



АНАЛИТИКА БОЛЬШИХ ДАННЫХ: ПРАКТИЧЕСКОЕ ПРИМЕНЕНИЕ

Кафедра «Анализ конкурентных систем»
ИНСТИТУТ МЕЖДУНАРОДНЫХ ОТНОШЕНИЙ МИФИ

Артамонов Алексей Анатольевич, к.т.н., доц.,
заведующий кафедрой «Анализ конкурентных систем» МИФИ

ТИПЫ ДАННЫХ

Структурированные – это данные, которые имеют формализованную структуру, например, база данных, где в искомой таблице строка представляет некий объект с характеристиками, представленными в соответствующих полях и заранее установленными типами данных.

Слабоструктурированные – это данные, которые имеют слабо формализованную структуру, например, текстовая информация, размеченная при помощи языков разметки, таких как HTML. Из-за вариативности визуального представления некоего объекта, исходный HTML код страницы информационного ресурса может кардинально различаться.

Неструктурированные – это данные, которые имеют неформализованную структуру, например, текст письма электронной почты, сообщение в социальной сети и т.д.

РАБОТА С BIG DATA

V7



VOLUME
(объем)



VELOCITY
(скорость)



VARIETY
(разнообразие)



VARIABILITY
(изменчивость)



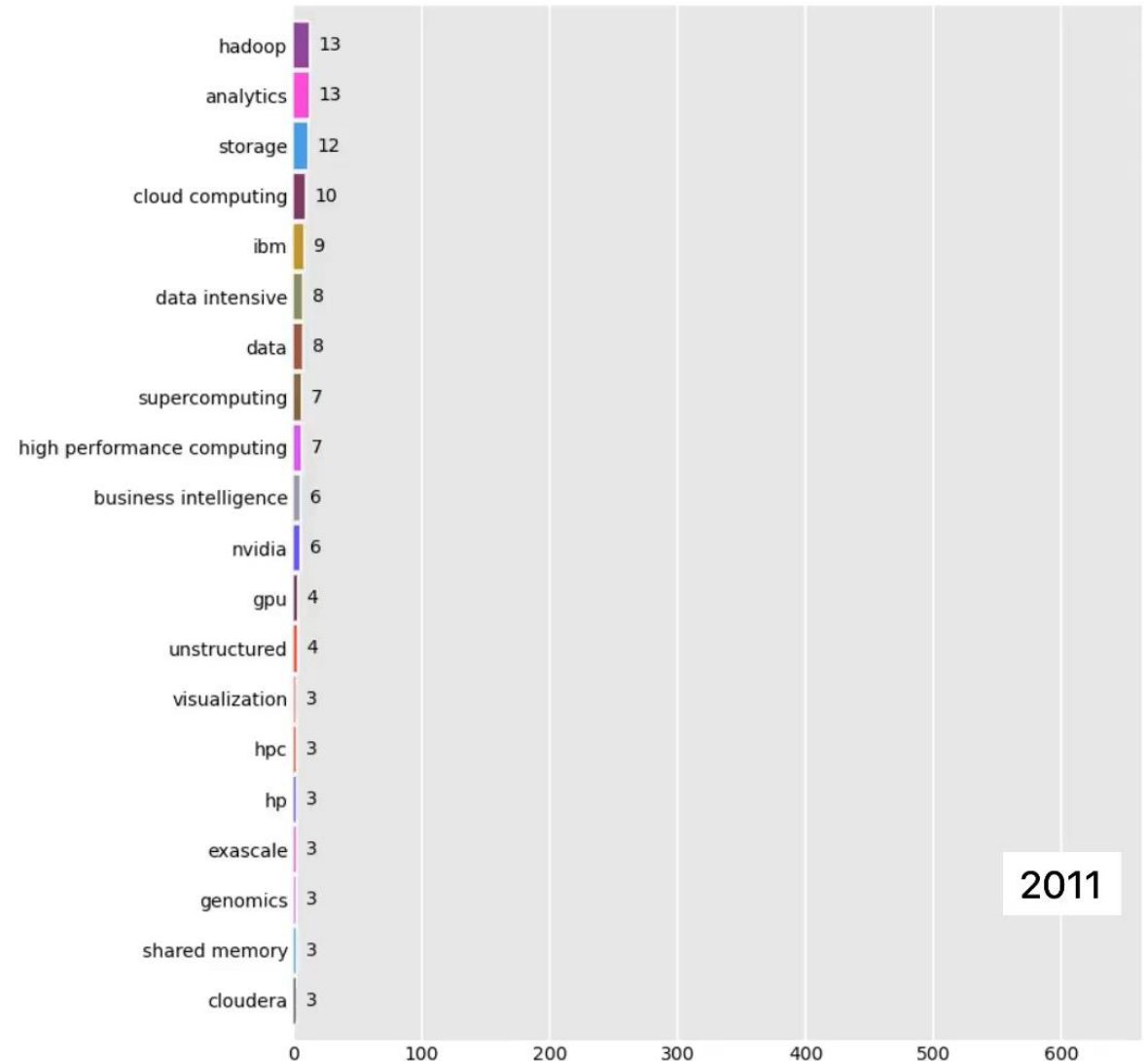
VALUE
(ценность)



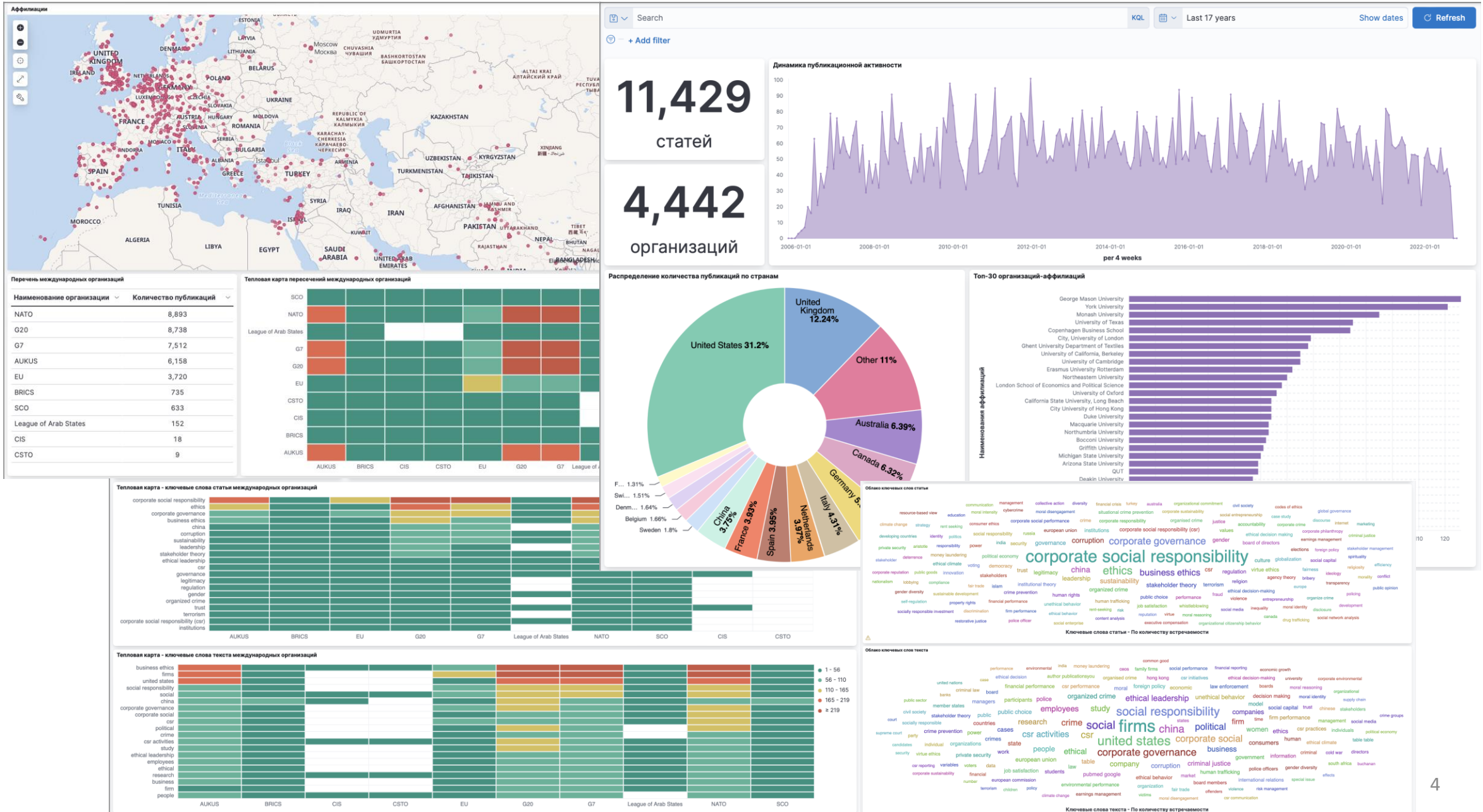
VERACITY
(достоверность)



VISUALIZATION
(визуализация)

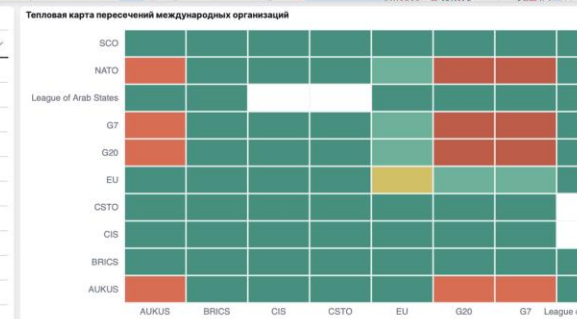


Озеро данных по публикациям в области Финансовой безопасности

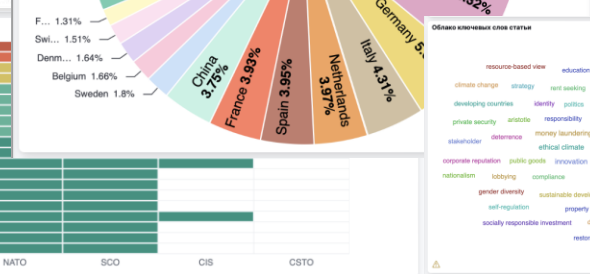
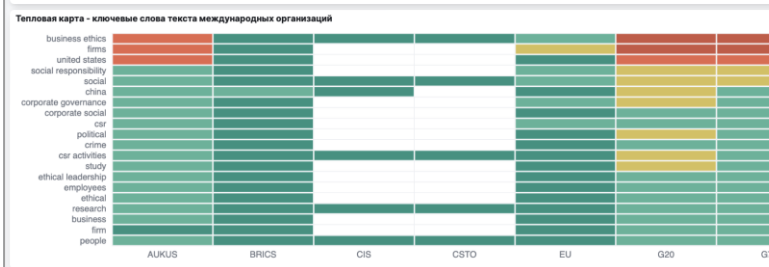
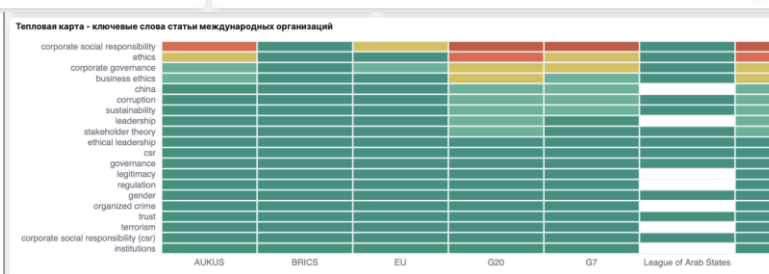
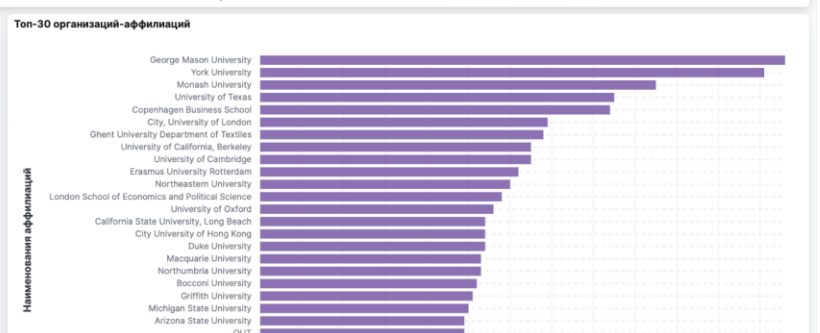
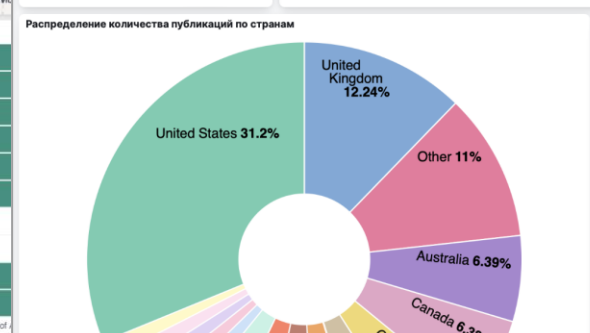
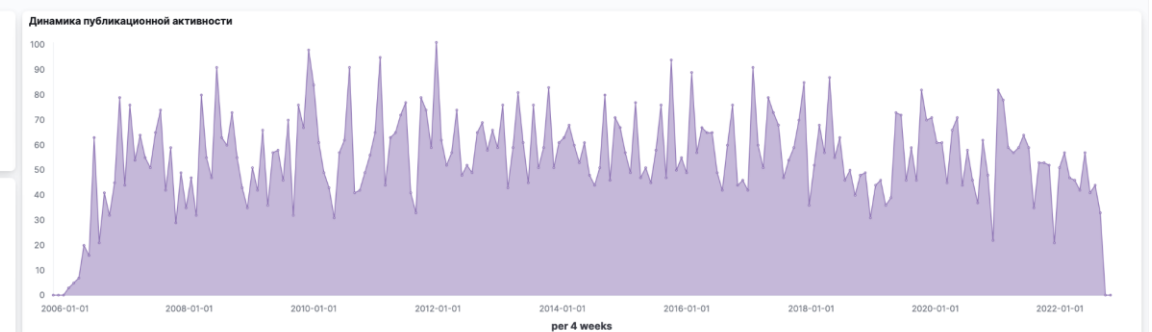


Перечень международных организаций

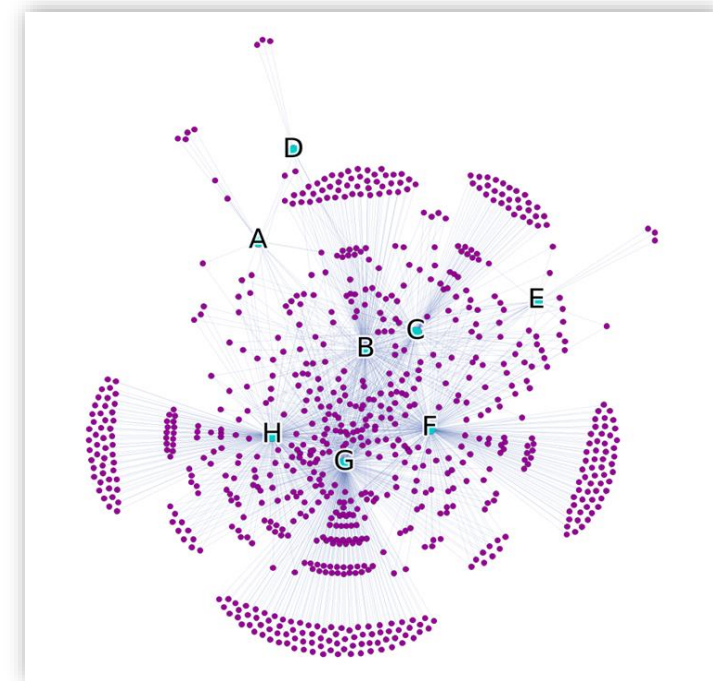
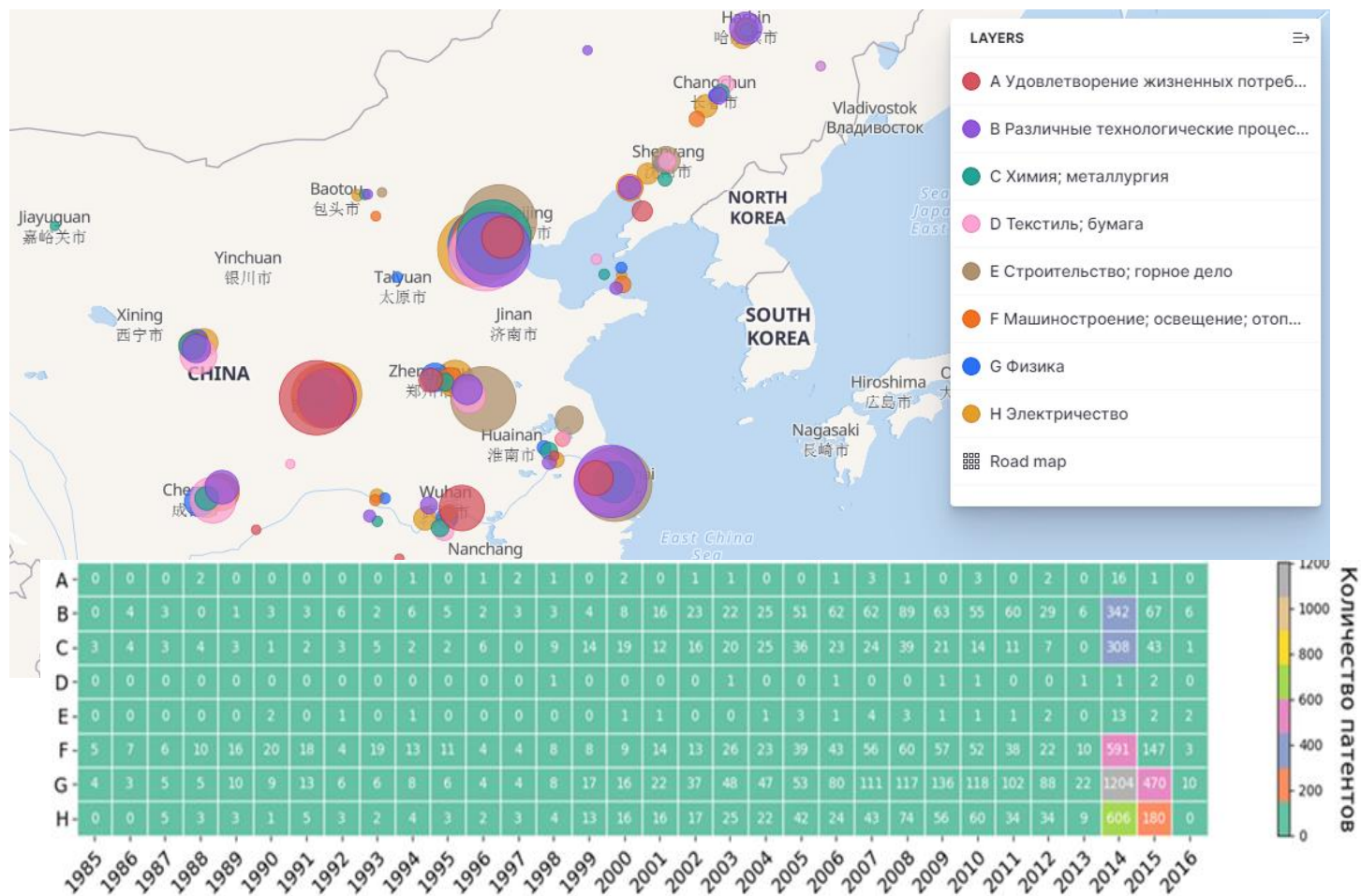
Наименование организации	Количество публикаций
NATO	8,893
G20	8,738
G7	7,512
AUKUS	6,158
EU	3,720
BRICS	735
SCO	633
League of Arab States	152
CIS	18
CSTO	9



11,429 статей
4,442 организаций



Исследование научно-технических направлений развития КНР по патентной информации (на китайском языке)

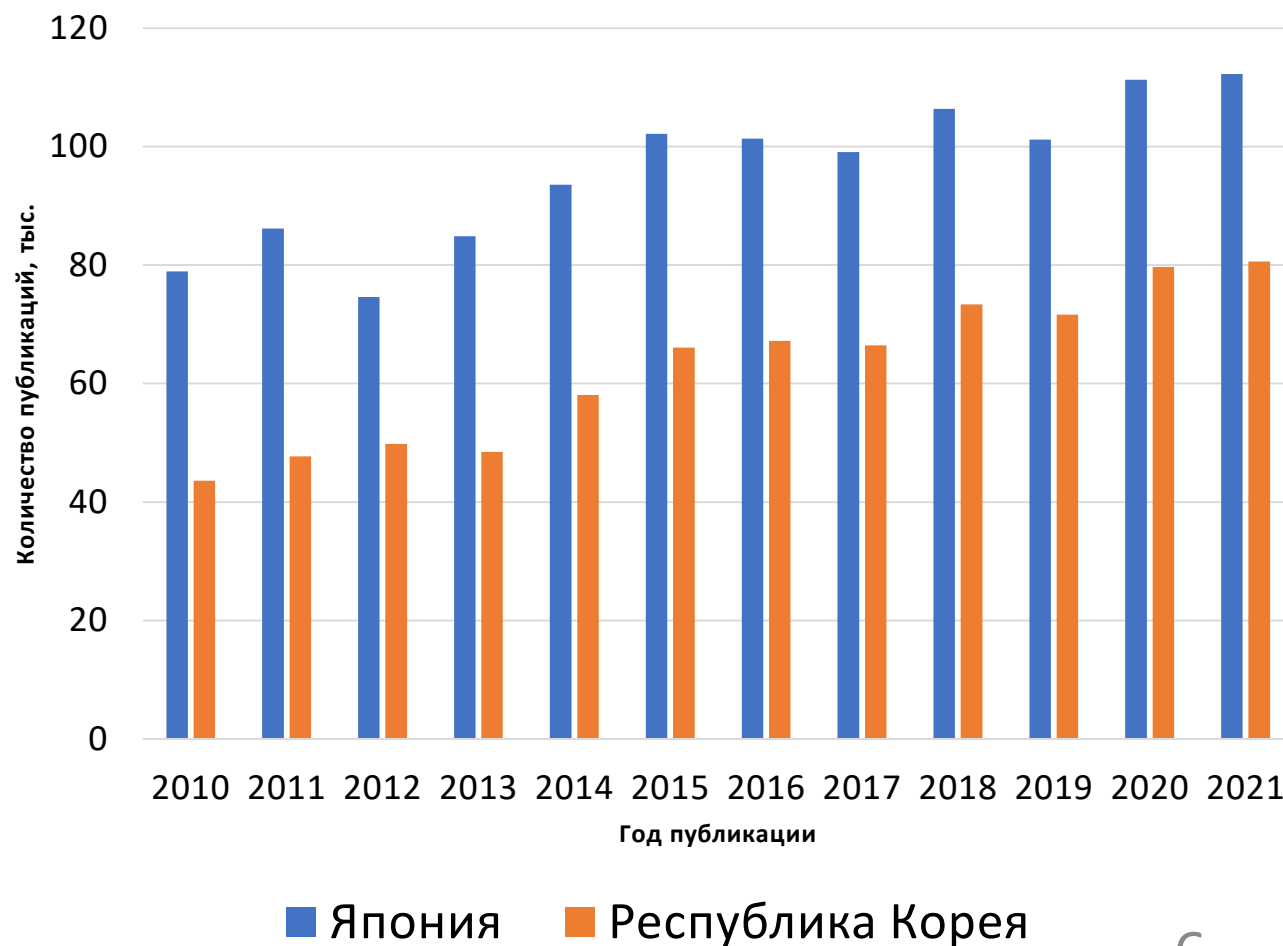


Формирование набора данных



	Япония	Республика Корея
Временной период	2010 – 2021 гг.	2010 – 2021 гг.
Количество научных публикаций	1 050 000	753 000
Объем данных	24.3 Гб	16.3 Гб

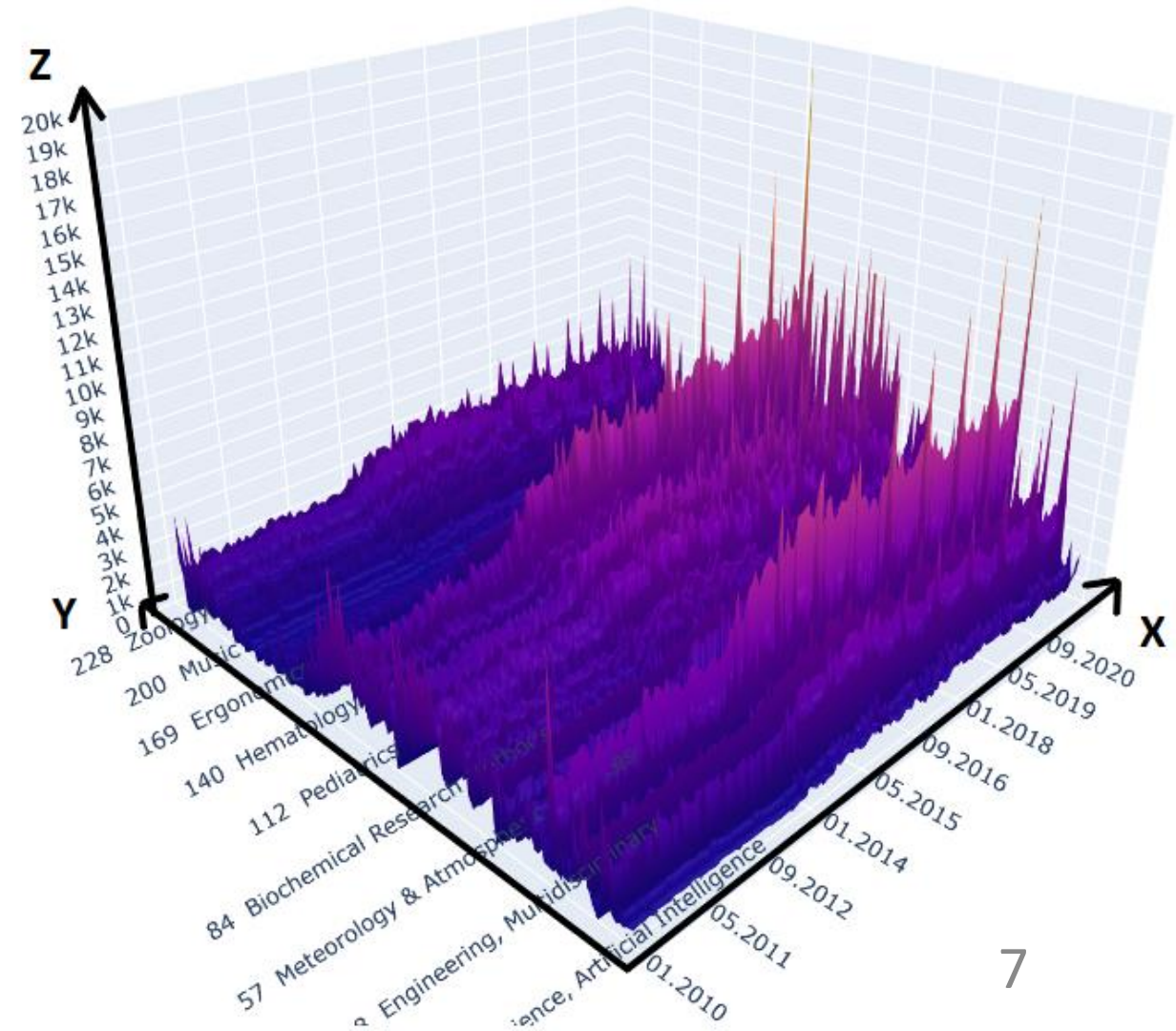
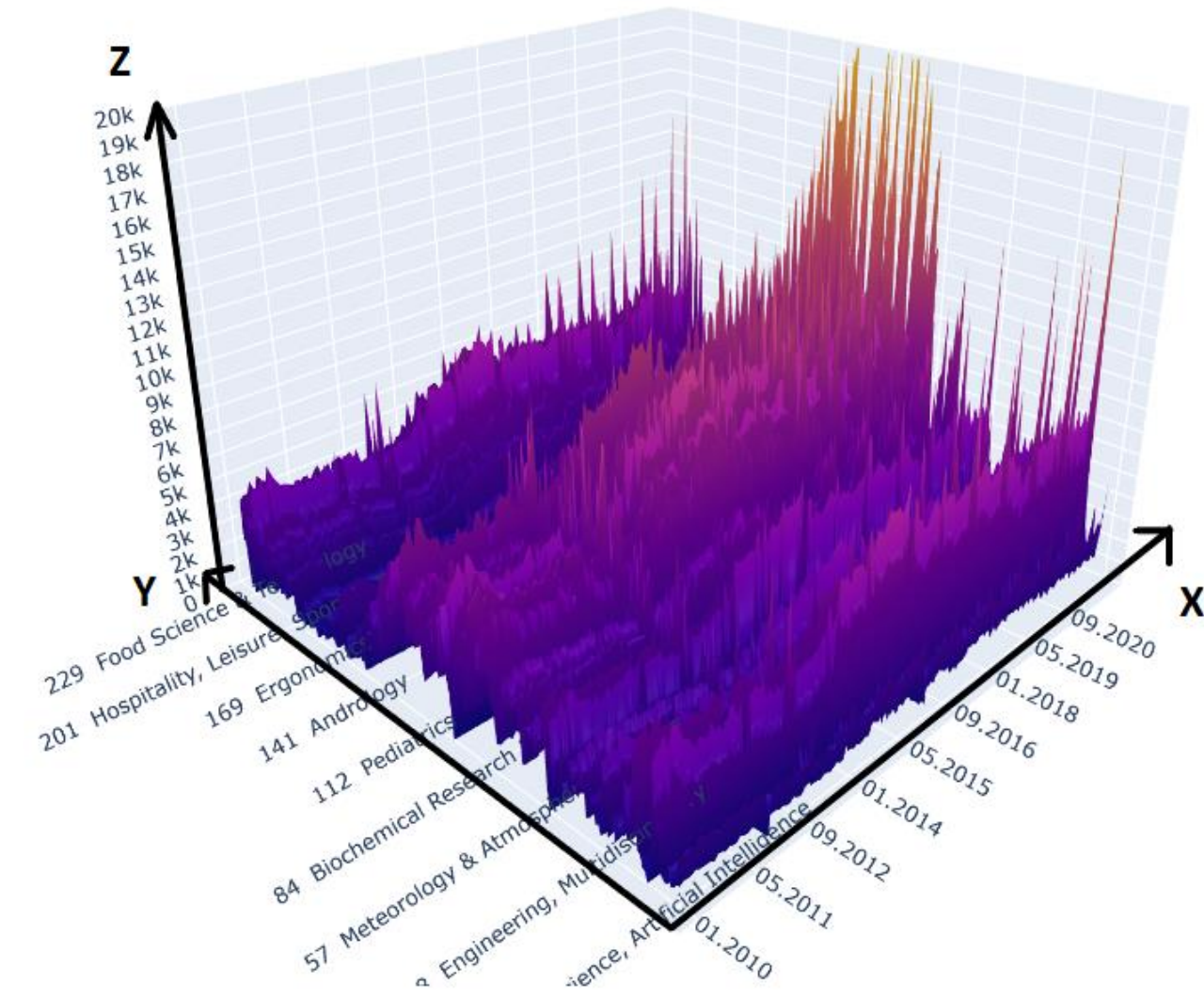
Распределение по годам научных публикаций
(Япония и Республика Корея)



Научно-технологический ландшафт страны

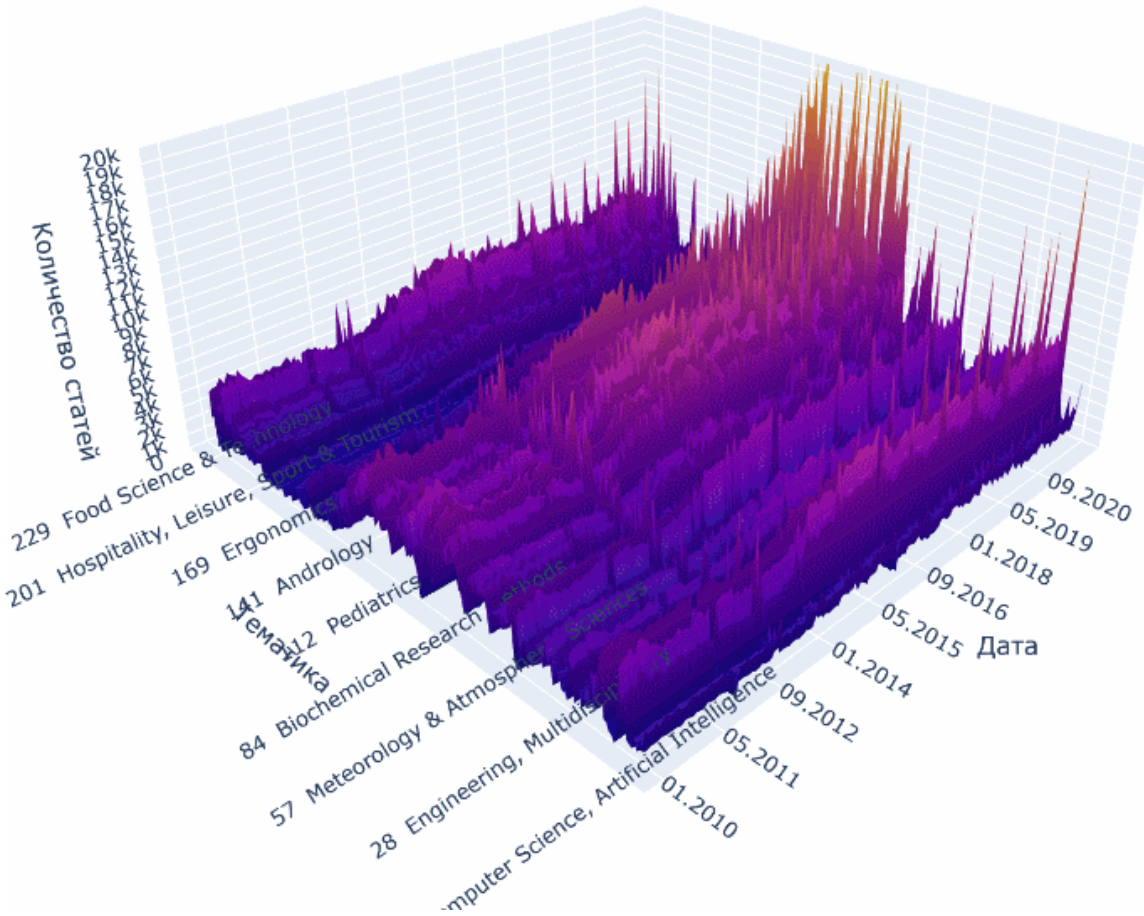
НТЛ страны (Япония)

НТЛ страны (Республика Корея)

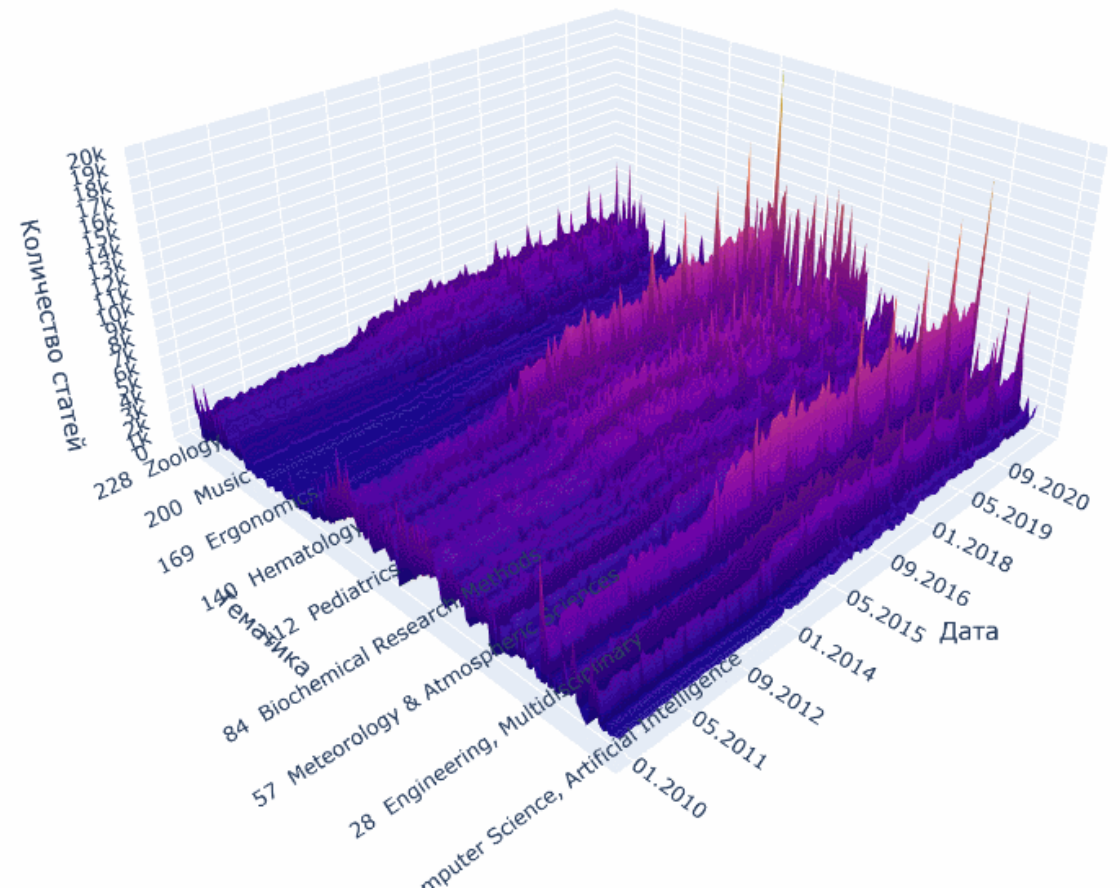


Научно-технологический ландшафт страны

НТЛ страны (Япония)

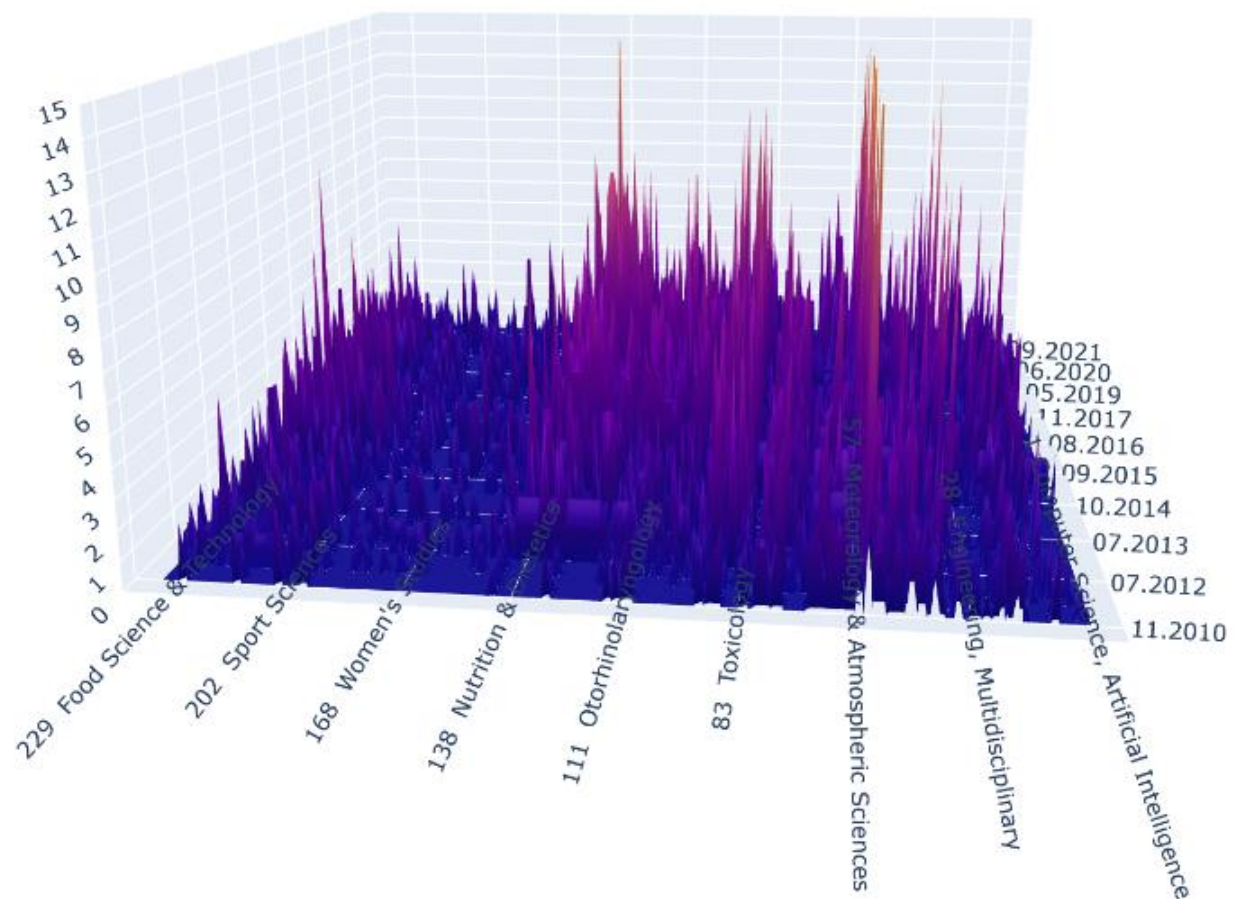


НТЛ страны (Республика Корея)

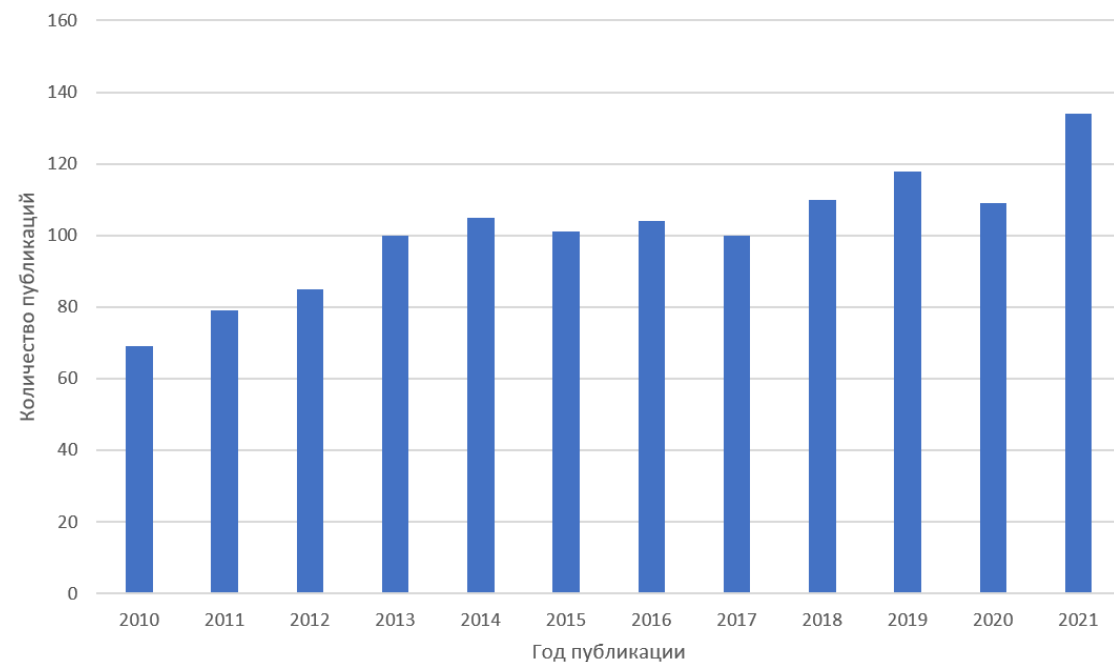


Научно-технологический ландшафт научной области

НТЛ области "FLASH терапия"
(Япония)



Распределение по годам научных
публикаций по тематике "FLASH терапия"



ВЫЯВЛЕНИЕ СВЯЗЕ МЕЖДУ ОБЪЕКТАМИ



Filters

Choose countries ▾

YYYY 📅 YYYY 📅

Choose areas ▾

Key word

Enter keywords separated by semicolon

Apply +

Reset graph ↻

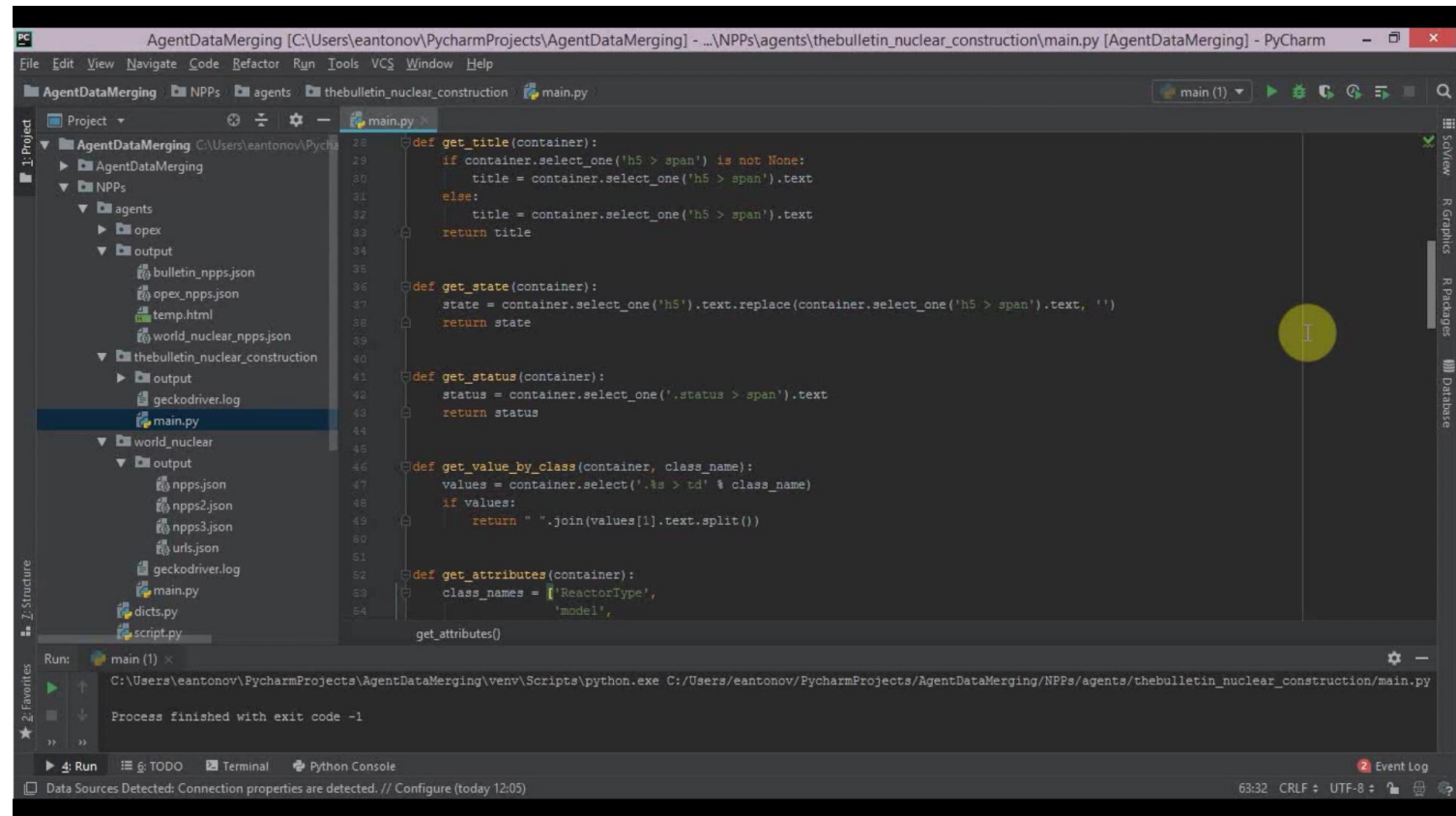
«

Node Information

»

АВТОМАТИЗАЦИЯ ДЕЯТЕЛЬНОСТИ

Разработка информационных агентов на языке программирования Python для формирования базы данных атомных станций



The screenshot displays the PyCharm IDE interface. The main editor window shows a Python script named `main.py` with the following code:

```
28 def get_title(container):
29     if container.select_one('h5 > span') is not None:
30         title = container.select_one('h5 > span').text
31     else:
32         title = container.select_one('h5 > span').text
33     return title
34
35
36 def get_state(container):
37     state = container.select_one('h5').text.replace(container.select_one('h5 > span').text, '')
38     return state
39
40
41 def get_status(container):
42     status = container.select_one('.status > span').text
43     return status
44
45
46 def get_value_by_class(container, class_name):
47     values = container.select('%s > td' % class_name)
48     if values:
49         return " ".join(values[1].text.split())
50
51
52 def get_attributes(container):
53     class_names = ['ReactorType',
54                   'model',
55                   'get_attributes()']
```

The left sidebar shows the project structure with the following folders and files:

- AgentDataMerging
- NPPs
 - agents
 - opex
 - output
 - bulletin_npps.json
 - opex_npps.json
 - temp.html
 - world_nuclear_npps.json
 - thebulletin_nuclear_construction
 - output
 - geckodriver.log
 - main.py
 - world_nuclear
 - output
 - npps.json
 - npps2.json
 - npps3.json
 - urls.json
 - geckodriver.log
 - main.py
 - dicts.py
 - script.py

The bottom status bar indicates the process finished with exit code -1.



СПАСИБО ЗА ВНИМАНИЕ!