Как и где искать открытые данные?

Ирина Радченко
кандидат технических наук, доцент
iradche@gmail.com
телеграм-канал @DataPlace

мастер-класс для Skills Lab
24 сентября 2020

План

- Определения
- Порталы и агрегаторы данных
- Репозитории данных
- Платформы распространения данных
- Информационные ресурсы по поиску данных
- Как разместить и аннотировать данные, чтобы их можно было найти

Определения

OPEN DATA HANDBOOK

Open Knowledge Foundation

GUIDE

Что такое Открытые данные?

Improve this page

Translate this page

Donate

Share

Languages: de el en es fr he hr id is it ja ko lt lv my ne nl_BE pt_BR ro ru zh_CN zh_TW

Это руководство посвящено *открытым данным*, но что же это в точности такое? В частности, что делает *открытые* данные открытыми, и о каких именно данных мы говорим.

Что значит открытость?

- **Доступность и читаемость:** данные должны быть доступны целиком не дороже разумной стоимости их воспроизведения; желательно через интернет. Формат данных должен быть удобным для чтения и изменения.
- **Повторное использование и распространение:** данные должны предоставляться на условиях, которые разрешают их повторное использование и распространение, в том числе - в комбинации с другими наборами данных.
- **Всеобщее участие:** каждый должен иметь возможность использовать и распространять данные. Не должно быть дискриминации областей применения, людей или групп. Например, ограничение «только для некоммерческого использования», которое запрещает «коммерческое» применение, или ограничение возможных областей применения (к примеру, только в образовании), недопустимы.

Если у вас возник вопрос, почему так важно чётко определить, что значит открытость и почему используется такое определение, есть простой ответ: **интероперабельность**.

Интероперабельность означает способность различных систем и организаций работать совместно (англ. inter-operate). В данном случае, речь идёт о способность к взаимодействию — или «перемешиванию» различных наборов данных.

Интероперабельность важна, поскольку позволяет различным компонентам работать вместе. Возможность выделения компонентов и их «сборки» в единое целое — ключевая для построения больших и сложных систем. Без интероперабельности это практически

способность

Improve this page

Translate this page

Donate

Share

https://opendatahandbook.org/guide/ru/what-is-open-data/

Пантонские принципы

Panton Principles

Principles for Open Data in Science

Endorse Principles About Comment FAQ Panton Fellowships

Contributors
========

* Peter Murray-Rust
* Rufus Pollock
* Bryan Bishop
* Mike Chelen
* Cameron Neylon
* Daniel Mietchen
* Andy Powell

Basics
======

Q1: What are the Panton Principles?

The Panton Principles are a set of recommendations that address how best to make published data from scientific studies available for re-use. In this context, “published” means “made public” and is not restricted to formal publication in the scholarly literature.

Web buttons

Get an **open data web button** for your project!

This material is Open Data

This material is Open Data

Related Links

- Open Science Working Group - Open Knowledge Foundation
- Open Definition - Defining the Open in Open Data and Content
- Is It Open Data?
- Science Commons - Protocol for Implementing Open Access Data

https://pantonprinciples.org/index.html

Принципы FAIR

scientific **data**

View all Nature Research journals | Search | Login

Explore our content | Journal information

nature > scientific data > comment > article

Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

143k Accesses | **1752** Citations | **1592** Altmetric | Metrics

An Addendum to this article was published on 19 March 2019

Abstract

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and

Download PDF

Associated Content

Collection
Scientific data

Collection
Metadata quality

Sections | References

Abstract
Comment
Additional Information
References

https://doi.org/10.1038/sdata.2016.18

Порталы
и агрегаторы данных

Порталы открытых данных

https://data.gov.ru/

https://www.data.gov/

https://data.gov.uk/

Порталы открытых данных по всему миру

https://opendatainception.io/

Портал открытых данных World Bank

https://data.worldbank.org/

Портал открытых данных OECD

https://data.oecd.org/

EU open data portal

https://data.europa.eu/euodp/en/data

European Data portal: сборщик метаданных

https://www.europeandataportal.eu/data/datasets?locale=en&minScoring=0

Портал ВОЗ: Global Health Data

https://www.who.int/data/gho

Centers for Disease Control and Prevention

https://data.cdc.gov/browse

Федеральная служба государственной статистики

https://rosstat.gov.ru/statistic

ЕМИСС государственная статистика

https://www.fedstat.ru/

Проект ГосЗатрат

https://clearspending.ru/

Платформа DrivenData

https://www.drivendata.org/

Платформа Kaggle

https://www.kaggle.com/datasets

NASA

Earth Data

https://earthdata.nasa.gov/

Space Data

https://pds.nasa.gov/datasearch/data-search/

Репозитории данных

Time Series Classification

Time Series Classification | Home | Datasets | Algorithms | Results | Researchers | Code | Bibliography | UEA Papers | About Us

Train Size

Less than 100 (49)
100 to 500 (85)
Greater than 500 (44)

Test Size

Less than 300 (84)
300 to 1000 (49)
Greater than 1000 (45)

Length

Less than 300 (91)
300 to 700 (40)
Greater than 700 (47)

Classes

Less than 10 (134)
10 to 30 (34)
Greater than 30 (10)

Type

Device (10)
ECG (9)
Image (34)
Motion (35)

Dataset listing

The univariate and multivariate classification problems are available in three formats: Weka ARFF, simple text files and sktime ts format. Weka does not allow for unequal length series, so the unequal length problems are all padded with missing values. ts format does allow for this feature.

Univariate Weka formatted ARFF files and .txt files (about 500 MB).

Univariate sktime formatted ts files (about 300 MB).

Multivariate Weka formatted ARFF files (and .txt files) (about 2 GB).

sktime formatted ts files (about 1.5 GB).

Lists of the data, including which are unequal length, can be found here. Details on loading sktime data with the Python package are here .

To store multivariate series in ARFF we take advantage of relational attributes. These are fairly unintuitive, so we have provided an overview of this and other basic features of loading data and building classifiers here.

These files provide a simple list of the data characteristics for univariate and multivariate problems.

To see how accurate different classifiers are on these data see the results page.

More information on the datasets is given below. Problems with variable length series are listed as length 0.

Search

Dataset	Train Size	Test Size	Length	No. of Classes	Type
AbnormalHeartbeat	303	303	3053	5	AUDIO
ACSF1	100	100	1460	10	DEVICE
Adiac	390	391	176	37	IMAGE

http://timeseriesclassification.com/dataset.php

NLP Database

The Big Bad NLP Database

Quantum stat

For database updates follow on or

Want to add a dataset, edit?

updated 09.19.20 584 datasets

search...

1 2 3 4 5 6

Dataset	Added	Lang	Description	Inst	Format	Task	Year	Creator	Source
Visual Genome	09.19.20	English	Dataset contains over 100K images where each image has an average of 21 objects, 18 attributes, and 18 pairwise relationships between objects.	100,000+	JSON	Visual Question Answering, Knowkedge Base	2016	Krishna et al.	LINK PAPER
CraigsListBargain	09.19.20	English	Dataset contains 6,682 human-human dialogues where 2 agents negotiate the sale/purchase of an item.	6,682	JSON	Dialogue	2018	He et al.	LINK PAPER
A Multi-Turn, Multi-Domain Dialogue Dataset (KVRET)	09.19.20	English	Dataset contains 3,031 multi-turn dialogues in three distinct domains appropriate for an in-car assistant: calendar scheduling, weather information retrieval, and point-of-interest navigation.	3,301	JSON	Dialogue	2017	Eric et al.	LINK PAPER
CMU_ARCTIC	09.19.20	English	Dataset contains 1,150 utterances carefully selected from out-of-copyright texts from Project Gutenberg. The databases include US English male (bdl) and female (slt) speakers (both experinced voice talent) as well as other accented speakers.	1,150	WAV	Speech Recognition	2004	CMU	LINK PAPER
			Dataset includes recordings from twenty-four (24) non-native speakers of						

https://datasets.quantumstat.com/

Machine Learning Repository

https://archive.ics.uci.edu/ml/index.php

Google: Проект Dataset Search

https://datasetsearch.research.google.com/

Google: BigQuery public datasets

https://cloud.google.com/bigquery/public-data/

Google: YouTube-8M Segments Dataset

YouTube | 8M

Dataset Explore Download Workshop About

YouTube-8M Segments Dataset

The YouTube-8M Segments dataset is an extension of the YouTube-8M dataset with human-verified segment annotations. In addition to annotating videos, we would like to temporally localize the entities in the videos, i.e., find out when the entities occur.

We collected human-verified labels on about 237K segments on 1000 classes from the validation set of the YouTube-8M dataset. Each video will again come with time-localized frame-level features so classifier predictions can be made at segment-level granularity. We encourage researchers to leverage the large amount of noisy video-level labels in the training set to train models for temporal localization.

We are organizing a Kaggle Challenge and The 3rd Workshop on YouTube-8M Large-Scale Video Understanding at ICCV 2019.

237K	1000	5.0
Human-verified Segment Labels	Classes	Avg. Segments / Video

In addition to annotating the topical entity of the full-video, we want to understand when the entity occurs in videos. Given a 5-second segment and a query class, our human raters are asked to verify whether the entity is identified within the segment. To speed up the annotation process, our human raters do not report presence or absence of non-query classes.

https://research.google.com/youtube8m/

Google: AudioSet

https://research.google.com/audioset/

Google: AVA, audiovisual annotations of video

AVA

Dataset Explore Download Challenge About

AVA is a project that provides audiovisual annotations of video for improving our understanding of human activity. Each of the video clips has been exhaustively annotated by human annotators, and together they represent a rich variety of scenes, recording conditions, and expressions of human activity.

We provide the following annotations.

AVA-Kinetics Dataset

AVA-Kinetics, our latest release, is a crossover between the AVA Actions and Kinetics datasets. In order to provide localized action labels on a wider variety of visual scenes, we've provided AVA action labels on videos from Kinetics-700, nearly doubling the number of total annotations, and increasing the number of unique videos by over 500x. We hope this will expand the generalizability of localized action models, and open the door to new approaches in multi-task learning.

AVA-Kinetics is described in detail in the arXiv paper.

AVA-Kinetics is now available for download. It was the basis of a challenge in partnership with the ActivityNet workshop at CVPR 2020.

AVA Actions Dataset

The AVA dataset densely annotates 80 atomic visual actions in 430 15-minute movie clips, where actions are localized in space and time, resulting in 1.62M action labels with multiple labels per human occurring frequently. A detailed description of our contributions with this dataset can be found in our accompanying CVPR '18 paper.

AVA v2.2 is now available for download. It was the basis of a challenge in partnership with the ActivityNet workshop at CVPR 2019.

https://research.google.com/ava/

Google: Open Image

https://storage.googleapis.com/openimages/web/index.html

Open Data on Amazon

https://registry.opendata.aws/

https://aws.amazon.com/ru/opendata/open-data-sponsorship-program/

AWS: Биологические данные

1000 Genomes Project and AWS

The 1000 Genomes Project is an international collaboration which has established the most detailed catalogue of human genetic variation, including SNPs, structural variants, and their haplotype context. The final phase of the project sequenced more than 2500 individuals from 26 different populations around the world and produced an integrated set of phased haplotypes with more than 80 million variants for these individuals.

The Amazon mirror contains the complete data set from the project and the data can be found in the `s3://1000genomes` bucket in the us-east-1 AWS region.

For more information please look at http://www.1000genomes.org. If you have any questions about the data, please email info@1000genomes.org.

Accessing 1000 Genomes Data

AWS is making the 1000 Genomes Project data publicly available to the community free of charge. Public Data Sets on AWS provide a centralized repository of public data hosted on Amazon Simple Storage Service (Amazon S3). The data can be seamlessly accessed from AWS services such Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Elastic MapReduce (Amazon EMR), which provide organizations with the highly scalable compute resources needed to take advantage of these large data collections. AWS is storing the public data sets at no charge to the community. Researchers pay only for the additional AWS resources they need for further processing or analysis of the data. Learn more about Public Data Sets on AWS.

The latest 1000 Genomes Project data is publicly available in the 1000genomes Amazon S3 bucket.

You can access the data via simple HTTP requests, or take advantage of the AWS SDKs in languages such as Ruby, Java, Python, .NET and PHP.

Analyzing 1000 Genomes Data

Researchers can use the Amazon EC2 utility computing service to dive into this data without the usual capital investment required to work with data at this scale. AWS also provides a number of orchestration and automation services to help teams make their research available to others to remix and reuse.

Making the data available via a bucket in Amazon S3 also means that customers can crunch the information using

https://docs.opendata.aws/1000genomes/readme.html

Спонсорская программа от AWS

https://aws.amazon.com/ru/opendata/open-data-sponsorship-program/

Microsoft Research Open Data

https://msropendata.com/

Платформы распространения данных

CKAN

https://ckan.org/about/instances/

DataHub

https://datahub.io/search

Проект Хаб открытых данных

https://hubofdata.ru/

Dataverse

https://dataverse.org/

Платформа figshare

https://figshare.com/account/home#/data

Платформа Zenodo

https://zenodo.org/search?page=1&size=20&type=dataset

Academic Torrents

https://academictorrents.com/

LOD cloud

https://lod-cloud.net/datasets

Проект Wiki DATA

https://www.wikidata.org/wiki/Wikidata:Main_Page

Информационные ресурсы по поиску данных

Коллекция KDnuggets

https://www.kdnuggets.com/datasets/index.html

Awesome Public Datasets

https://github.com/awesomedata/awesome-public-datasets

Открытые данные на русском языке

https://github.com/infoculture/awesome-opendata-rus

Наборы открытых данных с сертификатами

https://certificates.theodi.org/en/datasets

BuzzFeed

https://github.com/BuzzFeedNews

Pew Research Centre

https://www.pewresearch.org/internet/datasets/

Задать вопрос в ResearchGate

https://www.researchgate.net/post/Where_are_researchers_publishing_datasets_Especially_environmental_sustainability_and_geoscience_researchers

Как разместить
и аннотировать данные

Публикация на Hub of Data

https://hubofdata.ru/dataset/hse2019export20152019

Публикация данных в Springer Computer Proceedings

https://www.springer.com/gp/computer-science/lncs/data-publishing-in-computer-proceedings

Публикация данных на платформе Zenodo

zenodo | Search | Upload | Communities | iradche@gmail.com

April 25, 2017

Dataset | Open Access

A list with recommended journals by Russian VAK

Irina Radchenko

Processed list of recommended journals in social science and computer science by VAK (source of raw data: http://vak.ed.gov.ru/87, date: 19.04.2017).

Preview

Название журнала	Специализация	Дата включения в реестр
Alma mater (Вестник высшей школы)	13.00.00 – педагогические науки; 09.00.00 – философские науки; 22.00.00 – социологические науки.	2015-12-01
International Journal for Computation Civil and Structural Engineering (Международный журнал по расчету гражданских и строительных конструкций	01.02.00 – механика; 05.13.00 – информатика, вычислительная техника и управление; 05.23.00 – строительство и архитектура	2015-12-01

Edit

Upgrade to versioned record

52 views | 45 downloads

See more details...

	All versions	This version
Views	52	52
Downloads	45	45
Data volume	6.7 MB	6.7 MB
Unique views	51	51
Unique downloads	36	36

More info on how stats are collected.

Publication date:
April 25, 2017

DOI:
DOI 10.5281/zenodo.566976

Keyword(s):
scientific journals

License (for files):
Creative Commons Attribution 4.0 International

Share

Cite as

Irina Radchenko. (2017). A list with recommended journals by Russian VAK [Data set]. Zenodo. http://doi.org/10.5281/zenodo.566976

Start typing a citation style...

https://zenodo.org/record/566976#.X2vjRZP7QUs

Советы по публикации научных данных

- Не забывайте выкладывать свои исследовательские данные на **платформах распространения данных**
- Снабжайте выкладываемые данные **открытыми лицензиями**
- Снабжайте выкладываемые данные **понятной аннотацией** и **корректными метаданными**
- Снабжайте выкладываемые данные **постоянными идентификаторами (PID)**

ЕСЛИ:
— **смогут найти** ваши данные,
— **разобраться** в них,
— ваши данные окажутся **полезными**

Ваши коллеги вас оценят и процитируют!

Спасибо за внимание!

Телеграм-канал: @DataPlace

Сайт: http://iRadche.ru

@dadaistka

@iRadche

https://www.facebook.com/iRadche

http://www.slideshare.net/iRadche