Анализ данных и Машинное обучение в гидрологии

...

Неделя 3

План

Лекция

- открытые данные;
- источники открытых данных;
- почему вы обязаны делать свои исследовательские данные открытыми.

Практическое занятие

- как скачать весь интернет (и не умереть молодым);
- API;
- pandas - библиотека, способная вывести российскую науку из кризиса.

Что такое открытые данные?

Гос.портал открытых данных:

информация (в том числе документированная), созданная в пределах своих полномочий государственными органами, либо поступившая в указанные органы и организации, а также информационно-аналитическими организациями, участвующими в публикации собственных открытых данных на территории Российской Федерации, которая подлежит размещению в сети Интернет в формате, обеспечивающем ее автоматическую обработку в целях повторного использования без предварительного изменения человеком (машиночитаемый формат), и может свободно использоваться в любых соответствующих закону целях любыми лицами независимо от формы ее размещения (простая совокупность сведений, база данных и т.д.

Википедия:

концепция, отражающая идею о том, что определённые данные должны быть свободно доступны для машиночитаемого использования и дальнейшей републикации без ограничений авторского права, патентов и других механизмов контроля.

Что такое открытые данные?

Хартия открытых данных:

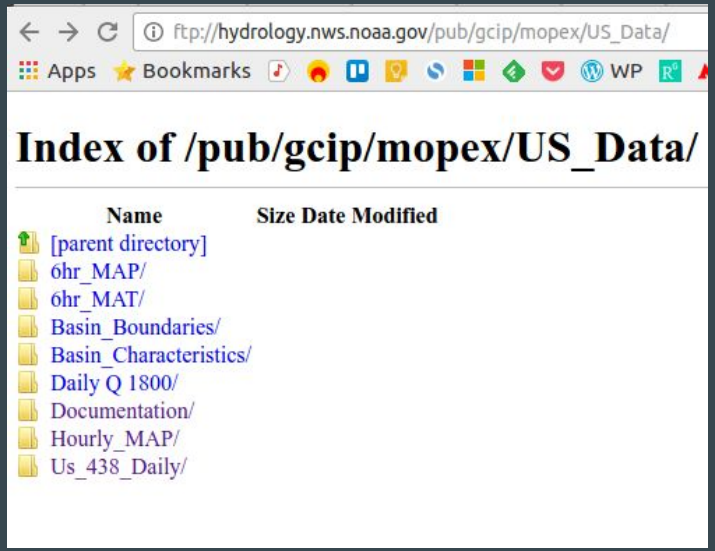
цифровые данные сделанные общедоступными с техническими и юридическими характеристиками обязательными для того чтобы они свободно использовались, использовались повторно и распространялись кем угодно, когда угодно и где угодно.

Принципы:

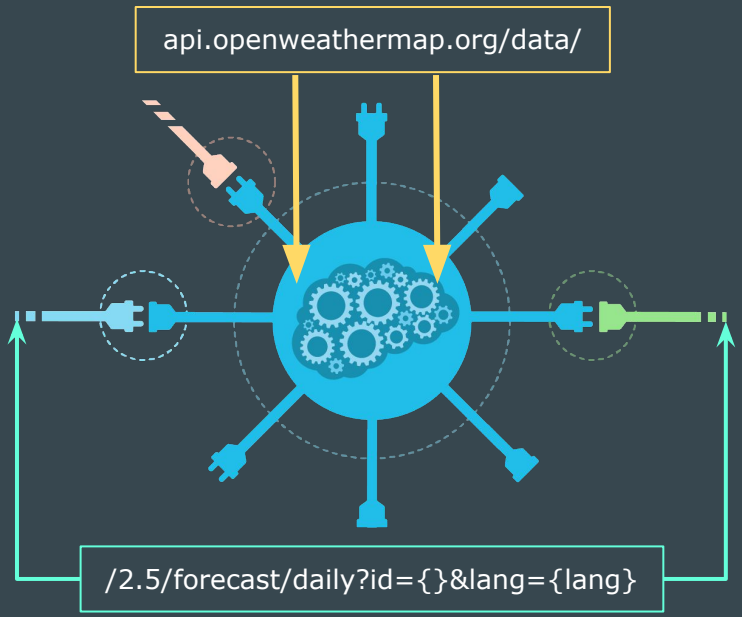
1. Открытость по умолчанию;
2. Своевременно и полно;
3. Доступно и удобно;
4. Сравнимо и интегрируемо;
5. Для улучшения управления и вовлечения граждан;
6. Для развития и инноваций.

Что такое открытые данные на самом деле?

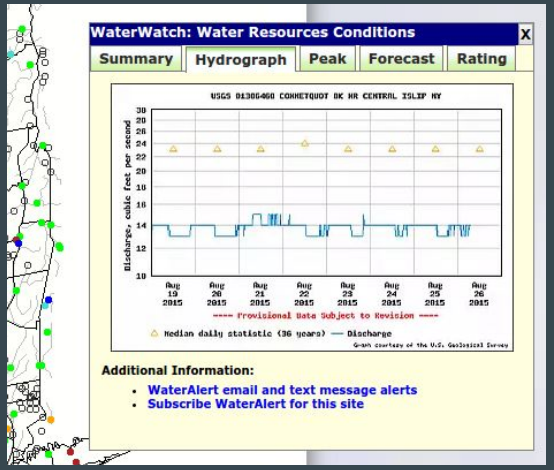
1. Удаленные репозитории (ftp, http, dropbox, yandex disk и т.д.).

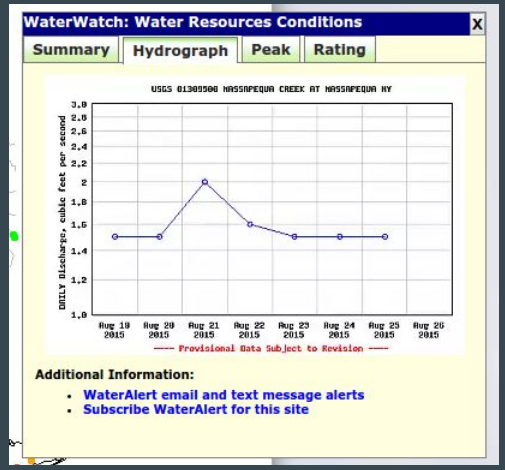


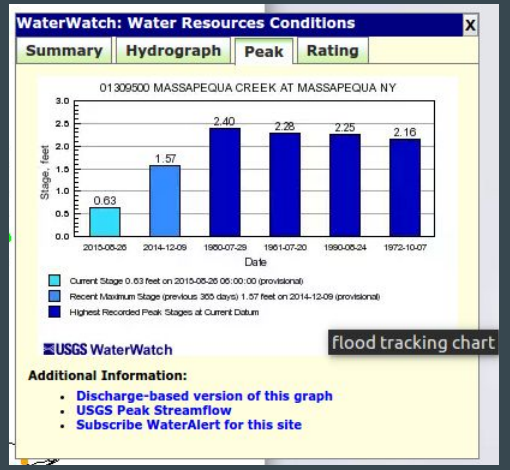
2. API (Application programming interface).



Открытые данные здорового человека

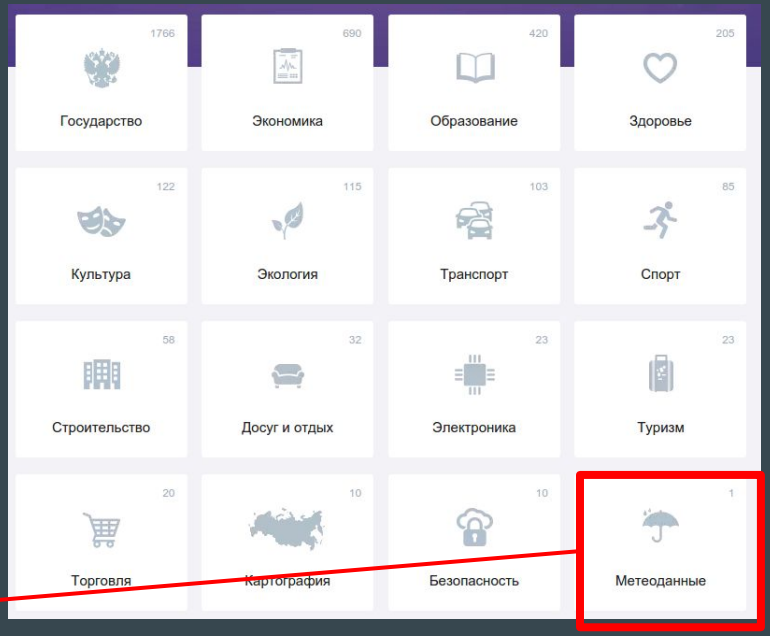


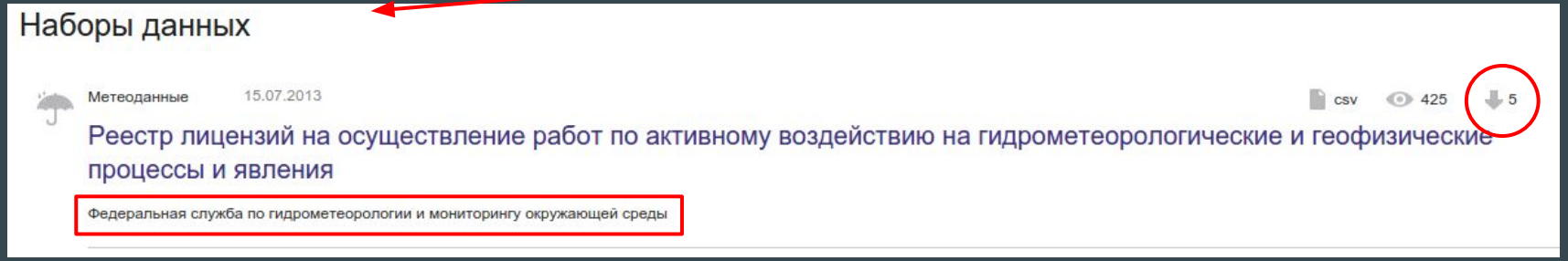




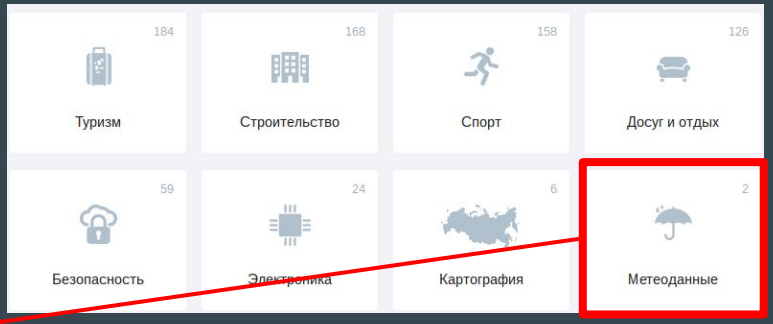
Открытые данные курильщика (2015 г.)

ОТКРЫТЫЕ ДАННЫЕ





Открытые данные курильщика (2016 г.)



Наборы данных

Метеоданные 19.10.2016

Неблагоприятные метеорологические условия

json(1.14 КБ) 224 47

Высший исполнительный орган государственной власти города Москвы-Правительство Москвы

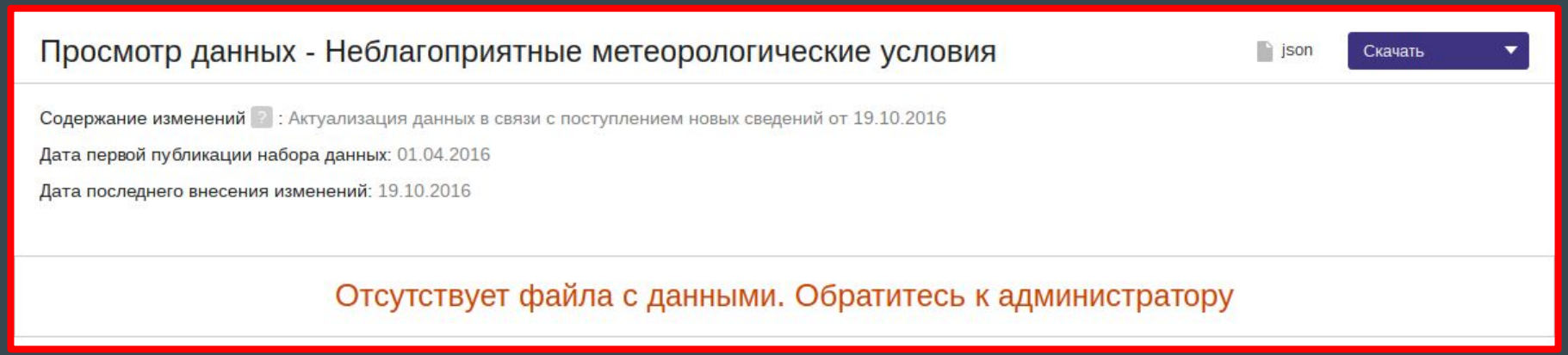
Метеоданные 15.07.2013

Реестр лицензий на осуществление работ по активному воздействию на гидрометеорологические и геофизические процессы и явления

csv(23.29 КБ) 1119 53

Федеральная служба по гидрометеорологии и мониторингу окружающей среды

Ну такое...


Просмотр данных - Неблагоприятные метеорологические условия
json
Скачать
Содержание изменений ? : Актуализация данных в связи с поступлением новых сведений от 19.10.2016
Дата первой публикации набора данных: 01.04.2016
Дата последнего внесения изменений: 19.10.2016
Отсутствует файла с данными. Обратитесь к администратору

Зачем делать данные открытыми?

- Обязательство;
- Пиар;
- Выгода;
- Стандарты (воспроизводимость исследований).

Зачем делать открытыми ваши научные данные?

Сами по себе вы никому не интересны, а у ваших данных есть шанс.

Хватит параноить

- С***ли?
- Украдет;
- Опубликует раньше;
- Найдет ошибку.

Пора продвигать свои исследования

- Найдет ошибку;
- Опубликует;
- Предложит дальнейшее сотрудничество.

Ну давай, расскажи мне

как сделать мои данные открытыми

Рецепт хороших открытых данных

1. Адекватный формат

- csv
- tsv
- netcdf
- grib
- json

- txt
- xls
- xlsx
- doc (!!!)

Рецепт хороших открытых данных

2. Следование конвенциям

- Одна переменная - один столбик;
- CF convention ver. 1.6;
- Code book;
- Неизменяемость.

- Кириллица в названиях;
- Разные обозначения одной переменной (R, Q);
- Тихая смена версии (изменения без декларации).

Рецепт хороших открытых данных

3. Доступное 24/7 хранилище

- серверы ftp, http;
- облачные хранилища (dropbox, yandex disk).

- флешка, HDD;
- google drive (ИМХО), mega.

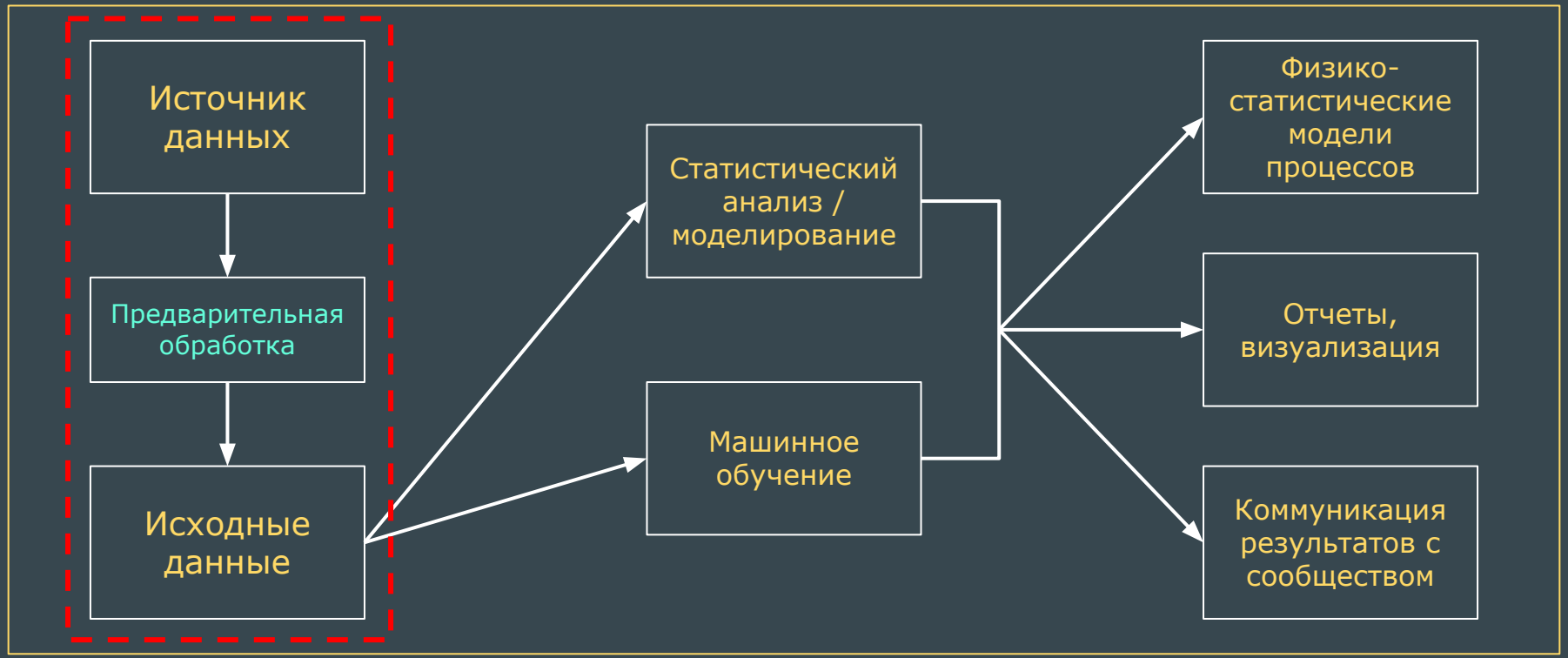
Рецепт хороших открытых данных

4. Лицензия

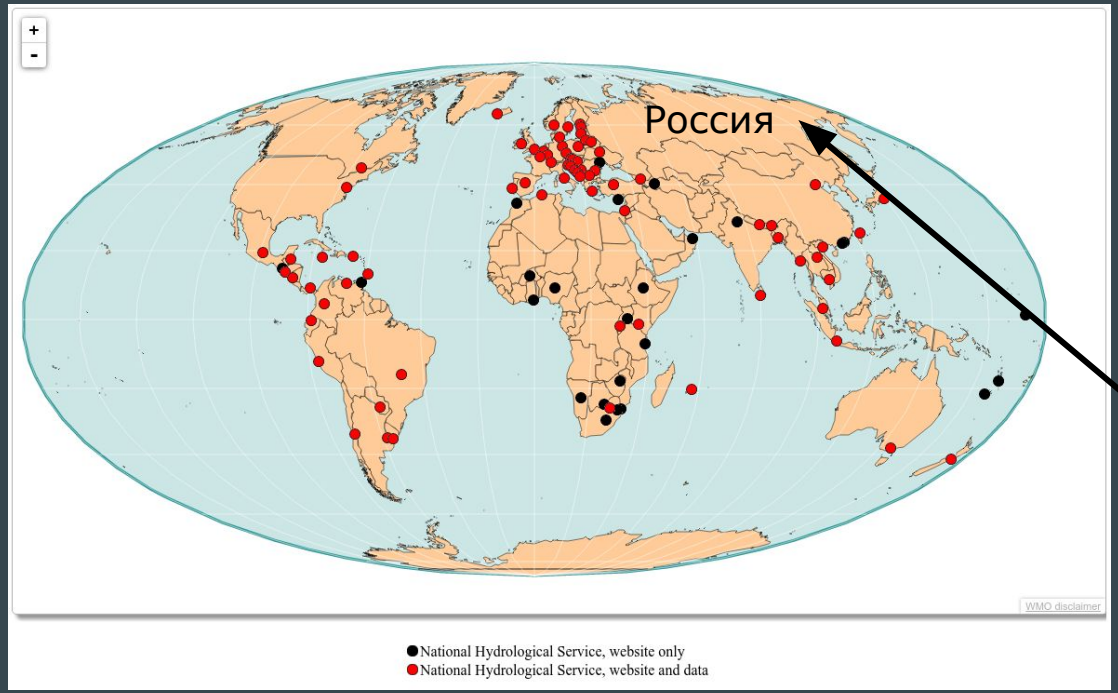
- Creative Commons;
- GNU, BSD, MIT.

- Лицензия?
- Да все будет ок, братко.

Среда воспроизводимых вычислений



Открытые данные национального масштаба



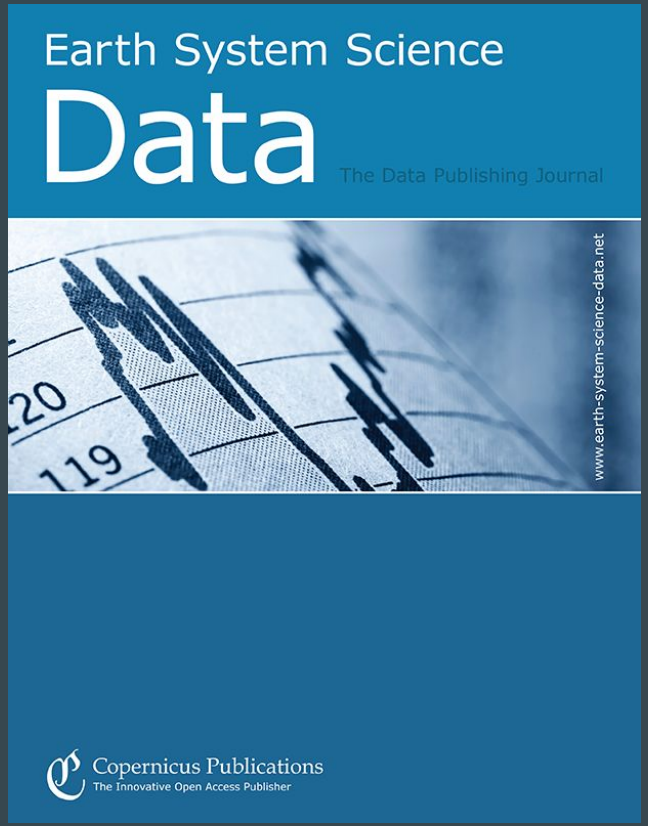
WMO Hydrological Observing System

Шах и мат, глобалисты!

Мы будем ездить в Женеву на совещания, подписывать декларации и резолюции, а данных все равно не дадим.

С любовью, ваш Росгидромет.

Лучший источник открытых данных



www.earth-system-science-data.net/

- An explicit GIS-based river basin framework for aquatic ecosystem conservation in the Amazon;
- The PRIMAP-hist national historical emissions time series;
- The integrated water balance and soil data set of the Rollesbroich hydrological observatory.

Важно

Вы можете помочь существенно улучшить этот курс!

- ayzelgv@gmail.com, hydrogo@yandex.ru
- vk.com/ayzelgv, facebook.com/ayzelgv
- ИВП РАН, кабинет 617