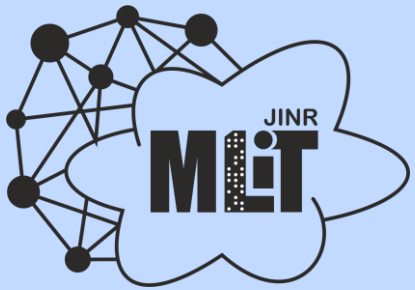




# Программа автоматизированного выделения значений и единиц измерения физических величин из полнотекстовых материалов

**Выполнила:**
Хвостова Мария Олеговна
НИЯУ МИФИ

**Научный руководитель:**
Антонов Евгений Вячеславович
ЛИТ ОИЯИ

16.04.2024

# Программа автоматизированного выделения значений и единиц измерения физических величин из полнотекстовых материалов

❖ **Цель**

разработать инструмент для ***выделения из текста, унификации записи*** физических величин

❖ **Применение**

***анализ содержания*** научно-технических текстов, ***конвертация*** физических величин ***в СИ***

## Физическая величина

числовое значение
+
единица измерения

| 5.8 | MW/m2 |
|---|---|

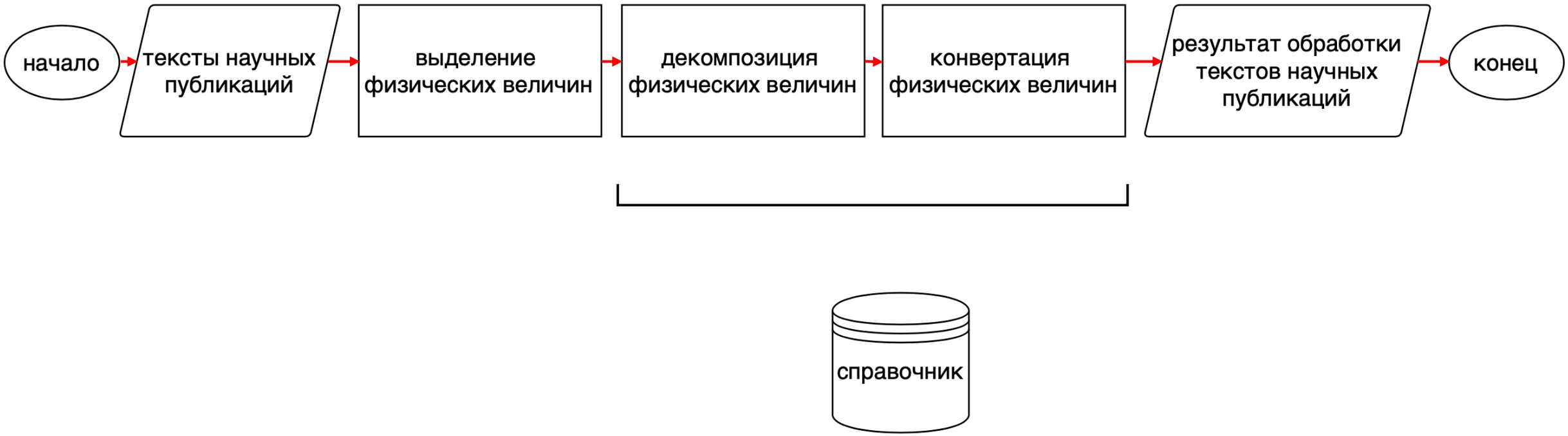
## Буквенная запись в СИ

приставка
*m; μ ...*

единица измерения

- *m kg s A K mol cd* — *базовая*
- *C= A*s ; ...* — *производная именованная*
- *m/s = m/s; ...* — *производная неименованная*

# Принцип работы



## Составляющие инструмента

- Справочник
- Алгоритм
- Программный интерфейс
- Средства контейнеризации

# Программа автоматизированного выделения значений и единиц измерения физических величин из полнотекстовых материалов

Составляющие инструмента

## Справочник

*приставки - 24*
*базовые единицы СИ - 7*
*производные именованные единицы СИ -14*
*добавленные единицы - 29*

```
'G':
        {'Name': 'giga',
        'Value': 1000000000
        }
```

```
'H': {
        'Name': 'henry',
        'Quantity': 'inductance',
        'SI': [1,'kg*m^2*s^-2*A^-2'],
        'Powers': True
```

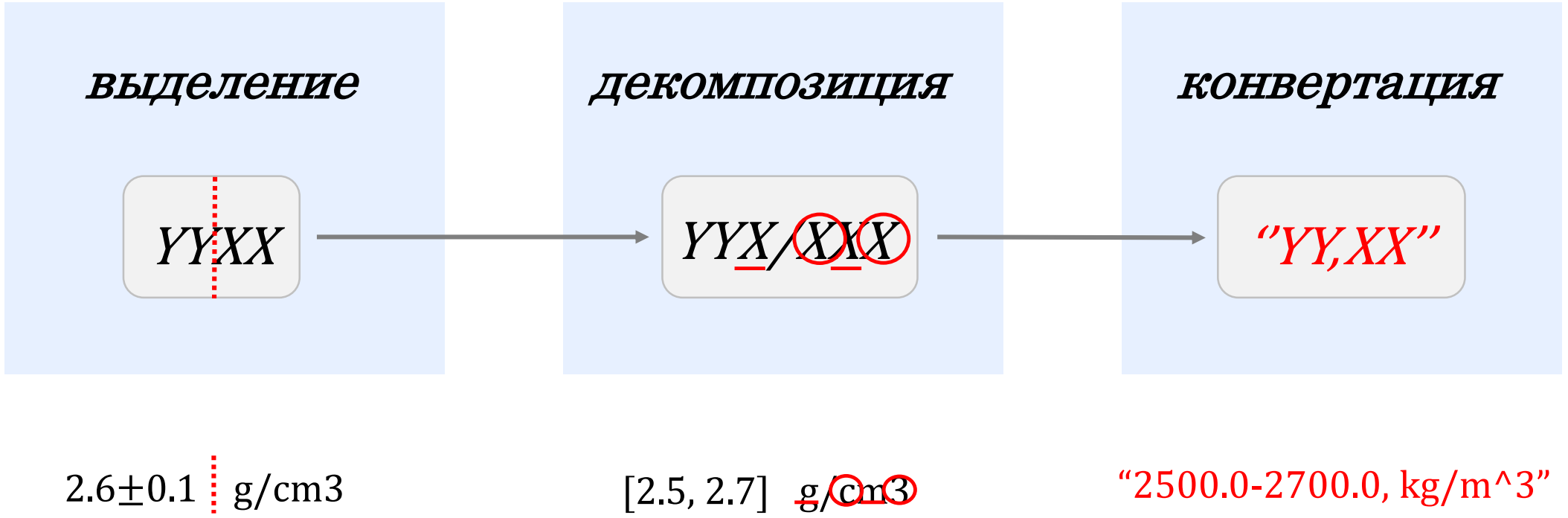
```
'J/kg': {
        'Name': ‘joule per kilogram',
        'Quantity': ‘specific energy',
        'SI': [1,'m^2*s^-2’]
```

```
'appm': {
        'Name’: ‘atomic parts per million',
        'Quantity’: ‘ratio',
        'SI': [1,’appm’],
        'Powers': False
      }
```

```
'°F': {
        'Name': 'degree Fahrenheit',
        'Quantity': 'temperature',
        'SI': ['(x-32)*5/9+273.15','K'],
        'Powers': False
      },
```

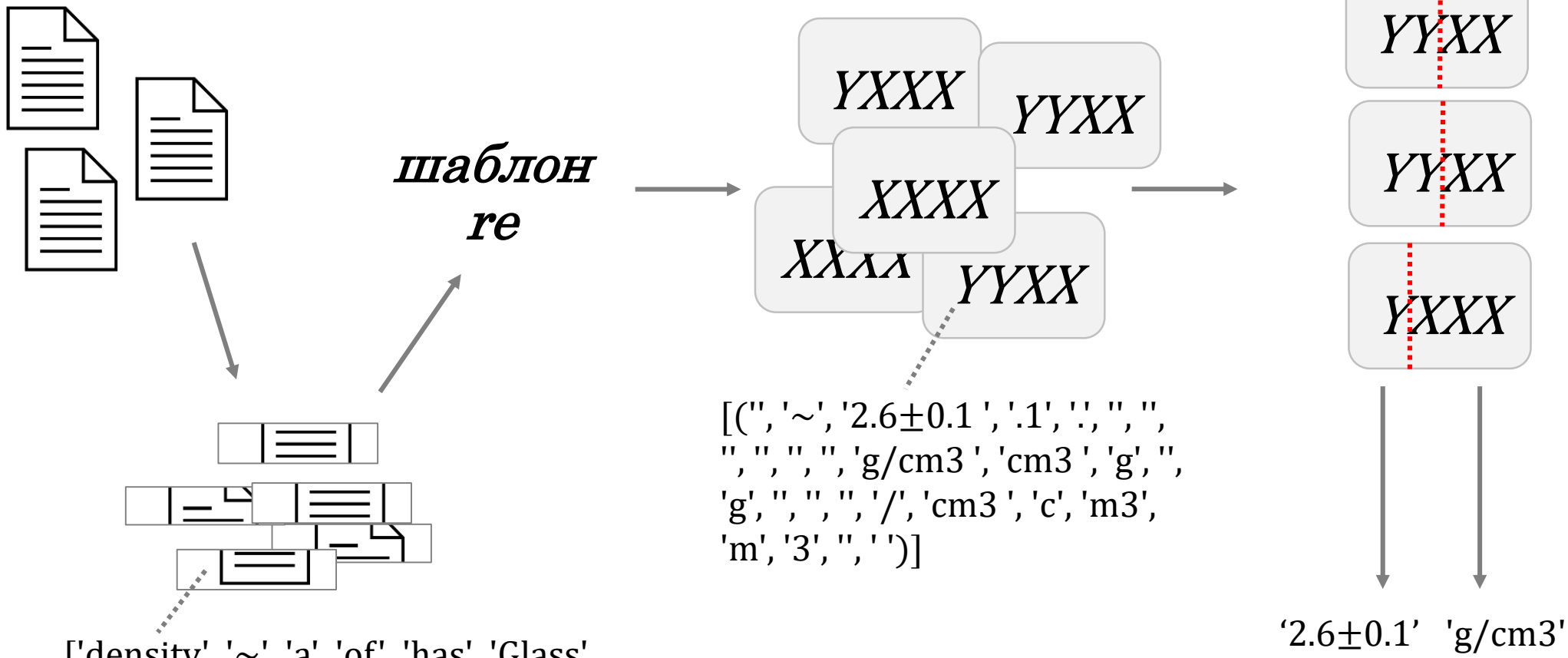
Составляющие инструмента

## Алгоритм



Составляющие инструмента

## Алгоритм выделения



Составляющие инструмента

# Алгоритм декомпозиции

## числовая группа

***шаблон***
***re***

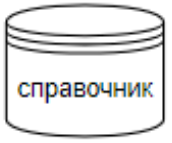
- интервал
- нормальная запись
- экспоненциальная запись

*33.0±0.8; 5*10-3; 5e-3*

[2.5, 2.7]

## символьная группа

***шаблон***
***re***




- простая/составная
- приставки
- степени

*g/cm3* — *{...}*

g [('', 'g', '', '')]
cm3 [('c', 'm', '3', '')]

{'Symbol': 'g/cm3',
'Type': 'compound', 'Description': ['NOT IN DATABASE'],
'BaseUnits': [
{'Symbol': 'g', 'Type': 'base', 'Root': 'g', 'Power': '1', 'Description': [{'Name': 'gram', 'Quantity': 'mass', 'SI': [0.001, 'kg'], 'Powers': True}], 'Prefix': {'Symbol': '', 'Description': []}},
{'Symbol': 'cm3', 'Type': 'base', 'Root': 'm', 'Power': '3', 'Description': [{'Name': 'metre', 'Quantity': 'length', 'SI': [1, 'm'], 'Powers': True}], 'Prefix': {'Symbol': 'c', 'Description': [{'Name': 'centi', 'Value': 0.01}]}}]}

Составляющие инструмента

# Алгоритм конвертации

**числовое значение**

*{...}*

коэффициент (K)

“*ч.зн.(конвертированное) = = ч.зн(обработанное) * K*”,

2500.0-2700.0 = 2.5*1000-2.7*1000

**единица измерения**

*{...}*

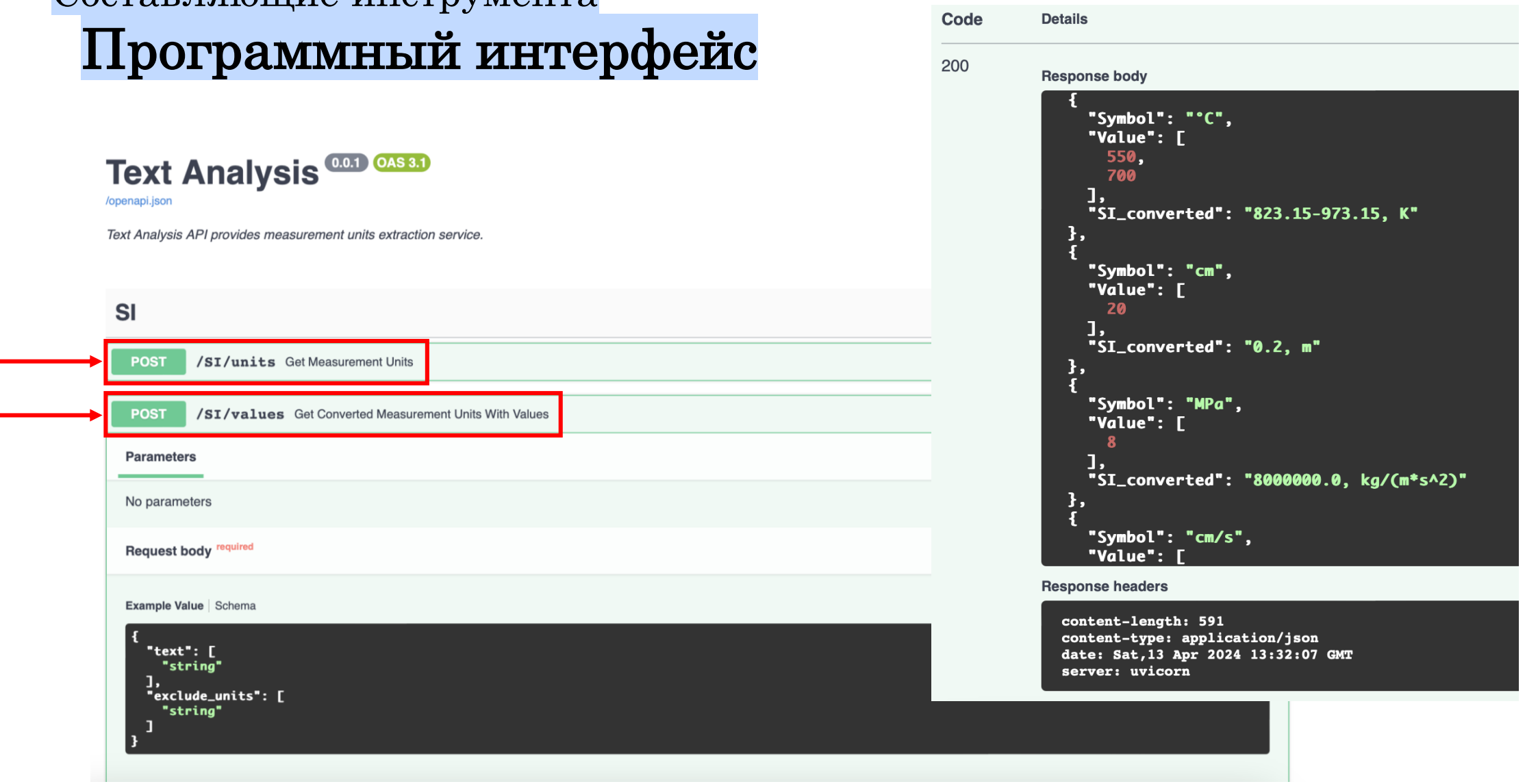
полная символьная запись в базовых единицах (ПСЗ)

“*ед.изм.(конвертированная) = = ПСЗ(приведённая)*”

kg/m^3 = (kg)^1/(m)^3

Составляющие инструмента

# Программный интерфейс



# Использование инструмента

## ❖ Cmd

```
curl -X 'POST' \
  'http://127.0.0.1:8008/SI/values' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "text": [
    "..."
  ],
  "exclude_units": [
  ]
}'
```

## ❖ Программный интерфейс

```
POST http://localhost:8008/SI/values
HTTP/1.1
```

```
{
  "text": [
    "..."
  ],
  "exclude_units": [
    "..."
  ]
}
```

# Пример работы инструмента: ввод

"Such LMs can provide sufficient tritium breeding ratio and have high thermal conductivity (~***101W/m*K***) and low viscosity ~ ***10-7m2/s*** that make them very favorable for heat removal. All LM blanket concepts have, however, feasibility issues associated with magnetohydrodynamic (MHD) interactions between the flowing high electrical conductivity LM ***106 S/m*** and a strong plasma-confining magnetic field. Only a weak flow is needed for a slow (***0.1-1mm/s***) circulation of the breeder toward the external ancillary system for tritium extraction and LM purification. In this concept, a high-temperature PbLi alloy flows slowly (velocity ***10cm/s***) in large poloidal rectangular ducts (duct size~***20 cm***) to remove the volumetric heat and produce tritium, while the pressurized He (typically to ***8MPa***) is used to remove the surface heat flux and to cool the ferritic first wall and other blanket structures to < ***550°C***. Thermal insulation is needed to thermally insulate the high-temperature self-cooled PbLi (which operates at ~***550-700°C*** depending on the variant of the design)"

# Пример работы инструмента: вывод

[ { "Symbol": "*mm/s*", "Value": [ "0.1", "1.0" ], "SI_converted": "0.0001-0.001, m/s" },
{ "Symbol": *"°C*", "Value": [ "550.0", "700.0" ], "SI_converted": ”823.15-973.15, K" },
{ "Symbol": "*W/m*K*", "Value": [ "101" ], "SI_converted": "101.0, kg*m/(s^3*K)" },
{ "Symbol": *"°C*", "Value": [ "550" ], "SI_converted": ”823.15, K" },
{ "Symbol": "*cm/s*", "Value": [ "10" ], "SI_converted": "0.1, m/s" },
{ "Symbol": "*cm*", "Value": [ "20" ], "SI_converted": "0.2, m" },
{ "Symbol": "*S/m*", "Value": [ "106" ], "SI_converted": "106.0, A^2*m^-1" },
{ "Symbol": "*MPa*", "Value": [ "8" ], "SI_converted": "8000000.0, kg/(m*s^2)" } ,
{ "Symbol": "*m2/s*", "Value": [ "1e-07" ], "SI_converted": "1e-07, m^2/s" }]

# Программа автоматизированного выделения значений и единиц измерения физических величин из полнотекстовых материалов

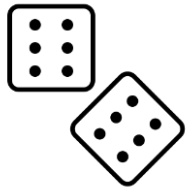

[ { "Symbol": "*mm/s*", "Value": [ "0.1", "1.0" ], "SI_converted": "0.0001-0.001, m/s" },
{ "Symbol": *"°C*", "Value": [ "550.0", "700.0" ], "SI_converted": ”823.15-973.15, K" },
{ "Symbol": "*W/m*K*", "Value": [ "101" ], "SI_converted": "101.0, kg*m^2/(s^3*K)" },
{ "Symbol": *"°C*", "Value": [ "550" ], "SI_converted": ”823.15, K" },
{ "Symbol": "*cm/s*", "Value": [ "10" ], "SI_converted": "0.1, m/s" },
{ "Symbol": "*cm*", "Value": [ "20" ], "SI_converted": "0.2, m" },
{ "Symbol": "*S/m*", "Value": [ "106" ], "SI_converted": "106.0, A^2*m^-1" },
{ "Symbol": "*MPa*", "Value": [ "8" ], "SI_converted": "8000000.0, kg/(m*s^2)" } ,
{ "Symbol": "*m2/s*", "Value": [ "1e-07" ], "SI_converted": "1e-07, m^2/s" }]


## Сравнение результатов работы инструмента и Quantulum3


[Quantity(0.55, "Unit(name="millimetre per second")",
Quantity(625, "Unit(name="degree Celsius")”,
Quantity(101, "Unit(name="watt per metre”)",
Quantity(550, "Unit(name="degree Celsius")”,
Quantity(10, "Unit(name="centimetre per second”)",
Quantity(20, "Unit(name="centimetre”)",
Quantity(106, "Unit(name="siemens per metre”)",
Quantity(8, "Unit(name="megapascal")"]

# Проблемы в работе инструмента

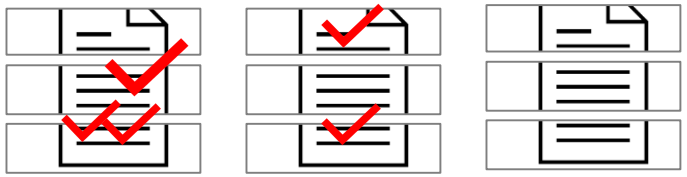
- ***Множественность Unicode-кодировок*** символов одного вида ..... *'MICRO SIGN' (U+00B5)* µ *'GREEK SMALL LETTER MU' (U+03BC)* µ
- ***Буквенное написание*** физических величин ..... *seven kilometers per second*
- ***Вариативность*** записи физических величин .....



# Перспективы применения инструмента

- ***Усовершенствованный поиск текстов с предварительно заданным набором*** единиц измерения и/или диапазонов их значений

- ***Индексация текстов для выделения технических параметров*** оборудования/ условий проведения экспериментов …



**Спасибо за внимание!**

# Приложение 1