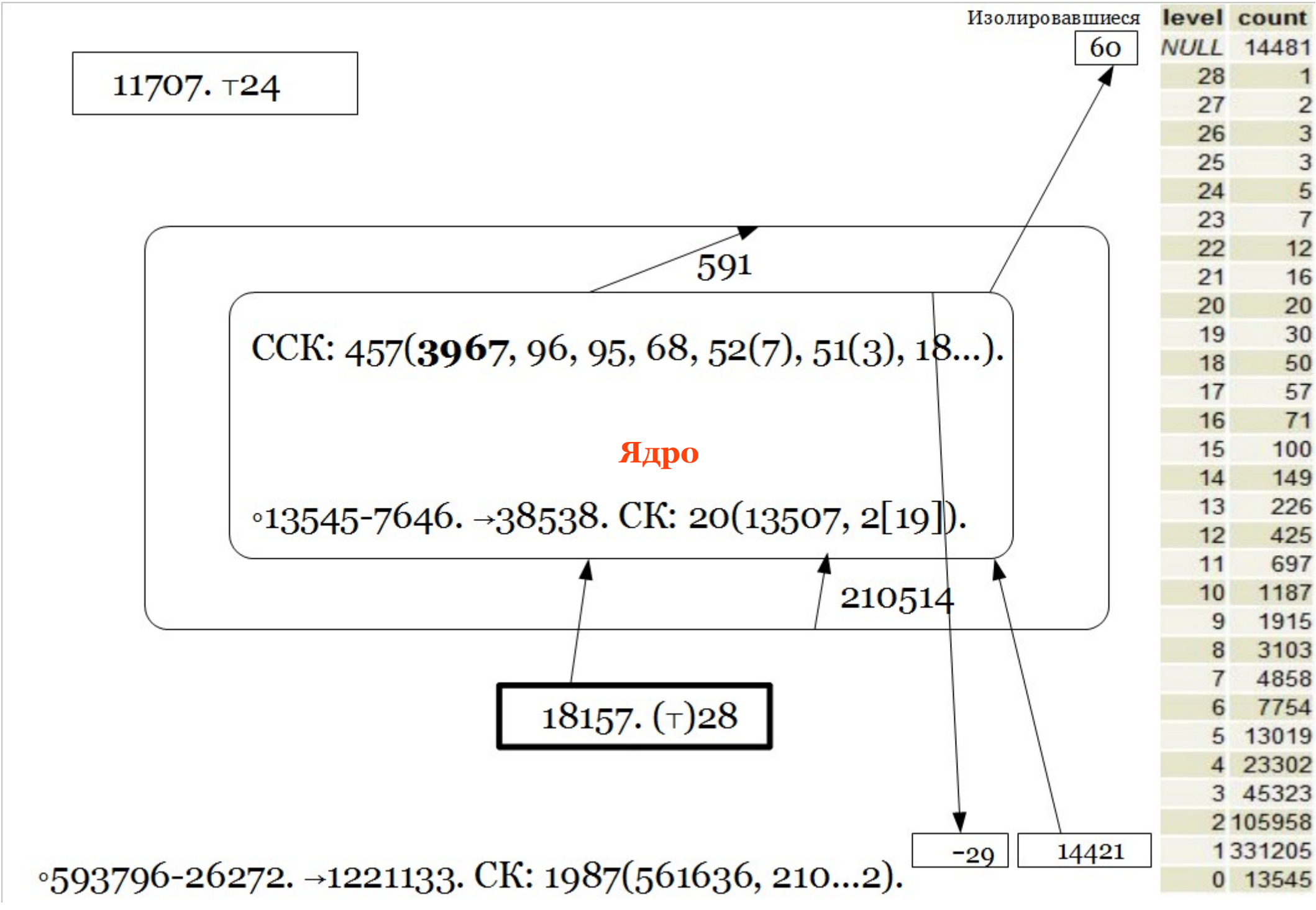
Ложных стоков - 18157.

Максимальная высота у ложного стока «Category:Creation myths».

Максимальная высота настоящего стока - 24.

В таблице: level = NULL - количество изолированных узлов мантии, level = 0 - количество узлов в ядре.

level	count
NULL	14481
28	1
27	2
26	3
25	3
24	5
23	7
22	12
21	16
20	20
19	30
18	50
17	57
16	71
15	100
14	149
13	226
12	425
11	697
10	1187
9	1915
8	3103
7	4858
6	7754
5	13019
4	23302
3	45323
2	105958
1	331205
0	13545



level	count
NULL	14481
28	1
27	2
26	3
25	3
24	5
23	7
22	12
21	16
20	20
19	30
18	50
17	57
16	71
15	100
14	149
13	226
12	425
11	697
10	1187
9	1915
8	3103
7	4858
6	7754
5	13019
4	23302
3	45323
2	105958
1	331205
0	13545

Обсуждение

ГКВ должен быть ациклическим графом. Таким образом аномалии значительны.

Можно создать средства уведомления о логическом противоречии. Для начала можно разобраться с 1269 линзами.

Интересно исследовать длинные пути.

Многие отношения между категориями попавшие в sub-category of следует перенести в See also.

Важная задача - получить и проанализировать русский ГКВ.

Литература

1. Christian Bizer, Jens Lehmann, Georgi Kobilarov, Soren Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann. DBpedia - A Crystallization Point for the Web of Data. May 25 2009.

2. Batagelj V., Mrvar A. Pajek reference manual. Ljubljana, April 16, 2012.

3. Исследовательский отчёт "Исследование графа категорий английской версии Wikipedia": текст.

Спасибо за внимание!

Александр Шкотин

ashkotin@acm.org