Садржај

Увод 5

Први део књиге: *Наука о подацима* 5

Други део књиге: *Jupyter notebook, matplotlib.pyplot, pandas, seaborn* 5

Трећи део књиге: *Једноставни примери – примена* 5

Четврти део књиге: *Рад са отвореним подацима* 6

Први део 7

Наука о подацима 7

Историјски осврт 8

О отвореним подацима 12

Формати отворених података 16

Други део 21

Џупитер радне свеске и рад са њима 21

Библиотека math 30

Библиотека matplotlib.pyplot 43

Табеларно представљање података 68

Библиотека Pandas 72

Библиотека Seaborn 117

Трећи део 119

Учитавање табеле из Excel-а 119

Пивотирање табеле 121

Четврти део 125

Рад са отвореним подацима 125

Пример 1 – наставници 125

Пример 2 – ненаставно особље 140

Увод

Ова књига је намењена свакоме ко жели да научи како да користи програмски језик Пајтон и његове библиотеке у интерактивном окружењу Џупитер за обраду, визуализацију и анализу података. Предуслов да бисте могли да пратите садржај ове књиге јесте елементарно познавање програмског језика Пајтон. Уколико до сада нисте имали прилике да изучавате овај програмски језик, на порталу отворених ресурса petlja.org налазе се интерактивни бесплатни курсеви који вам могу помоћи у савладавању овог програмског језика. Књига *Пајтон за основце* такође вам може дати добру основу за даљи рад.

Књига је подељена у четири дела.

Први део књиге: *Наука о подацима*

Први део књиге је историјског карактера и представља кратак осврт на то како се наука о подацима развијала, како је и зашто постала тако изузетно популарна данас. Такође, у овом делу ћемо се осврнути на неколико изузетно актуелних појмова данашњице, као што су велики подаци и отворени подаци, направићемо и кратак осврт на то да ли је наука о подацима само другачији назив за примењену статистику.

Други део књиге: *Jupyter notebook, matplotlib.pyplot, pandas, seaborn*

У другом делу књиге нас очекује упознавање са интерактивним окружењем Џупитер (енг. *Jupyter*), као и библиотекама програмског језика Пајтон:

- matplotlib.pyplot
- Pandas
- seaborn

Трећи део књиге: *Једноставни примери – примена*

Трећи део књиге је онај у коме ћемо све оно чега смо се подсетили у вези са детаљима програмског језика Пајтон, заједно са новим садржајима са којима смо се упознали, применити на једноставним малим примерима које ћемо сами креирати.

Четврти део књиге: *Рад са отвореним подацима*

Четврти део књиге намењен је за примену наученог над отвореним подацима. У овом делу ћемо обрађивати, визуализовати и анализирати неколико скупова података који су преузети са портала отворених података https://data.gov.rs/sr/. Овај део књиге је сиже свега споменутог до тог тренутка и добро је место да се на конкретном примеру види како се теоријска знања могу применити на конкретним подацима.

Приликом писања ове књиге коришћени су материјали који се налазе на порталу отворених ресурса petlja.org у оквиру курсева net.kabinet. Материјали који су интензивно коришћени су „Курс програмирања за 2. разред гимназије" и приручник „Анализа отворених података".

Први део

Наука о подацима

Наука о подацима је интердисциплинарно поље усмерено ка закључивању из скупова података, који су обично велики (велики подаци), и примени знања и увида из података за решавање проблема који су распрострањени свуда око нас. Наука о подацима је изузетно сложена област, која се веома брзо развија у последње време. Треба напоменути да је то „жива" област, односно да смо савременици њеног свакодневног напредовања и мењања. Ова наука обухвата припрему података за анализу, формулисање проблема, анализу података, развој решења заснованих на подацима и представљање решења за доношење одлука на високом нивоу у различитим доменима примене.

С обзиром на све ово, ова наука укључује вештине из рачунарства, статистике, информатике, математике, визуализације информација, интеграције података, графичког дизајна, сложених система, комуникације и пословања. Статистичар Нејтан Јау (енг. *Nathan Yau*) повезује науку о подацима са интеракцијом човек–рачунар: корисници би требало да су у стању да интуитивно контролишу и истражују податке. Америчко статистичко удружење је 2015. идентификовало управљање базама података, статистику и машинско учење као три нове гране у науци од изузетног значаја за ову област.

Ова наука је изузетно важна, јер се примењује на различитим доменима и утиче на животе многих људи.

Наука о подацима игра важну улогу у готово свим аспектима пословања и стратегија. На пример, пружа информације о купцима, које потом помажу компанијама да креирају снажније маркетиншке кампање и циљано оглашавање за повећање продаје производа. Помаже у управљању финансијским ризицима, откривању лажних трансакција и спречавању кварова опреме у производним погонима и другим индустријским окружењима. Помаже у блокирању сајбер напада и других безбедносних претњи у ИТ системима.

Са друге стране, применом ове науке може се оптимизовати управљање ланцима снабдевања, залихама производа, дистрибутивним мрежама и корисничком подршком. На фундаменталнијем нивоу, примена ове науке, тачније њених области, указује на пут ка повећаној ефикасности и смањењу трошкова. Наука о подацима такође омогућава компанијама да креирају пословне планове и стратегије засноване на информисаној анализи понашања купаца, тржишних трендова и конкуренције. Без тога, предузећа могу пропустити прилике и доносити погрешне одлуке.

Наука о подацима је такође витална у областима изван редовног пословања. У здравственој заштити њена употреба укључује дијагностику здравствених стања, анализу слике, планирање лечења и медицинска истраживања. Академске институције користе науку о подацима за праћење учинка студената, креирање стратегија за помоћ ученицима приликом учења и побољшање њиховог маркетинга за будуће студенте. Спортски тимови анализирају перформансе играча и планирају стратегије игре помоћу науке о подацима. Владине агенције и организације за јавну политику такође су велики корисници.

Уопштено говорећи, једна од највећих предности науке о подацима јесте оснаживање и омогућавање бољег одлучивања. Организације које улажу у то могу у своје пословне одлуке укључити мјерљиве доказе засноване на подацима. У идеалном случају, такве одлуке засноване на подацима довешће до јачих пословних перформанси, уштеде трошкова и уједначених пословних процеса и токова рада.

Историјски осврт

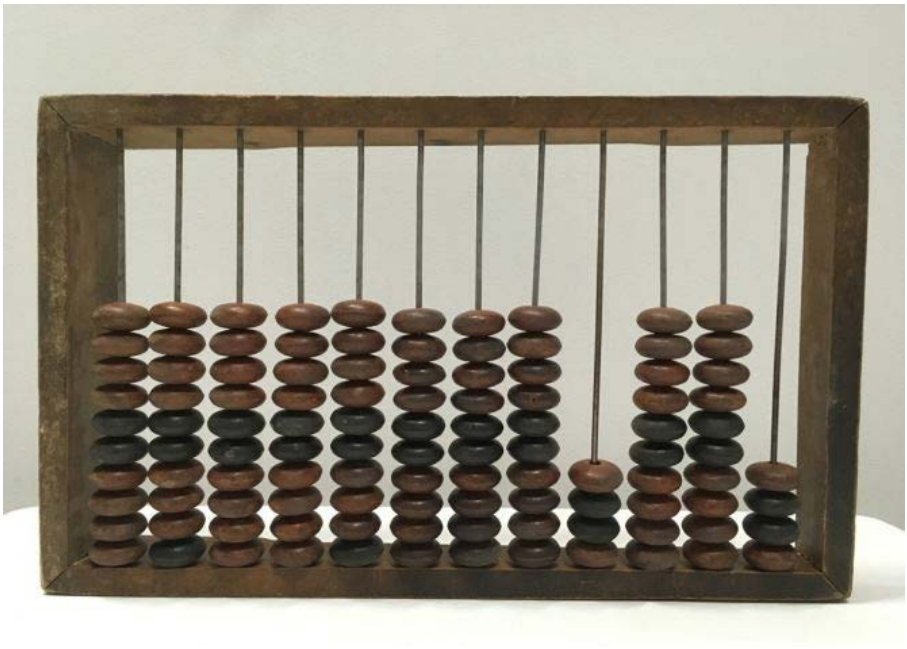
Историјски гледано, процес прикупљања и обраде података почео је веома давно. Тренутно је то изузетно жива област, али своје корене има далеко у прошлости, још док рачунари и нису постојали у оваквом облику. Идеја да се анализирају подаци и долази до нових сазнања управо на основу њихове анализе била је прво присутна у академским круговима.

Ако погледамо далеко у прошлост, можемо рећи да је прикупљање података почело још у палеолиту. Наиме, још тада су племенски народи, пратећи своје активности приликом трговине, урезивали цртице у дрвене штапове или кости животиња. Касније сравњивање дугова сводило би се упоређивањем урезаних цртица на костима. Једна оваква кост је пронађена у Африци, у националном парку Ишанго у Конгу.



Кост из Ишанга

Абакус, једноставна справа за рачунање слична дечијој рачунаљки, појављује се први пут када и библиотеке у Вавилону. Управо те библиотеке представљају људску потребу за масовно складиштење података. Александријска библиотека је исто једно такво место за складиштење података.

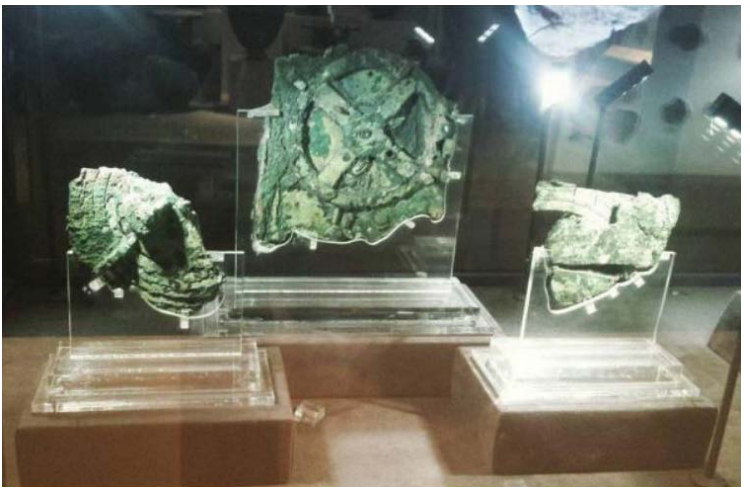


Абакус

Споменимо и *антикитеру*, механизам који је претеча данашњим рачунарима, а који датира из старог века. Чува се у Националном историјском музеју у Атини. Другачије се још назива *Антикитера механизам* или *Механизам са Антикитере*[1]. Антикитера је оство у Грчкој. Значај овог механизма је у томе што нам показује колико су антички Грци изузетно добро владали науком и још тада направили механизам који се по својим особинама може сматрати претечом данашњих рачунара. Њиме су израчунавали астрономске позиције.

Остаци механизма пронађени су 1901. године, али се тек педесетак и више година након тога сазнало шта он заправо представља. Занимљиво је да га је открио ловац на морске сунђере у олупини једног трговачког брода. Археолог Валериос Стаис први је докучио намену овог механизма, откривши да уређај има уграђен точак са видљивим натписима астрономских термина. Још једна од занимљивости је и та да се технолошке направе сличне по својој сложености нису појављивале наредних 1000 година.

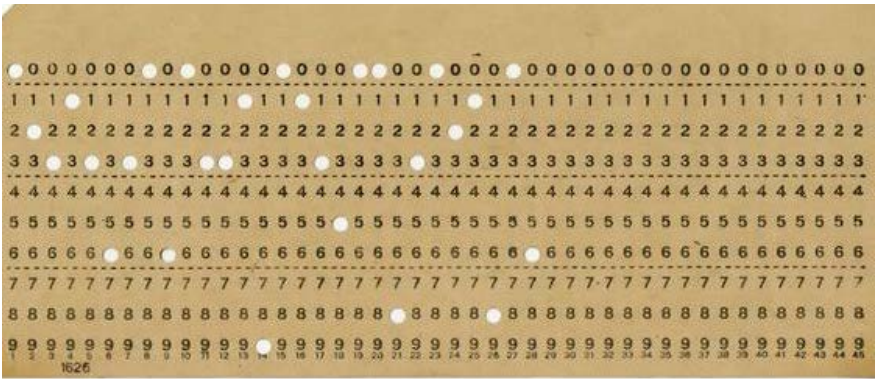
[1] Антикитера је острво у Грчкој.



Механизам са Антикитере

За прве елементе појаве стастистике можемо сматрати рад Џона Граунта, *једноī од ӣрвих демоīрафа* (енг. *John Graunt*), који је 1663. године у Европи бележио информације у вези са смртношћу од бубонске куге. Он је покушавао да анализом забележених података предвиди потенцијалне нове исходе.

Херман Холерит (енг. *Herman Hollerith*) је такође значајно име у свету аутоматизације обраде података. Наиме, САД су имале проблем приликом обраде података на попису становништва 1880. године. Прогноза је била да се сви подаци који су том приликом прикупљени могу обрадити у наредних 8 година, што је било неприхватљиво дуго. Такође, претпоставка у том тренутку је била да би се подаци са наредног пописа могли обрађивати добрих 10 година. Тада је Холерит осмислио машину за табелирање која се базирала на примени тзв. бушених картица. Холерит је познат и по томе што је касније основао чувену компанију IBM.



Бушена картица

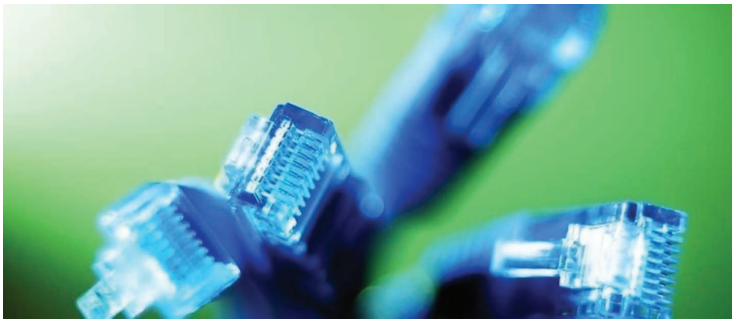
Година 1928. историјски је значајна за складиштење података, јер је те године Фриц Флојмер (гер. *Fritz Pfleumer*), немачко–аустријски инжењер, изумео метод магнетног складиштења информација на траци. Принципи које је тада развио добрим делом су и данас у употреби.

Године 1944. Фремонт Ридер (енг. *Fremont Rider*), библиотекар са Универзитета у Конектикату, објавио је рад у коме је анализирао колико ће се у будућности ширити библиотеке због количине књига које ће годинама расти. Овим радом се заправо квантификује количина произведених информација и њихов раст у наредном периоду. Током писања овог рада библиотекар је закључио да би америчке библиотеке морале да удвоструче своју величину на сваких 16 година уколико желе да похране све књиге које ће у будућности настати. Приликом процењивања количине информација узете су у обзир само академске и популарне књиге. У овом раду се износи невероватан податак да би библиотека Јејла 2040. године могла да има 200 милиона књига, за које би било потребно 6 миља[2] полица да се распореде. Назив овог занимљивог рада у оригиналу је The Scholar and the Future of the Research Library.

Назив *велики подаци* (у истом значењу као и данас, али написано малим словима), први пут је употребио Ерик Ларсон (енг. *Erik Larson*) 1989. године. Наиме, Ларсон је овај израз употребио анализирајући зашто се појављује нежељена пошта (енг. *spam*) коју је добијао.

Године 1997. јавља се *Гугл претрага* (енг. *Google Search*), која од тада постаје синоним за претрагу података на интернету.

Сигурно сте до сада пуно пута чули како живимо потпуно окружени подацима, као и да сами свакодневно креирамо велику количину података. Како то појединац направи велику количину података? Сваким изласком на интернет, једним провлачењем платне картице, писањем мејла, једним твитом, било којим кликом на интернету, сваким оваквим или сличним потезом креирамо податке. Међутим, неупоредиво већу количину података од самог појединца креирају различите државне институције, фондови, организације. Овакви подаци се налазе у машински читљивом облику.



[2] 9, 66 км.

Такође, чули сте како је то изузетно драгоцено знање и вештина, која појединцу омогућава да на основу претходно прикупљених података предвиди одређене догађаје у будућности. Ово слободно можемо назвати савремена видовитост заснована на добром познавању математике.

Оно што вероватно не знате, будући да читате ову књигу, јесте да ли и како свако од вас може да научи да, анализирајући податке, предвиди будућа дешавања. Најбоље је да кренемо редом.

Основни елементи науке која се бави обрадом података (енг. *Data Science*), тј. *науке о ɵодацима,* заправо су једноставни. Чини је скуп методологија за прикупљање велике количине података који су нам данас доступни за доношење одговарајућих закључака. Како бисмо успели да на основу података дођемо до неких закључака, потребно је да имамо прикупљене податке, да знамо да их обрадимо, да умемо да их визуализујемо, и на крају, да умемо да их тумачимо.

Споменули смо потенцијалну везу између статистике (примењене статистике) и науке о подацима. Незахвално је давати дефиниције и није чак ни нужно; значајно је једино имати неки општи контекст шта представља једна, а шта друга област, и где је граница где се ове две области сусрећу. Статистика је математички засновано поље које настоји да прикупи и тумачи квантитативне податке (податке представљене помоћу бројева). Наука о подацима је мултидисциплинарно поље које користи научне методе, процесе и системе за доношење закључака из података у различитим областима.

Овде треба напоменути да, на основу неких других дефиниција, постоји много преклапања између поља статистике и науке о подацима, до те мере да многе дефиниције једне дисциплине могу исто тако лако описати другу дисциплину. Можемо се свакако сагласити да је за успешно предвиђање нових трендова на основу анализе података потребно добро познавање математике.

Данас је присутан термин *ɵ̄творени ɵодаци* и већина значајних закључака и предвиђања у сфери маркетинга, индустрије, здравства, глобалне политике доноси се на основу обраде и анализе отворених података. Добро питање у овом тренутку јесте шта су то отворени подаци.

О отвореним подацима

Када су подаци које сакупља нека институција или појединац који то чини јавно у одређене сврхе (научна истраживања, редовна мерења стања животне околине, индустријска мерења и др.) објављени тако да могу да их користе и други људи и организације, они тада постају отворени подаци.

Тиме се постиже да се отворени подаци користе вишеструко, у различитим контекстима и од стране много ширег круга корисника него што су то организације или особе које су прикупиле податке.

Зашто су отворени подаци битни за појединца и шта ће они нама?

Да ли знате тачно колико се новца од пореза троши на улично осветљење или на истраживања за откриће лека против канцера? Која је најкраћа, најсигурнија и најсликовитија бициклистичка рута од ваше куће до оближњег парка? Каквог је састава ваздух који успут удишемо? Где ћете у свом региону пронаћи најбоље могућности за посао и највећи број воћака по глави становника? Када и како можете утицати на одлуке о темама до којих вам је стало и са ким би требало разговарати на ту тему?

Нове технологије сада омогућавају изградњу услуга које аутоматски одговарају на ова питања. Велики део података који су вам потребни да бисте одговорили на ова питања прикупљају јавне установе. Међутим, често потребни подаци још увек нису доступни у облику који је лак за употребу. Такође је значајно рећи да није велики број људи обучен да се позабави обрадом и анализом отворених података уколико им то није баш уска специјалност. Помак се дешава у нашем образовању увођењем ове области у градиво информатике за основне школе и гимназије, али сигурно је да ће проћи још доста времена пре него значајан број појединаца дође до ступња да уме самостално да тумачи велике количине података. Овде треба рећи да наше познавање руковања са подацима није на ништа нижем ступњу од земаља у окружењу.

Како је почело отварање података у свету? Појам отворених јавних података, којима је приступ слободан и који се могу употребљавати у било коју сврху, постоји већ неколико година. У 2009. години отворени подаци почели су да постају доступни у многим земљама, а разне владе (попут САД, Велике Британије, Канаде и Новог Зеланда) најављивале су нове иницијативе да још већа количина података постане јавна и доступна. Отворени подаци су изузетно значајни и могу имати позитиван утицај у многим различитим областима.

Отворени подаци су подаци које свако може слободно да користити, може да их обрађује и даље дистрибуира. Нећемо се бавити прецизном дефиницијом отворених података, али је добро да се упознамо шта нам доносе отворени подаци.

- Подаци морају бити потпуно доступни, по могућству да се могу преузети путем интернета. Подаци такође морају бити доступни у одговарајућем, машински читљивом формату.
- Могуће је обрађивати ове податке и дистрибуирати их, морају бити дати под условима који дозвољавају поновну употребу и

прерасподелу, укључујући спајање више различитих скупова података и њихово такво коришћење.

- Сви морају бити у могућности да их користе, поново употребљавају и поново дистрибуирају, не би требало бити дискриминације према областима деловања или према особама или групама. На пример, „некомерцијална" ограничења која би спречила „комерцијалну" употребу или ограничења употребе у одређене сврхе (нпр. само у домену образовања) нису дозвољена.

Ако се сада питате зашто се толико инсистира на појму отвореност и приступачност за све, одговор је интероперабилност.

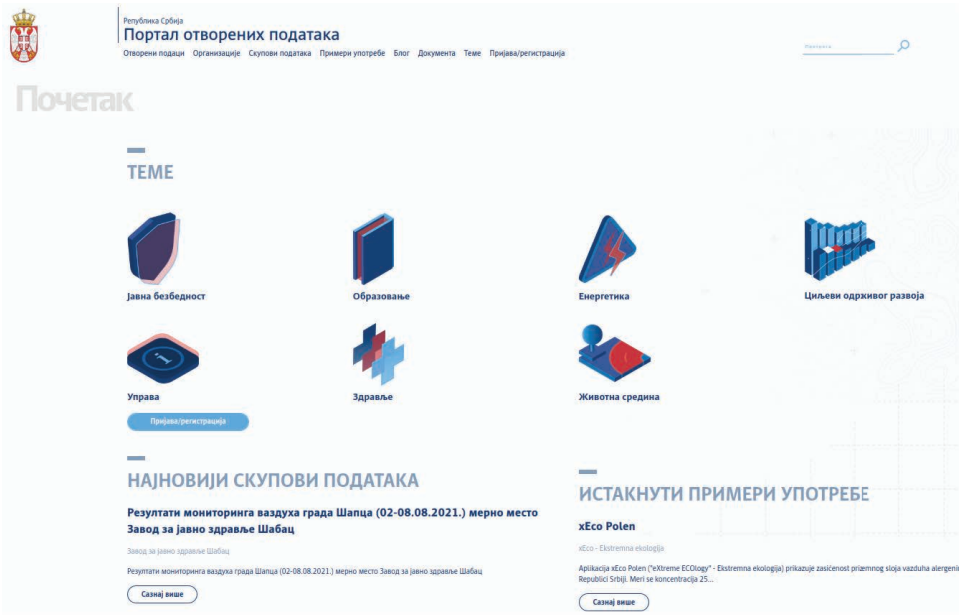
Интероперабилност означава способност различитих система и организација да раде заједно (међусобно сарађују, колаборирају). У овом случају то је могућност интероперације – или мешања – различитих скупова података. Интероперабилност је важна јер омогућава да различите компоненте раде заједно. Ова способност састављања и „повезивања" компоненти је неопходна за изградњу великих, сложених система. Без интероперабилности ово постаје готово немогуће, о чему сведочи најпознатији мит о Вавилонској кули, где је (не)способност комуникације (интероперације) резултирала потпуним сломом изградње торња. Суочени смо са сличном ситуацијом у погледу података. Срж „заједничког" податка (или кода) јесте да се један део „отвореног" материјала који се у њему налази може слободно користити унутар другог или других отворених материјала. Ова интероперабилност је апсолутно кључна за остваривање главних практичних предности „отворености": драматично побољшане способности комбиновања различитих скупова података, а тиме и развој већег броја бољих производа и услуга. Омогућавање јасне слике о томе шта је то отвореност података осигурава да ћете, када добијете два отворена скупа података из два различита извора, моћи да их комбинујете заједно, уколико је то потребно за ваше даље истраживање. Чак и ако имамо приступ многим скуповима података, али немамо могућност да их комбинујемо у веће системе, можемо остати без праве вредности поседовања ових ресурса.

Јесу ли сви подаци отворени подаци? Досад су неки од вас видели, док други нису, примере отворених података.

Видећете их у сваком случају ускоро унутар ове књиге. Међутим, биће корисно брзо описати које врсте података су, или би могле бити, отворене. Такође, једнако је важно знати и шта то нису отворени подаци. Кључна ствар јесте да се приликом отварања података води рачуна о заштити личних података. Слично, за неке посебне податке који су посед влада неких од држава, може се прогласити да су од националног значаја и да се не смеју објавити, попут података који су безбедносног карактера.

Многе државе су, зарад добробити својих грађана, заинтересоване да се подаци које су прикупиле што више искористе. Неке фирме чувају своје податке да би себи обезбедиле предност на тржишту (иако има и фирми које налазе интерес у отварању података). Зато су највећи портали на којима се објављују отворени подаци државни портали, портали међународних организација и портали великих државних институција. Национални Портал отворених података (data.gov.rs) представља место на коме се објављују информације о скуповима отворених података државних органа Републике Србије, али и свих осталих субјеката са подацима који задовољавају услове да буду отворени. То обухвата органе државне управе, аутономних покрајина и локалних самоуправа, као и организације којима је поверено вршење јавних овлашћења – влади, министарствима, општинама, агенцијама, регулаторним телима, установама, комуналним предузећима и другима. Успостављање оваквих портала је стандардна пракса широм света.

Отвореним подацима можете приступити на адреси https://data.gov.rs/sr/.



Такође, адреса која је релевантна за отворене податке код нас је сајт Завода за статистику https://opendata.stat.gov.rs/odata/.