Відкриті дані в Україні

Навчальний посібник

Олександр Краковецький

Відкриті дані в Україні

Навчальний посібник

Олександр Краковецький

This book is for sale at http://leanpub.com/opendatainua

This version was published on 2017-03-20

Зміст

Вступ

Мета курсу - ознайомити представників державних структур та представників громадського сектору (активістів, членів громадських організацій, журналістів тощо) з основними поняттями, принципами та підходами щодо роботи з даними з метою їх підготовки та викладення у форматах, що відповідають принципам відкритості та зручності обробки програмними засобами, показати переваги окремих, найбільш популярних, форматів відкритих даних, з огляду на міжнародну практику роботи з даними, а також особливості роботи з відкритими даними в Україні.

Також ці методичні матеріали будуть цікавими представникам IT сектору та бізнесу, які мають намір аналізувати публічні дані для працювати з суспільно-корисними ініціативами.

Про стан розвитку відкритих даних в Україні

За останні два роки Україна досягла великого прогресу в сфері відкритих даних, але все ще знаходиться досить далеко в міжнародних рейтингах.

Запущено єдиний державний портал відкритих даних data.gov.ua

Перша версія порталу була розроблена волонтерами із організації Socialboost за підтримки міжнародних організацій і компанії Microsoft. На сьогоднішній день портал переданий на баланс Державного агентства з питань електронного урядування. На порталі доступно майже 6000 наборів даних від 700+ розпорядників, хоча якісних наборів даних у відсотковому співвідношенні не дуже багато.

Запущено систему держваних закупівель ProZorro

ProZorro[1] – електронна система публічних закупівель яка прийшла на зміну паперовим держтендерам.

У Законі України «Про публічні закупівлі» передбачається:

- Запровадження обов'язковості проведення процедур через електронну систему. (Перший етап — обов'язковість проведення процедур через електронну систему поширюється на головних розпорядників коштів та монополістів (з 1 квітня 2016 року), на другому (з 1 серпня) – на всіх замовників;
- Запровадження електронного аукціону, який передбачає автоматичну оцінку тендерних пропозицій;
- Визначення нових понять «авторизований електронний майданчик», «електронна система закупівель», «централізована закупівельна організація», «система хмарних обчислень»;
- Замість 5 процедур залишити 3 (відкриті торги, конкурентний діалог, переговорна процедура»;
- Зміна термінології, зокрема, замість терміну «державна закупівля» вводиться термін «публічна закупівля»; замість термінів «конкурс», «документація конкурсних торгів», «пропозиція конкурсних торгів», «комітет з конкурсних торгів», вводяться поняття «тендер», «тендерна документація», «тендерна пропозиція», «тендерний комітет».

[1] https://prozorro.gov.ua

ProZorro отримала міжнародну премію у сфері публічних закупівель Public Sector Procurement Award за створення і впровадження електронної системи з унікальною архітектурою. Розробка цієї системи на базі програмного забезпечення з відкритим кодом була здійснена у партнерстві влади, бізнесу та громадськості та адмініструвалася антикорупційною організацією Transparency International Україна[2]. (с)

З 1 серпня 2016 року ProZorro є обов'язковою системою для всіх державних замовників при закупівлі від 200 тис. грн для товарів і від 1,5 млн грн для робіт.

Запущено офіційний портал публічних фінансів України Є-Data

Є-Data — це офіційний державний інформаційний портал у мережі Інтернет, на якому оприлюднюється інформація про використання публічних коштів та реалізується ідея «Прозорого бюджету». Задача - забезпечити повну прозорість державних фінансів та задовольнить право громадськості на доступ до інформації.

Основні законодавчі та нормативні засади для створення проекту:

- Закон України «Про відкритість використання публічних коштів»[3]
- План заходів з виконання Програми діяльності Кабінету Міністрів України та Стратегії сталого розвитку "Україна - 2020" у 2015 році п. 95: «Впровадження інтегрованої інформаційно-аналітичної системи "Прозорий бюджет" з метою забезпечення доступності інформації про державні фінанси для суспільних потреб із забезпеченням відкритої звітності за всіма коштами, використаними отримувачами бюджетних коштів..»
- Коаліційна угода Парламенту VIII скликання п. 2.5.10.: «Запровадження системи «Прозорий бюджет» з метою забезпечення доступності інформації про державні фінанси для суспільних потреб.

З 15 вересня 2015 року на порталі оприлюднюються всі трансакції Державної казначейської служби, з листопада на ньому доступна інформація про використання коштів державного і місцевих бюджетів, а у січні інформацію почали розкривати суб'єкти господарювання державної і комунальної власності, у статутному капіталі яких державна або комунальна частка акцій (часток, паїв) перевищує 50 відсотків.

До 2018 року планується забезпечити повний функціонал Інтегрованої інформаційно-аналітичної системи «Прозорий бюджет», яка своєю чергою торкнеться змін бюджетних процесів Міністерства фінансів, автоматизації систем ДФС та Казначейства, автоматизації систем обліку та звітності на місцевих рівнях.

Запущено пошуково-аналітичну систему 007

[2] https://uk.wikipedia.org/wiki/%D0%A2%D1%80%D0%B0%D0%BD%D1%81%D0%BF%D0%B5%D1%80%D0%B5%D0%BD%D1%81%D1%96_%D0%86%D0%BD%D1%82%D0%B5%D1%80%D0%BD%D0%B5%D1%88%D0%BD%D0%BB

[3] http://zakon3.rada.gov.ua/laws/show/183-19

Пошуково-аналітична система 007[4] – це веб-сервіс на основі відкритих даних про використання публічних коштів. Проект передбачає сервіс пошуку та візуалізації даних з відкритих джерел про використання державою бюджетних коштів. Його ідея – дати громадськості інструмент контролю влади та можливість стежити за бюджетними витратами, збирати докази зловживань і швидко переводити боротьбу з корупцією в правове поле. Основний акцент зроблено на простоті використання та представлення специфічної інформації з масивів великих даних. Сайт був відкритий 8 квітня 2016 року.

Нормативно-правове забезпечення

9 квітня 2015 року Верховна Рада України ухвалила Закон України № 319 «Про внесення змін до деяких законів України щодо доступу до публічної інформації у формі відкритих даних[5]». Зазначеним Законом внесені зміни до Закону України «Про доступ до публічної інформації» з метою визначення базових норма та засад розвитку відкритих даних в Україні, а саме:

1. Публічна інформація у формі відкритих даних - це публічна інформація у форматі, що дозволяє її автоматизоване оброблення електронними засобами, вільний та безоплатний доступ до неї, а також її подальше використання;
2. Розпорядники інформації зобов'язані надавати публічну інформацію у формі відкритих даних на запит, оприлюднювати і регулярно оновлювати її на єдиному державному веб-порталі відкритих даних та на своїх веб-сайтах;

3.Будь-яка особа може вільно копіювати, публікувати, поширювати, використовувати, у тому числі в комерційних цілях, у поєднанні з іншою інформацією або шляхом включення до складу власного продукту, публічну інформацію у формі відкритих даних з обов'язковим посиланням на джерело отримання такої інформації.

21 жовтня 2015 року затверджено Постановою КМУ № 835 «Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних[6]», якою визначено вимоги до формату і структури наборів даних, а також затверджено перелік пріоритетних наборів даних, які підлягають оприлюдненню (більше 300 наборів). Постановою чітко визначений перелік форматів для оприлюднення відкритих даних в залежності від їх виду:

Текстові дані TXT, RTF, ODT, DOC(X), PDF (с текстовим змістом, скановане зображення), (X)HTML

Структуровані дані RDF, XML, JSON, CSV, XLS(X), ODS, YAML

Графічні дані GIF, TIFF, JPG (JPEG), PNG

Видеодані MPEG, MKV, AVI, FLV, MKS, MK3D

Аудіодані MP3, WAV, MKA

[4] https://www.facebook.com/pointOOSeven

[5] http://zakon3.rada.gov.ua/laws/show/319-19

[6] http://zakon5.rada.gov.ua/laws/show/835-2015-%D0%BF

Дані, розроблені з використанням програми Macromedia Flash SWF, FLV

Архіви даних ZIP, 7z, Gzip, Bzip2

Також, постановою передбачено, що державні реєстри, які постійно оновлюються, мають бути відкриті за допомогою API.

З метою забезпечення ефективного функціонування Єдиного державного веб-порталу відкритих даних та підвищення відкритості і прозорості діяльності органів виконавчої влади та місцевого самоврядування, Державним агентством з питань електронного урядування в лютому 2016 року розроблено проект постанови Кабінету Міністрів України «Деякі питання оприлюднення публічної інформації у формі відкритих даних», які знаходяться зараз на етапі обговорення.

Оцінка готовності України до розвитку відкритих даних

З метою визначення поточного стану розвитку та готовності України до розвитку відкритих даних, а також планування подальший дій, проведено оцінку готовності України до розвитку відкритих даних за методикою Всесвітнього банку ODRA.

Оцінка проводилась за восьма напрямами:

1. Зобов'язання Уряду;
2. Політичні і правові засади;
3. Інституційна структура, розподіл відповідальності та спроможність урядових структур;
4. Політика та процедури Уряду стосовно обробки даних;
5. Попит на відкриті дані;
6. Залучення громадського сектору та можливості для відкритих даних;
7. Фінансування політики відкритих даних;
8. Національна технологічна інфраструктура та навички.

Якщо коротко, з готовністю громадськості використовувати відкриті дані все добре, більшменш хороша ситуація з готовністю уряду відкривати дані, технічною інфраструктурою, але все погано в плані фінансування, індустрією розробки продуктів на базі відкритих даних і захистом приватності. Детальний звіт знаходиться за посиланням[7] (є звіти на українській і англійській мовах).

Дорожня карта розвитку відкритих даних

З метою забезпечення комплексного розвитку відкритих даних Державним агентством з питань електронного урядування напрацьовано Дорожню карту розвитку відкритих даних в Україні, яка містить 41 завдання по 5 напрямкам:

1. Підвищення доступності та якості відкритих даних;

[7] http://dhrp.org.ua/uk/publikatsii1/1071-20160227-ua-publication

2. Розвиток спроможності органів влади щодо публікації відкритих даних;
3. Посилення ролі відкритих даних в реалізації державної політики;
4. Нормативно-правове забезпечення;
5. Розвиток попиту та спроможності цільових аудиторій щодо використання відкритих даних.

Зазначений документ затверджено наказом Мінрегіону від 04.02.2016 року № 19. Повний текст розміщено за посиланням[8].

Хартія відкритих даних

Наразі Україною ініціюється приєднання до міжнародної Хартії відкритих даних.

Проект Постанови КМУ знаходиться за посиланням[9].

Розробка Міжнародної хартії була ініційована представниками урядів Канади, Мексики, Великобританії, впливових міжнародних організацій у травні 2015 року під час міжнародної конференції з питань відкритих даних у Канаді.

Головною метою Міжнародної Хартії відкритих даних є покращення та сприяння співпраці та взаємоузгодженості для прийняття та реалізації спільних принципів, стандартів та кращих практик відкритих даних по всьому світу. Цілями Хартії є поширення демократії, боротьба з корупцією та сприяння економічному зростанню по всьому світу. Хартія визначає 6 головних принципів та шляхи розвитку відкритих даних для країни.

Світові та українські компетенції й рейтинги

Сьогодні найбільш важливими є наступні два рейтинги оцінювання стану розвитку відкритих даних:

1. Open Data Barometer[10]. За цим рейтингом Україна посідає 62 місце.
2. Open Data Index[11]. За цим рейтингом Україна посідає 54 місце (із 122).

Україна співпрацює з такими міжнародними організаціями в сфері відкритих даних:

1. Open Data Institute[12]. Займається розвитком відкритих даних по всьому світу, формуванням стандартів та єдиних підходів, розвитком компетенцій.
2. Open knowledge foundation[13]. Займається підтримкою громадських інституцій щодо розвитку відкритих даних, а також розвитком відкритої платформи для побудови порталів відкритих даних CKAN.

[8] https://drive.google.com/file/d/0B1kGsKt9XV_QaFZVaTZiT19aRTA/view.
[9] https://drive.google.com/file/d/0B1kGsKt9XV_QMEFNNklzRXdaSDA/view?usp=sharing
[10] http://opendatabarometer.org/data-explorer/?_year=2015&indicator=ODB&open=UKR
[11] http://index.okfn.org/place/ukraine
[12] http://theodi.org
[13] https://okfn.org

3. Open Data for Development[14]. Займається підтримкою ініціатив з розвитку відкритих даних по всьому світу, а також організацію міжнародної співпраці.

В останньому рейтингу E-Government Development Index (EGDI) 2016 Україна зайняла 62 місце серед 193 країн, покращивши свою позицію на 25 пунктів.

Професійні українські організації і ініціативи

В Україні запущений інкубатор відкритих даних 1991, який системно займається відбором проектів, їх інкубацією і пошуком інвестицій. На сьогодні уже було два набори в інкубаційну програму.

На початку року розпочався EGAP Challenge – конкурс IT проектів в області електронної демократії, соціальної сфери і проектів в сфері відкритих даних. Спільна ініціатива Державного агентства з питань електронного урядування, Фонду Східної Європи в рамках реалізації Програми EGAP, що фінансується Швейцарською Конфедерацією, має на меті не тільки запровадити нові інструменти e-democracy в чотирьох регіонах України (Вінницький, Волинській, Дніпровській та Одеській областях), а також показати українським стартапам нову нішу, в якій можливо створювати якісні проекти, що мають змогу безпосередньо впливати на місцеву і центральну влади. А влада – у свою чергу – отримає набір нових інструментів для росту ефективності її роботи і прозорості взаємодії з платниками податків.

Пріоритетними є напрямки:

- створення нових інструментів взаємодії влади і суспільства, особливо в частині надання можливості громадянам напряму впливати на процеси прийняття управлінських рішень;
- підвищення прозорості і відкритості діяльності органів влади, особливо в частині формування і виконання бюджетів, надання дозвільних документів і тому подібне;
- створення нових якісних сервісів для громадян і бізнесу, особливо в частині надання публічних послуг;
- розвиток проектів на базі відкритих даних;
- вирішення соціальних проблем;
- об'єднання і налагодження ефективного співробітництва громадян для вирішення загальних проблем;
- проекти в області Smart City;
- галузеві проекти, спрямовані на підвищення ефективності державного управління і обслуговування громадян та бізнесу (е-екологія, е-медицина, е-освіта і т.д.).

В чотирьох областях пройшли креативні уікенди на яких було відібрано 15 проектів для проходження двомісячної інкубації. Партнерами конкурсу виступили компанії Cisco, IBM, DeNovo, Intel.

[14] http://od4d.net

В Україні діють потужні громадські організації, що працюють з відритими даними— ОПОРА, Чесно, Канцелярська сотня, Vox Ukraine і інші. Результат — десятки досліджень, проекти по візуалізації даних, моніторинг відкритих наборів даних, десятки заходів по усій країні. Декілька проектів уже імплементовані в державні служби і проекти “смарт сіті”.

Наприклад, відкриті дані Укрзалізниці дали можливість провести масштабне дослідження – хто куди подорожує, які ключові станції, як розподіляється потік пасажирів і інші візуалізації.

Рекомендації по роботі з відкритими даними

На сайті Верховної Ради є інструкція[15] по роботі з відкритими даними.

Також спільно з Агентством з питань електронного врядування та ПРООН були розроблені методичні рекомендації щодо оприлюднення наборів даних у формі відкритих даних[16].

Метою рекомендацій є ознайомлення відповідальних осіб розпорядників інформації з ключовими питаннями, що постають у процесі оприлюднення наборів даних у формі відкритих даних. Також, методичні рекомендації будуть корисні розробникам застосунків на базі відкритих даних, громадським організаціям, засобам масової інформації та громадянам, які мають намір працювати з відкритими даними.

Проект першої версії Методичних рекомендацій підготовлено на основі найбільш популярних загальних запитань від розпорядників інформації, але, безумовно, ще не охоплює усіх піднятих розпорядниками інформації та громадськістю питань. Методичні рекомендації будуть систематично оновлюватися з метою забезпечення розпорядників інформації необхідною інформацією для оприлюднення наборів даних.

В цілому, Україна знаходиться лише на початку з точки зору відкритості державних служб, створення інститутів моніторингу діяльності депутатів, чиновників і міських служб, дигіталізації державних послуг. За наявності політичної волі, допомоги міжнародних організацій, а також за умови координації громадських організацій, бізнесу та IT сектору Україна здатна на досить короткий термін значно поліпшити свої показники в області електронного врядування та відкритих даних.

[15] http://data.rada.gov.ua/open/main/opendata

[16] https://drive.google.com/file/d/0B1kGsKt9XV_QWjhaZ0ZMVmFiUE0/view

Що таке відкриті дані?

Визначення відкритих даних

Відкриті дані (англ. Open Data) – це концепція, яка відображує ідею, що визначені дані мають бути доступні для легкої обробки програмними засобами (machine readable) та подальшого використання і розповсюдження без жодних обмежень і контролю, в тому числі й для комерційного використання. Відкриті дані – це не просто інформація, а концепція, тобто система поглядів, підходів, процесів, які мають одну ідею та мету – вільного використання і розповсюдження даних про діяльність державних органів та органів місцевого самоврядування через мережу Інтернет.

Згідно загальноприйнятої в світі концепції, обов'язково безоплатними можуть бути тільки ті дані, які знаходяться у власності держави («ліцензійно чисті»), і якщо вони подаються в первинному необробленому вигляді. Додаткова обробка або доступ до API частіше за все лімітуються або коштують грошей. Крім того, хто завгодно може (з обов'язковим посиланням на джерело) використати дані в комерційний спосіб, створити на базі них власну програму чи обробити та надати нову цінність (розкласифікувати, встановити зв'язки тощо). Безумовно, всі витрати на додаткову роботу редакторів, програмістів та інше оплачує замовник (або кінцевий користувач).

Відкриті дані дозволяють повторно і необмежено використовувати інформацію, поєднувати її між собою, зменшити або виключити зайві витрати на дублювання та опрацювання великих масивів даних, реєстрів, довідників, баз даних тощо, створених в різних органах влади.

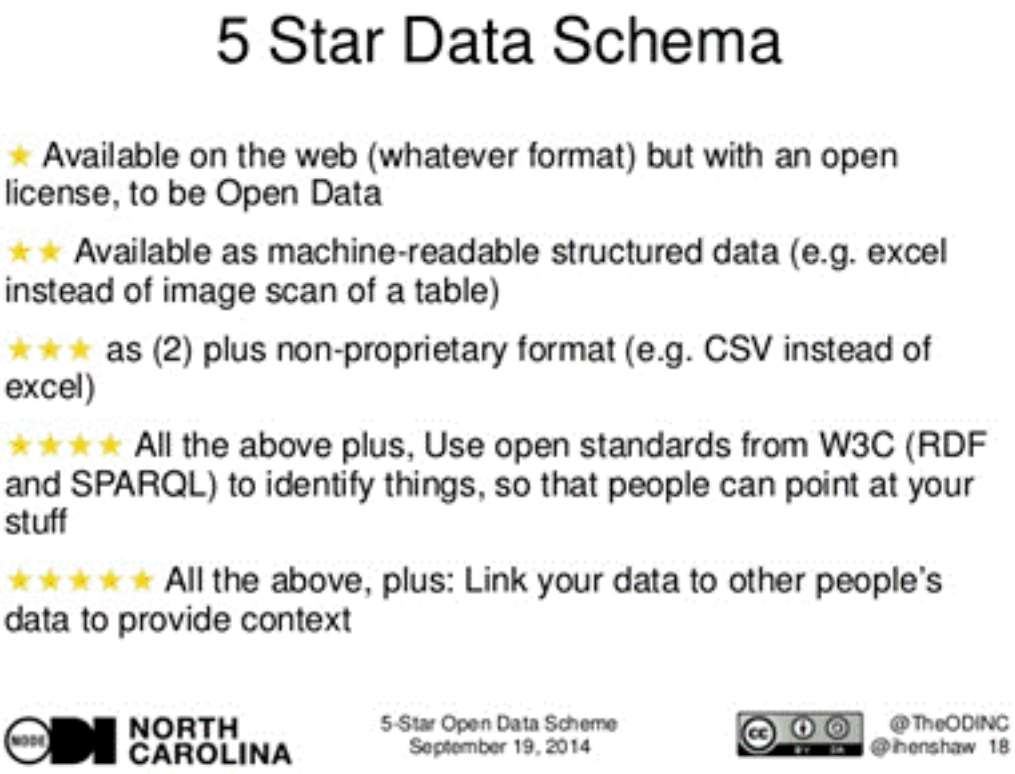
Закон "Про внесення змін до деяких законів України щодо доступу до публічної інформації у формі відкритих даних" №2171 від 19.02.2015 (прийнятий 09.04.2015) визначає наступний термін:

Публічна інформація у формі відкритих даних – це публічна інформація у форматі, що дозволяє її автоматизоване оброблення електронними засобами, вільний та безоплатний доступ до неї, а також її подальше використання.

Класифікація відкритих даних

В законопроекті одночасно використовуються поняття «форма» та «формат», що не є тотожними. Розглянемо детально.

Для того, щоб зрозуміти, які можуть бути форми відкритих даних, ми також звернемося до відомої класифікації **«5 зірок Open Data»** (http://5stardata.info/[17]), де якість даних та рівень відкритості визначається кількістю зірок від 1 до 5, чим більше – тим краще. Відкритість даних залежить від способів доступу, форматів та кількості додаткових дій, які потрібні для отримання кінцевої інформації, її обробки та збереження у власному сховищі або базі даних.



Одну зірку (*) отримує будь-яка інформація вільно доступна через Інтернет в будь-якому форматі. Під цю класифікацію підпадає файл в форматі PDF або інша (сканована) копія документу, на який веде пряме посилання на офіційному сайті державного органу. Якщо цей файл можна відкрити на власному екрані, прочитати, роздрукувати та отримати звідти потрібну інформацію, то це відкриті дані з однією зіркою.

Дві зірки ()** отримує структурована інформація, яку можна обробляти автоматично, наприклад, в форматах для веб-браузерів чи офісних програм (відкриті формати – TXT, HTML, RSS; пропрієтарні формати, Excel – XLS, Word – DOC, RTF). Якщо дані знаходяться в тілі вихідної веб-сторінки, але не мають чіткої структури, містять зайві елементи оформлення, навігації, якщо дані потрібують додаткових дій – спеціального розбору (парсингу), то вони вважаються «з двома зірками».

[17] http://5stardata.info/

Три зірки (*)** може отримати інформація, представлена у відомих, добре описаних відкритих структурованих форматах (наприклад, CSV, JSON, XML, YAML) і якщо автоматизована її обробка не потребує від користувача особливих ліцензій та додаткових плат. До відкритих форматів також відносяться пов'язані дані (HTML+RDFa) з узгодженою розміткою елементів в атрибутах (див. сніппет для пошукових систем) або текстові файли таблиць, поля яких розділені табуляцією, комами, крапками з комою або іншими символами.

Чотири зірки () надаються у випадку, якщо можна отримати первинні необроблені набори відкритих даних у вигляді файлів (довідники, списки, таблиці у відкритому форматі, зліпок бази даних, архів документів тощо) або фільтровані дані у запиті до API за вказаними параметрами. Це дає змогу отримувати тільки потрібну інформацію, актуальну на момент запиту, заощаджує ресурси та час користувача. Безумовно, API має бути описаний так само, як і формати даних, а доступ до нього може бути анонімний без обмежень або з реєстрацією, за вказаним ідентифікатором, лімітами на кількість одночасних запитів тощо.

Останній рівень – **п'ять зірок (*)** – надається інформації, коли набори відкритих даних пов'язані між собою (мають спільні довідники, класифікатори, ідентифікатори, посилання між документами та іншими елементами тощо) і представляють собою семантичну мережу, що постійно оновлюється й змінюється відповідно до сучасних запитів.

Формати даних

В залежності від специфіки даних, їх розміру та тематики (геологія, тендери, реєстри, судові документи тощо), одні проекти відкритих даних створювались на базі наборів PDF чи DOC файлів, таблиць XLS, що перетворювались на прості текстові таблиці CSV, а інші брали за основу формат розмітки XML, проектували власні схеми XSD і використовували складні структури.

Як свідчить остання статистика використання форматів відкритих даних, найбільш поширений в світі формат (як по кількості, так і по об'єму даних) – PDF. Для українських органів влади, де найбільш розповсюджені операційні системи Microsoft Windows, переважають формати DOC та XLS. Разом з новими версіями офісних програм в Інтернет почали з'являтися документи DOCX та XLSX, рідко ODF (*Open Document Format*). Як надбання DOS'івського періоду, поки ще зустрічаються документи в старому форматі Lexicon (TXT) або ранніх версій Word (RTF). Після поширення ініціативи відкриття державних даних та створення порталів, кількість наборів в форматі XML та інших відкритих форматах почала суттєво збільшуватись.

Необроблені дані, сформовані державними структурами за багато років, можуть бути досить неоднорідними, а деякі набори навіть дублюються в різних форматах для зручності користування.

Серед доступних в цей час форматів відкритих даних, які можна автоматично обробляти електронними засобами, є: CSV (текстові дані, розділені комами або іншими розділовими

символами), JSON (формат, орієнтований на обробку складних структурованих даних за допомогою javascript) та XML (універсальний текстовий формат розмітки). Є багато форматів, що основані на XML, зокрема, HTML, KML та інші.

Табличні формати

CSV (Comma-Separated Values)

CSV (від англ. *Comma-Separated Values* – значення, що розділені комами) – текстовий відкритий формат, призначений для представлення таблиць (масивів, наборів) даних, де кожний рядок – це запис таблиці, а значення окремих полів у рядку розділені спеціальними символами (delimiter), зазвичай комами. Щоб завантажити записи таблиці за найменуванням полів (а не за порядковими номерами), додатково потрібно мати опис її структури – назви та формат полів.

Більшість програм широко трактують цей формат і допускають використання інших розділових символів, наприклад, табуляції (TSV) чи коми з крапкою.

Приклад даних в форматі CSV:

```
1997,Ford,E350,”ac, abs, moon”,3000.00
1999,Chevy,”Venture ””Extended Edition””””,””,4900.00
1996,Jeep,Grand Cherokee,”MUST SELL! air, moon roof, loaded”,4799.00
```

Таблиця в форматі CSV має наступні параметри: кодування (Windows-1251, KOI-8, UTF-8 тощо) і формат кінця рядків. Різниця у символах розділення рядків залежіть від операційної системи (для Unix – один символ CR (CHR 0x0D), для Mac – LF (CHR 0x0A), а для Windows – пара символів CR LF).

Інститут відкритих даних Open Data Institute випустив оновлену версію 0.4 додатка Comma Chameleon https://github.com/theodi/comma-chameleon/releases/tag/0.4.0[18] - з валідації CSV файлів. Ця версія найбільш стабільна з усіх попередніх і існує в версіях для Mac, Linux, Windows і просто як відкритий код в репозиторії https://github.com/theodi/comma-chameleon[19].

Корисний інструмент для всіх хто готує дані для публікації і думає про автоматизацію і спрощення очищення даних.

Корисні сервіси та інструменти для роботи з CSV файлами:

1. Валідація CSV файлів: Comma Chameleon https://github.com/theodi/comma-chameleon[20].
2. CSVLint http://csvlint.io/[21] - онлайн сервіс з валідації CSV файлів і з відкритим кодом https://github.com/theodi/csvlint[22].

[18] https://github.com/theodi/comma-chameleon/releases/tag/0.4.0
[19] https://github.com/theodi/comma-chameleon
[20] https://github.com/theodi/comma-chameleon
[21] http://csvlint.io/
[22] https://github.com/theodi/csvlint

3. CSVkit https://github.com/wireservice/csvkit[23] - бібліотека для Python за численними маніпуляціям з CSV файлами і безліччю інструментів для командного рядка.
4. Textql https://github.com/dinedal/textql[24] - інструмент із запуску SQL запитів на CSV / TSV файлах.
5. PapaParse https://github.com/mholt/PapaParse[25] - парсер дуже великих CSV файлів.
6. https://mledoze.github.io/countries/[26] - країни світу в JSON, CSV, XML і YAML.
7. Tablib https://github.com/kennethreitz/tablib[27] - бібліотека для роботи з будь-якими табличними даними включаючи CSV.

XLS/XLSX (Document Office Open XML)

Файл XLSX - електронна таблиця, створена в Microsoft Excel - додатку для роботи з таблицями. Дані в документі зберігаються в комірках, кожна з яких має певну адресу (колонки нумеруються англійськими літерами, рядки - цифрами, наприклад: A1 - ліва верхня клітинка).

Кожна клітинка може містити як фіксовані дані, так і формули, часто пов'язані з даними в інших осередках. Формат XLSX дозволяє користувачеві змінювати форматування тексту: його шрифт, колір, вирівнювання (в комірці) і інші параметри. Крім того, документ може містити зображення, а також діаграми, побудовані на основі даних в певних комірках.

Програми, які дозволяють працювати з XLSX файлами: File Viewer Lite, Microsoft Excel, Microsoft Excel Viewer, Microsoft Works, Corel WordPerfect Office X7, Apache OpenOffice, Kingsoft Spreadsheets, Nuance OmniPage Ultimate, Gnumeric, LibreOffice.

Спочатку формат XLSX створювався як заміна раніше бінарного формату документів (.xls), який раніше використовували додатки Microsoft Office. Але в 2006 році формат Office Open XML був оголошений вільним і відкритим форматом.

Графічні формати

JPEG (Joint Photographic Experts Group)

JPEG (англ. *Joint Photographic Experts Group*, за назвою організації-розробника) - один з популярних графічних форматів, застосовуваний для зберігання фотозображень і подібних до них зображень. Файли даних JPEG, зазвичай мають розширення .jpg, .jfif, .jpe або .jpeg. Однак з них .jpg є найпопулярнішим на всіх платформах.

Алгоритм JPEG дозволяє стискати зображення як з втратами, так і без втрат (режим стиснення lossless JPEG). Підтримуються зображення з лінійним розміром не більше 65535 × 65535 пікселів.

[23] https://github.com/wireservice/csvkit
[24] https://github.com/dinedal/textql
[25] https://github.com/mholt/PapaParse
[26] https://mledoze.github.io/countries/
[27] https://github.com/kennethreitz/tablib

Алгоритм JPEG найбільшою мірою придатний для стиснення фотографій і картин, що містять реалістичні сцени з плавними переходами яскравості і кольору. Найбільшого поширення JPEG отримав в цифровій фотографії і для зберігання і передачі зображень з використанням мережі Інтернет.

Формат JPEG в режимі стиснення з втратами малопридатний для стиснення креслень, текстової та знакової графіки, де різкий контраст між сусідніми пікселями приводить до появи помітних артефактів. Такі зображення доцільно зберігати в форматах без втрат, таких як JPEG-LS, TIFF, GIF, PNG або використовувати режим стиснення Lossless JPEG.

JPEG не підходить для стиснення зображень багатоетапної обробки, так як спотворення в зображення будуть вноситися кожен раз при збереженні проміжних результатів обробки. JPEG не повинен використовуватися і в тих випадках, коли недопустимі навіть мінімальні втрати, наприклад, при стисненні астрономічних або медичних зображень.

Детальніше про формат: https://ru.wikipedia.org/wiki/JPEG[28]

PNG (portable network graphics)

PNG (англ. *Portable network graphics*) - растровий формат зберігання графічної інформації, що використовує стиснення без втрат за алгоритмом Deflate. Формат PNG розроблений для заміни форматів GIF та TIFF. Формат PNG позиціонується передусім для використання в Інтернеті і редагування графіки. Формат PNG зберігає графічну інформацію в стислому вигляді, причому це стиснення графічної інормації проводиться без втрат.

PNG підтримує три основних типи растрових зображень:

- Півтонування (з глибиною кольору 16 біт)
- Кольорові індексовані зображення (палітра 8 біт для кольору глибиною 24 біт)
- Повнокольорове зображення (з глибиною кольору 48 біт)

Формат має наступні основні переваги:

- практично необмежену кількість кольорів в зображенні;
- опциональная підтримка альфа-каналу;
- можливість гамма-корекції;
- двовимірна черезстрокова розгортка;
- можливість розширення формату для користувача блоками (на цьому заснований, зокрема, APNG).
- будь-яке збережене зображення PNG може бути прочитано в будь-якому іншому додатку, що підтримує PNG.

[28] https://ru.wikipedia.org/wiki/JPEG

PNG є хорошим форматом для редагування зображень, навіть для зберігання проміжних стадій редагування, так як відновлення і повторне збереження зображення здійснюється без втрат в якості.

Детальніше про формат: https://ru.wikipedia.org/wiki/PNG[29]

TIFF (Tagged Image File Format)

TIFF (англ. *Tagged Image File Format*) - формат зберігання растрових графічних зображень. TIFF став популярним форматом для зберігання зображень з великою глибиною кольору. Він використовується при скануванні, відправлення факсів, розпізнаванні тексту, в поліграфії, широко підтримується графічними додатками. TIFF був обраний в якості основного графічного формату операційної системи NeXTSTEP і з неї підтримка цього формату перейшла в Mac OS X.

Спочатку формат підтримував стиснення без втрат, згодом формат був доповнений для підтримки стиснення з втратами в форматі JPEG. Файли формату TIFF, як правило, мають розширення .tiff або .tif.

Структура формату гнучка і дозволяє зберігати зображення в режимі кольорів з палітрою, а також в різних колірних просторах:

- Бінарному (двокольоровому, іноді неправильно званому чорно-білим)
- напівтоновому
- з індексованої палітрою
- RGB
- CMYK
- YCbCr
- CIE Lab

Підтримуються режими 8, 16, 32 і 64 біт на канал при цілочисельному, а також 32 і 64 біт на канал при поданні значення пікселя числами з плаваючою комою.

TIFF є тяговим форматом і в ньому є такі види міток:

- основні мітки;
- розширені мітки;
- спеціальні мітки.

Детільніше про формат: https://ru.wikipedia.org/wiki/TIFF[30]

[29] https://ru.wikipedia.org/wiki/PNG
[30] https://ru.wikipedia.org/wiki/TIFF

Текстові формати

TXT (Textfile)

TXT - це формат, що містить текстові дані, які, як правило, організовані в вигляді рядків; служить основою для багатьох більш спеціалізованих форматів, таких як CHM, PHP, XML, CSV. Може бути переглянутий будь-яким текстовим редактором.

Текстові файли розбиваються на кілька рядків. На сучасних платформах розбивка на рядки кодується керуючим символом переведення рядка, а іноді послідовністю двох керуючих символів. Взагалі, текстові файли можуть містити друковані символи, такі як букви, цифри і розділові знаки і деяку кількість керуючих символів, таких як знаки табуляції і переведення рядка.

Файли TXT містять дуже мало елементів форматування, однак дозволяють співвіднести прийняті набори елементів форматування з системним терміналом або простим текстовим редактором. Файли TXT універсальні за своєю природою, тому що відкрити такі файли в стані будь-який текстовий редактор. При цьому файли TXT можуть використовувати Unicode, яка дозволяє полегшити користувачам використання файлів TXT, які пишуть на різних мовах. Файли TXT, що використовують текст тільки в кодуванні ASCII, можна переносити на різні комп'ютери і пристрої на ОС Unix, Mac і Windows.

Markdown

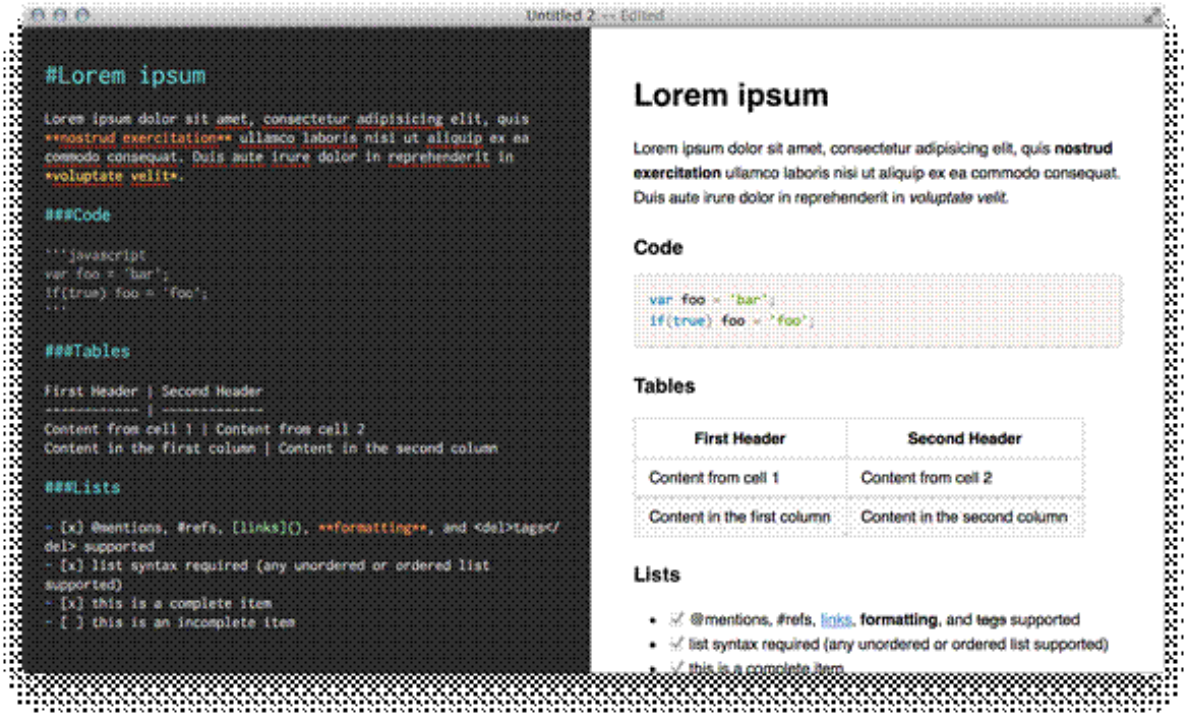
Markdown - полегшена мова розмітки, створена з метою написання максимально читабельного і зручного для редагування тексту, але придатного для перетворення в мови для публікацій (HTML, Rich Text і ін.).

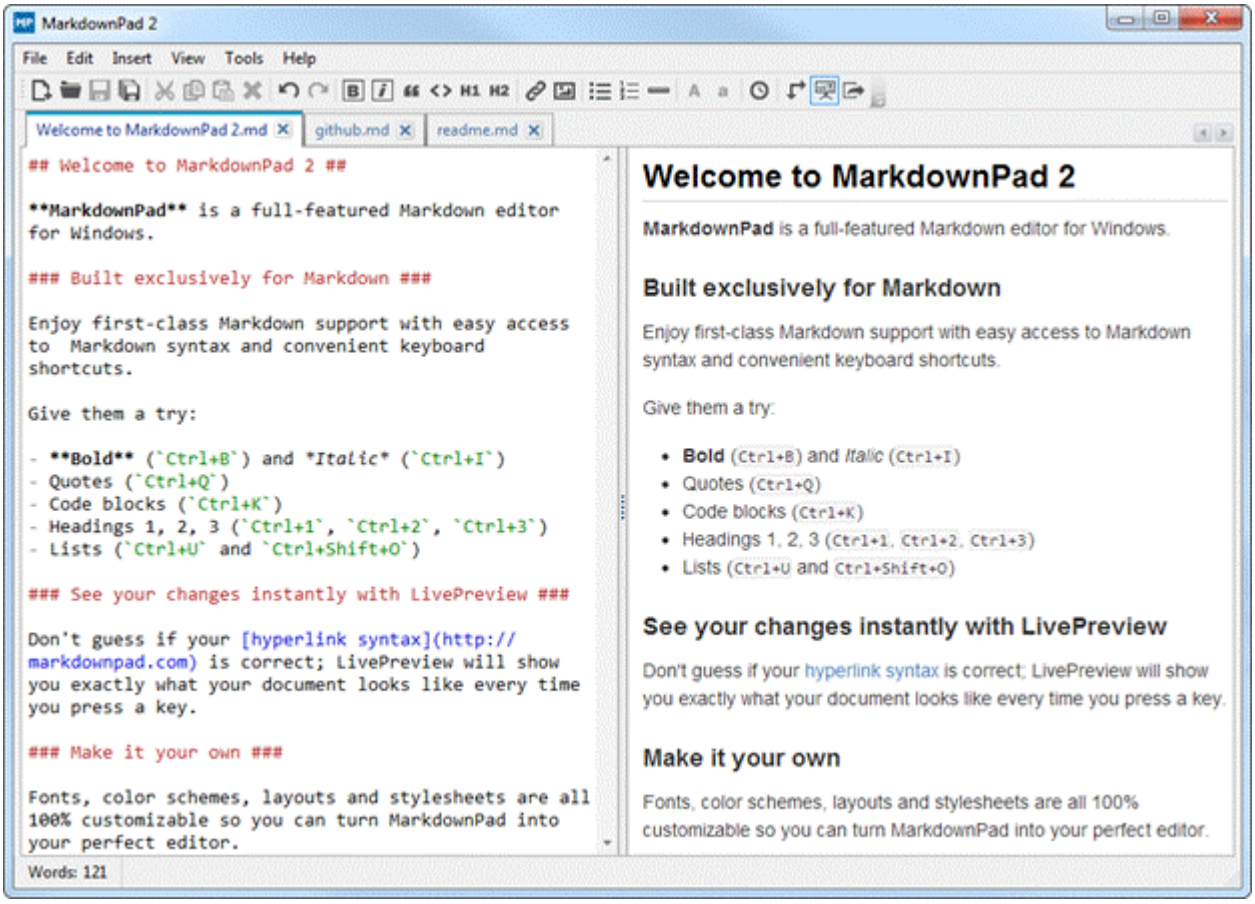
Таким чином, "Markdown" це дві речі:

-простий синтаксис форматування тексту;
-програмний інструмент, написаний на Perl, який перетворює звичайний форматування тексту в HTML.

Першочерговою метою розробки для форматування синтаксису Markdown є зробити його максимально читабельним, наскільки це можливо. Ідея полягає в тому, що Markdown-форматований документ повинен бути опублікуванний як звичайний текст, використовуючи теги та інструкції форматування.

Приклади форматування тексту мовою розмітки Markdown:


Untitled 2 — Edited
#Lorem ipsum
Lorem ipsum dolor sit amet, consectetur adipisicing elit, quis **nostrud exercitation** ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in *voluptate velit*.
###Code
```javascript
var foo = 'bar';
if(true) foo = 'foo';
```
###Tables
First Header | Second Header
Content from cell 1 | Content from cell 2
Content in the first column | Content in the second column
###Lists
- [x] @mentions, #refs, [links](), **formatting**, and ~~tags~~ supported
- [x] list syntax required (any unordered or ordered list supported)
- [x] this is a complete item
- [ ] this is an incomplete item
Lorem ipsum
Lorem ipsum dolor sit amet, consectetur adipisicing elit, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit.
Code
var foo = 'bar';
if(true) foo = 'foo';
Tables
First Header
Second Header
Content from cell 1
Content from cell 2
Content in the first column
Content in the second column
Lists
@mentions, #refs, links, formatting, and tags supported
list syntax required (any unordered or ordered list supported)
this is a complete item



Детальніше про формат та його синтаксис:

1. Markdown http://daringfireball.net/projects/markdown/[31]
2. Wikipedia https://uk.wikipedia.org/wiki/Markdown[32]

Текстово-графічні формати

HTML (HyperText Markup Language)

HTML (від англ. *HyperText Markup Language* - «мова гіпертекстової розмітки») - стандартна мова розмітки документів у Всесвітній павутині. Більшість веб-сторінок містять опис розмітки на мові HTML (або XHTML). Мова HTML інтерпретується браузерами; отриманий в результаті інтерпретації форматований текст відображається на екрані монітора комп'ютера або мобільного пристрою.

У всесвітній павутині HTML-сторінки, як правило, передаються браузерам від сервера по протоколам HTTP або HTTPS, у вигляді простого тексту або з використанням шифрування.

[31] http://daringfireball.net/projects/markdown/
[32] https://uk.wikipedia.org/wiki/Markdown

Текстові документи, що містять розмітку на мові HTML (такі документи зазвичай мають розширення .html або .htm), обробляються спеціальними додатками, які відображають документ в його форматованому вигляді. Такі додатки, звані «браузерами», зазвичай надають користувачеві зручний інтерфейс для запиту веб-сторінок, їх перегляду (і виведення на інші зовнішні пристрої) і, при необхідності, відправки введених користувачем даних на сервер.

HTML - тегова мова розмітки документів. Будь-який документ на мові HTML являє собою набір елементів, причому початок і кінець кожного елемента позначається спеціальними позначками - тегами. Елементи можуть бути порожніми, тобто не містять ніякого тексту та інших даних (наприклад, тег переносу рядка
). В цьому випадку зазвичай не вказується закриваючий тег. Крім того, елементи можуть мати атрибути, що визначають будь-які їх властивості (наприклад, розмір шрифту для елемента font). Атрибути вказуються в відкриваючому тегові.

Детальніше про формат: https://ru.wikipedia.org/wiki/HTML[33]

DOCX (Document Office Open XML)

DOCX (*Document Office Open XML*) - формат файлу для зберігання електронних документів пакетів офісних додатків - зокрема, Microsoft Office. Формат є zip-архів, що містить текст у вигляді XML, графіку і інших даних, які можуть бути переведені в послідовність бітів (серіалізовані) із застосуванням захищених патентами довічних форматів, специфікації яких були опубліковані Microsoft для користувачів на умовах Microsoft Open Специфікація Promise (англ.).

Спочатку формат створювався як заміна раніше бінарного формату документів, який використовували додатки Microsoft Office аж до версії Office 2003 включно. У 2006 році формат Office Open XML був оголошений вільним і відкритим форматом Ecma International. Він є форматом за умовчанням для додатків Microsoft Office 2007 і пізніших.

Файл з розширенням .DOCX - документ, створений Microsoft Word, програмою обробки тексту; містить текст документа, зображення, форматування, стилі, намальовані об'єкти та інші параметри документа.

Файли DOCX створюються за допомогою відкритого формату XML, в якому зберігаються документи, як збори окремих файлів і папок в стислому пакеті. DOCX-файли містять XML-файли і три папки, docProps, Word, і _rels, які містять властивості документа, зміст і відносини між файлами.

Детальніше про формат: https://ru.wikipedia.org/wiki/Office_Open_XML[34]

OpenDocument format (LibreOffice/OpenOffice)

Формат OpenDocument - це формат для текстових, табличних докуметів та презентацій, що був створений з метою підвищити інтероперабельність між різними офісними застосунками

[33] https://ru.wikipedia.org/wiki/HTML
[34] https://ru.wikipedia.org/wiki/Office_Open_XML

(програмами, офісними пакетами).

Історична довідка: процес стандартизації форматів офісних документів почався на початку 2000-х років та був обумовлений становищем на ринку офісних пакетів програм. Microsoft на той момент займала домінантне становище. Європейска комісія за скаргою Sun Microsystems, що тоді розробляла альтернативний офісний пакет OpenOffice.org, вирішила, що Microsoft зловживає своїм становищем та веде нечесну конкуренцію. Microsoft змушена була відкрити специфікацію своїх форматів. Параллельно з цим, відбувалась стандартизація офісних форматів. Результатом зусиль зі стандартизації стало створення двох конкуруючих форматів - Office Open XML (Microsoft) та OpenDocument (Sun Microsystems).

Файли формату OpenDocument мають розширення ODT (OpenDocument Text), ODS (OpenDocument Spreadsheet). Його підтримують наступні пакети програм:

- Apache OpenOffice
- LibreOffice
- MS Office (2013 - тільки читання, 2016 - читання та редагування)
- Google Docs (тільки імпорт/експорт)

PDF (Portable Document Format)

Формат переносного документа (PDF) - це формат файлу, який використовується для надійного уявлення і обміну документами, незалежно від програмного і апаратного забезпечення або операційної системи. В даний час формат PDF, винайдений компанією Adobe, є відкритим стандартом, підтримуваним Міжнародною організацією зі стандартизації (ISO). Файли PDF можуть містити посилання і кнопки, поля форм, аудіо- та відеоелементи, а також бізнес-логіку. Вони також підтримують можливість електронного підпису і можуть переглядатися за допомогою безкоштовного програмного забезпечення Acrobat Reader DC.

Можливості формату:

- мають такий же вигляд, як вихідні документи, незалежно від використовуваної платформи або пристрою;
- відповідають стандартам ISO 32000 з обміну електронними документами, включаючи спеціальні стандарти для PDF;
- можуть бути підписані електронно будь-яким користувачем за допомогою безкоштовного програмного забезпечення Adobe Acrobat Reader DC або мобільного застосування Acrobat DC;
- можуть захищатися паролем для запобігання копіювання та редагування;
- можуть виправлятися для видалення конфіденційної інформації без можливості відновлення;
- зберігають всю інформацію вихідного файлу, включаючи текст, зображення, аудіо, 3D-карти і багато іншого, об’єднану в одному файлі;

- забезпечують зручний пошук, включаючи пошук по сканованому тексту, який був перетворений за допомогою технології оптичного розпізнавання Символів (OCR);
- працюють з допоміжними технологіями, які роблять файли PDF доступнішими для людей з обмеженими можливостями, такими як слабкий зір або сліпота.

Детальніше про формат: https://ru.wikipedia.org/wiki/Portable_Document_Format[35]
та https://acrobat.adobe.com/ru/ru/products/about-adobe-pdf.html[36]

Формати представлення даних через API

JSON (JavaScript Object Notation)

JSON (від англ. *JavaScript Object Notation*) – текстовий відкритий формат, оснований на Javascript представлені та призначений для обміну даними в Інтернет між сервером та клієнтом або сервером і сервером. Хоча він позиціонується, як незалежний від системи і мови програмування, частіше за все використовується за допомогою програм на Javascript, але як і інші текстові формати, легко читається людиною.

Найбільшу популярність JSON набув після створення інтерактивних веб-сторінок, дані до яких через API передавались під час взаємодії користувача з елементами інтерфейсу (т.з. технологія AJAX). За рахунок своєї лаконічності, на відміну від XML, простоті й швидкості використання саме в програмах на Javascript, широкими можливостями в серіалізації даних – рекурсивного перетворення в текстовий вигляд складних об'єктів, формат активно використовується для формування «на льоту» та передачі структур даних в Інтернет в різних інформаційних системах і сервісах.

Синтаксис JSON (див. https://ru.wikipedia.org/wiki/JSON[37]) представляє собою текст у вигляді програмного коду Javascript, який описує кілька варіантів структур:

1) масив в квадратних дужках (значення через кому),
2) іменований масив (чи об'єкт) у фігурних дужках (пари ключ-значення, розділені двома крапками, через кому),
3) просте значення (число, рядок, true, false, null),
4) функцію.

Для перетворення формату JSON в об'єкт в пам'яті (десеріалізації) в Javascript використовується функція eval(), яка виконує завантажений в текстовий рядок програмний код.

Якщо для обробки використовувати тільки мову Javascript, виникають як мінімум дві проблеми. По-перше, щоб обробити дані, потрібно мати приблизно в два рази більше вільної пам'яті, ніж файл (спочатку завантажується повний текст JSON, а після його виконання, дані об'єкту завантажуються повторно у вигляді відповідних структур). По-друге, можливість

[35] https://ru.wikipedia.org/wiki/Portable_Document_Format
[36] https://acrobat.adobe.com/ru/ru/products/about-adobe-pdf.html
[37] https://ru.wikipedia.org/wiki/JSON

передачі програмного коду, який автоматично стартує, дозволяє зловмисникам отримати несанкціонований доступ до системи. Цю вразливість також можна уникнути, якщо спочатку перевірити і очистити текст JSON від заборонених об'єктів, а дані перетворити на коректні значення.

Через нераціональне використання пам'яті і ресурсів в JSON краще працювати зі об'єктами будь-якою складністю, але не дуже великих розмірів, в швидкому режимі «запит-відповідь».

Приклад формату JSON:

```
{
   "firstName": "Иван",
   "lastName": "Иванов",
   "address": {
       "streetAddress": "Московское ш., 101, кв.101",
       "city": "Ленинград",
       "postalCode": 101101
   },
   "phoneNumbers": [
       "812 123-1234",
       "916 123-4567"]
}
```

XML (eXtensible Markup Language)

XML (від англ. *eXtensible Markup Language* – мова розмітки, що розширюється) – мабуть найстаріший текстовий відкритий формат, створений в 1994 році та рекомендований Консорціумом Всесвітньої павутини (W3C), як основний для обміну інформацією в Інтернет. Гіпертекстова розмітка (HTML) – це один з різновидів XML. Разом з таблицями каскадних стилів CSS, які формують зовнішній вигляд документів, вони є тими основними форматами, що обумовлюють розвиток технологій.

Насправді XML (див. https://ru.wikipedia.org/wiki/XML[38]) – універсальний зручний для людини формат для збереження, передачі структурованих даних і їх автоматичної обробки у формі машиночитних документів. З нього починали розвиток інші відкриті формати, на ньому вдосконалювались підходи обміну даними. Сама ж мова розмітки XML є похідною від ще старішого та більш складного формату SGML (Standard Generalized Markup Language – стандартна узагальнена мова розмітки), стандартизованого за ISO 8879:1986 Information processing—Text and office systems—Standard Generalized Markup Language (SGML).

Перевагами XML є простота та гнучкість розмітки, яка не вимагає формальних, фіксованих назв тегів чи параметрів, і будь-який розробник може доповнювати та змінювати формат, створювати власну схему XSD (XML Schema Definition). Фактично, це мова, яка описує

[38] https://ru.wikipedia.org/wiki/XML

сама себе і будь-які за розміром і складністю структури даних. Безумовно, в цей формат можна конвертувати інші формати (наприклад, XLS). Крім того, формат не залежіть від операційної системи чи бази даних. Для простого перегляду чи редагування достатньо текстового редактора. Але є й недоліки – дані в форматі XML займають значно більше місце, ніж це потрібно, за рахунок повторення тегів та відступів, а парсинг значно складніший, ніж CSV чи JSON.

За довгий час існування XML на його базі було розроблено багато форматів і стандартів зі схожим синтаксисом (див. приклади стандартних схем на http://schema.org[39]). Зазвичай цю групу форматів називають загальною назвою – XML, тому що вони мають єдині механізми опису схем XSD, перевірки правильності даних (валідації), перегляд DOM (Document Object Model), доступу до елементів XPath та трансформації для автоматичного конвертування у інші схеми чи формати (наприклад, альтернативні JSON та YAML) за допомогою мови перетворення XSLT (*eXtensible Stylesheet Language Transformations*).

Використання формату XML (а саме LegalXML) у якості відкритого стандарту нормативно-правового документа – це сучасний спосіб забезпечити обмін інформацією (документами, картками, довідниками тощо) між інформаційними системами або в межах однієї системи при підготовці (опрацюванні) документів.

В одному файлі XML в текстовому вигляді, крім основних даних та тексту електронного нормативного документа, можна розміщати метадані (характеристики, реквізити, опис, класифікацію тощо), вкладені файли (картинки, стилі тексту, таблиці Excel, документи Word і т.і.), необхідні структури чи довідники. Це дозволяє зручно не тільки зберігати, передавати, обробляти документ, отримувати PDF версію для друку, формувати зміст чи робити посилання на конкретну главу, статтю, пункт, підпункт тощо, а й автоматизовано вносити зміни, підготовлені у вигляді, що дозволяє їх програмну обробку.

Приклад даних у форматі XML:

```
<person firstName="Иван" lastName="Иванов">
        <address streetAddress="Московское ш., 101, кв.101"
                city="Ленинград" postalCode="101101"/>
        <phoneNumbers>
                <phoneNumber>812 123-1234</phoneNumber>
                <phoneNumber>916 123-4567</phoneNumber>
        </phoneNumbers>
</person>
```

RDF (Resource Description Framework)

RDF - це розроблена консорціумом Всесвітньої павутини модель для представлення даних, особливо - метаданих. RDF представляє твердження про ресурсах у вигляді, придатному для машинної обробки. RDF є частиною концепції семантичної павутини.

[39] http://schema.org

Ресурсом в RDF може бути будь-яка сутність - як інформаційна (наприклад, веб-сайт або зображення), так і неінформаційна (наприклад, людина, місто або якесь абстрактне поняття). Твердження, що висловлюється про ресурс, має вигляд «суб'єкт - предикат - об'єкт» і називається триплетом. Затвердження «небо блакитного кольору» в RDF-термінології можна представити таким чином: суб'єкт - «небо», предикат - «має колір», об'єкт - «блакитний». Для позначення суб'єктів, відносин і об'єктів в RDF використовуються URI.

RDF сам по собі є не форматом файлу, а тільки лише абстрактною моделлю даних, тобто описує пропоновану структуру, способи обробки та інтерпретації даних. Для зберігання і передачі інформації, покладеної в модель RDF, існує цілий ряд форматів запису.

Для обробки RDF-даних пропонується реалізувати мови запитів: SPARQL (стандарт W3C), RQL, RDQL.

Для запису і передачі RDF використовується кілька форматів, в тому числі:

- RDF / XML - запис у вигляді XML-документа;
- RDF / JSON - запис у вигляді JSON-даних;
- RDFa - запис всередині атрибутів довільного HTML- або XHTML-документа;
- N-Triples, Turtle, N3 - компактні форми запису тверджень.

Формати даних для роботи з геопросторовими даними

Формати GIS

Файли геопросторових даних можуть містити таку інформацію:

- Просторова або геометрична інформація, яка забезпечує положення і форми конкретних географічних об'єктів.
- Інформація про атрибути, яка забезпечує описову інформацію (числову, текстову, логічну) про кожний об'єкт.
- Інформація про відображення, яке може описувати спосіб оформлення об'єкту.

Деякі цифрові карти не містять всі три типи інформації. Наприклад, растрові карти, як правило, не включають в себе інформацію про атрибути. В свою чергу багато джерел векторних даних не включають інформацію про оформлення.

Просторова або геометрична інформація

Географічна інформація описує у цифровій формі положення та форму кожного зображуваного об'єкта. На векторній карті, властива об'єкту позиція зазвичай виражається у вигляді наборів координат x, y або x, y, z з використанням системи координат, визначеній для набору геоданих. Більшість векторних географічних інформаційних систем підтримують три основних геометричні об'єкти: * Точка: одна пара координат, 0-вимірний об'єкт. *

Лінія: дві або більше точок у певній послідовності, одновимірний об'єкт. * Полігон: площа, обмежена лінією, двовимірний об'єкт. * Тіло, об'єм обмежений гранями, тривимірний об'єкт.

Векторні дані характеризуються топологією, яка задає правила просторової взаємодії між векторними об'єктами. В залежності від масштабу даних один і той же об'єкт може зберігатися у вигляді точки, лінії або полігону.

На растровій карті властива об'єкту позиція визначається комірками з відповідними атрибутами. Кожна комірка характеризується розміром та значенням атрибутів, яке може бути відображене у вигляді значень кольорової моделі (супутниковий знімок), або значень атрибуту (GRID), на основі яких растр може бути класифікований за кількісними або якісними характеристиками.

Деякі системи також підтримують більш складні об'єкти, такі як регіони, кола, еліпси, дуги і криві.

Інформація про атрибути

Атрибути описують специфічні характеристики об'єктів, які не є графічними елементами. Для прикладу, атрибутом, що характеризує дорогу може бути її назва або дата, коли її побудували. Зазвичай, атрибути зберігаються окремо від графічної частини карти за допомогою геореляційних структур даних. Атрибути є частиною опису векторних карт і рідко застосовуються разом із растровими зображеннями.

Інформація про відображення

Інформацію, що відображається може описувати цифрова карта або накресена (план). Загальна інформація на карті водображається за допомогою різних кольорів: озера, моря, вершини, гори, рівнин; ширини і типів ліній: назви доріг, вулиць, нанесення висотної поясності на планах тощо.

Інформація про відображення задає графічні стилі, за допомогою яких відображаються дані на карті. Основними графічними стилями є графічні перемінні, до яких відноситься колір, товщина, форма, орієнтація, символізація, підписи. Разом графічні перемінні утворюють способи умовних позначень на карті географічних об'єктів та формують легенду карти.

GIS-формати створені для стандартного кодування географічної інформації в файл. Файли цього формату створюються державними агентствами з картографування, комерційними організаціями або розробниками програмного забезпечення.

Ці формати є розповсюдженими, але існує програмне забезпечення, що не підтримує файли цього формату. В таких випадках шукають способи перетворення ГІС-файлу або використовують інший файл. Майже кожен ГІС має свій власний внутрішній формат файлу. Ці формати призначені для оптимального використання розробником і часто є запатентованими. Вони не призначені для використання поза системою, що їх розробляла.

Векторні формати

Багато додатків ГІС засновані на векторній технології через її широке застосування. Вони є складними, тому що інформацію можливо описати безліччю способів: для зберігання

координат, атрибутів, зв'язків, структур баз даних та відображення інформації. Найбільш поширеним форматом гепросторових даних є формат Shapefile[40] (shp). Цей формат розроблений компанією ESRI та є пропрієтарним форматом даних. Проте сьогодні він став дефакто стандартом обміну та публікації даних. Більшість програмних продуктів працюють з цим форматом, в тому числі ПЗ QGIS, яке є найбільш популярним OpenSource додатком для ГІС. Специфікація формату визначає набір файлів із загальним префіксом імені файлу, які зберігаються в тому ж каталозі. Три обов'язкових файли мають розширення файлів shp, .SHX і .dbf.

> .shp - головний файл; містить набори геометричних об'єктів
> .shx - індексний файл; який икористовується для зв'язку між файлами .dbf і .shp
> .dbf - атрибутивний файл; містить атрибути об'єктів, описаних в .shp файлі у форматі dBase IV.

На сучасному етапі більшість геоданих зберігаються в СУБД, які підтримують збереження просторових даних. Зараз до них відносяться Oracle, SQLserver, Postgres та ін. ПЗ ArcGis підтримує два типи баз даних: персональна та геобаза даних з відповідними форматами: mdb, gdt.

Значного поширення набули формати ГІС: MapInfo TAB format[41], Spatialite[42], Geography Markup Language (GML)[43]. Також розповсюдженими в Україні є формати даних ПЗ Панорама – SXF, Digital – підтримка формату in4, який довгий час був обмінним файлом кадастрової інформації.

Файли AutoCAD не є ГІС форматом, проте сучасні професійні ГІС підтримують пряму конвертацію з формату DXF та навпаки. Доволі рідкими форматами в Україні є формати National Transfer Format[44] (NTF), SOSI[45], Spatial Data File[46], TIGER[47], Vector Product Format[48] (VPF).

Для обміну просторовими даними також використовують формат ASCII, який у більшості випадків використовують для збереження растрових даних.

Для прикладу наведено найпоширеніші формати: AutoCAD DXF[49], Cartesian coordinate system[50] (XYZ), Digital Line Graph[51] (DLG), Esri TIN[52], Geography Markup Language[53] (GML),

[40] https://en.wikipedia.org/wiki/Shapefile
[41] https://en.wikipedia.org/wiki/MapInfo_TAB_format
[42] https://en.wikipedia.org/wiki/Spatialite
[43] https://en.wikipedia.org/wiki/Geography_Markup_Language
[44] https://en.wikipedia.org/wiki/National_Transfer_Format
[45] https://en.wikipedia.org/wiki/SOSI
[46] https://en.wikipedia.org/wiki/Spatial_Data_File
[47] https://en.wikipedia.org/wiki/TIGER
[48] https://en.wikipedia.org/wiki/Vector_Product_Format
[49] https://en.wikipedia.org/wiki/AutoCAD_DXF
[50] https://en.wikipedia.org/wiki/Cartesian_coordinate_system
[51] https://en.wikipedia.org/wiki/Digital_line_graph
[52] https://en.wikipedia.org/wiki/Esri_TIN
[53] https://en.wikipedia.org/wiki/Geography_Markup_Language

GeoJSON[54], GeoMedia[55], ISFC, Keyhole Markup Language[56] (KML), MapInfo TAB format[57], National Transfer Format[58] (NTF), Spatialite[59], Shapefile[60], Simple Features[61], SOSI[62], Spatial Data File[63], TIGER[64], Vector Product Format[65] (VPF).

Растрові формати

Растрові файли зазвичай використовуються для зберігання інформації зображення у вигляді сканованих паперових карт чи аерофотозйомки. Вони також використовуються для передачі даних дистанційного зондування Землі. На відміну від інших растрових файлів, які характеризуються розрізненням у розмірах точок на дюйм, розрізнення на зображеннях дистанційного зондування виражається в метрах на піксел растру, що вказує розмір земельної ділянки певної області. Прикладами растрових форматів є: ADRG, RPF, CADRG, CIB[66], Digital raster graphic[67] (DRG), ECRG, ECW[68], Esri grid[69], GeoTIFF[70], IMG – ERDAS IMAGINE[71], JPEG2000[72], MrSID[73], netCDF[74]-CF.

Детальніше про формати та GIS-формат: https://en.wikipedia.org/wiki/GIS_file_formats[75]

KML - геоінформаційний формат Google Maps

KML (від англ. *Keyhole Markup Language*) - мова розмітки на основі XML для представлення тривимірних геопросторових даних в програмі «Google Earth».

Підмножина мови KML 2.0 може використовуватися і для відображення двовимірних карт в сервісі «Карти Google».

KML-файли зазвичай поширюються в ZIP-архіві: KMZ

Зміст файлів KML визначає один або кілька об'єктів для відображення в Google Earth. Цими об'єктами можуть бути:

[54] https://en.wikipedia.org/wiki/GeoJSON
[55] https://en.wikipedia.org/wiki/GeoMedia
[56] https://en.wikipedia.org/wiki/Keyhole_Markup_Language
[57] https://en.wikipedia.org/wiki/MapInfo_TAB_format
[58] https://en.wikipedia.org/wiki/National_Transfer_Format
[59] https://en.wikipedia.org/wiki/Spatialite
[60] https://en.wikipedia.org/wiki/Shapefile
[61] https://en.wikipedia.org/wiki/Simple_Features
[62] https://en.wikipedia.org/wiki/SOSI
[63] https://en.wikipedia.org/wiki/Spatial_Data_File
[64] https://en.wikipedia.org/wiki/TIGER
[65] https://en.wikipedia.org/wiki/Vector_Product_Format
[66] https://en.wikipedia.org/wiki/Controlled_Image_Base
[67] https://en.wikipedia.org/wiki/Digital_raster_graphic
[68] https://en.wikipedia.org/wiki/ECW_(file_format)
[69] https://en.wikipedia.org/wiki/Esri_grid
[70] https://en.wikipedia.org/wiki/GeoTIFF
[71] https://en.wikipedia.org/wiki/ERDAS_IMAGINE
[72] https://en.wikipedia.org/wiki/JPEG2000
[73] https://en.wikipedia.org/wiki/MrSID
[74] https://en.wikipedia.org/wiki/NetCDF
[75] https://en.wikipedia.org/wiki/GIS_file_formats

- **Позначки на карті**. Задається знаковий символ, який відображається в певному місці на карті, і його параметри: величина, текст і колір, величина написів, яка відображається біля знаку. Є можливість створити позначку без знаку. Наприклад, поставити номери будинків району. Для цього потрібно просто в стилі створити порожнє посилання на значок:

<Icon> <href> </href> </Icon>

- **Багатокутник або набір ліній**. Визначається колір ліній і колір підпису.
- **Зображення**. Визначається положення зображення на поверхні Землі, а також його масштаб. Також можна розмістити зображення на екрані, що не переміщається разом з картою - наприклад, логотип.
- **Тривимірна модель**. Версія мови KML 2.1 (яка відповідає четвертій версії програми Google Earth) дозволяє підключати опис тривимірних об'єктів (наприклад, будівель і споруд). Тривимірну модель можна задати двома способами: завданням висоти плоских фігур (витягуванням) і посиланням на повноцінну модель в форматі Collada.

Крім згаданих характеристик, для кожного об'єкта також задаються основні геоінформаційні властивості (географічна широта і довгота, а також висота над рівнем моря або над рівнем поверхні Землі). Може бути наведено короткий опис об'єкта (який в подальшому відображається в підказці за запитом користувача). Може бути зазначений рекомендований ракурс спостереження зазначеного на карті місця, тобто рекомендована висота, азимут і кут нахилу «віртуальної камери».

Об'єкти всередині KML-файлу можна організувати в ієрархічні структури папок і підпапок, щоб було зручніше спільно ввімкнути або вимкнути відображення логічно взаємопов'язаних груп об'єктів.

KML-файл може містити (в URL-формі) посилання на інші файли в форматі KML або KMZ, розташовані де-небудь в мережі, і задавати умови і регулярність завантаження і відображення даних з цих зовнішніх джерел. Таке мережеве посилання також видно в якості своєрідної підпапки.

Мова KML багато в чому наслідує структуру мови GML - географічного мови розмітки.

Приклад KML-розмітки:

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.1">
<Placemark>
    <name>Київ</name>
    <description><![CDATA[<p>Київ, Київська область, Україна.</p>
        Столиця України. Місто побудоване на березі річки Дніпро.]]>
    </description>
    <LookAt id="khLookAt540_copy0">
            <longitude>38.0576198113139</longitude>
            <latitude>44.56963150481845</latitude>
            <altitude>0</altitude>
            <range>14693.40972993507</range>
            <tilt>49.10268313434742</tilt>
            <heading>37.85562764777833</heading>
    </LookAt>
    <Style>
        <IconStyle>
            <scale>0.9</scale>
                <Icon>
                <href>root://icons/palette-4.png</href>
                <x>32</x>
                <y>128</y>
                <w>32</w>
                <h>32</h>
                </Icon>
        </IconStyle>
        <LabelStyle>
             <scale>0.9</scale>
        </LabelStyle>
    </Style>
        <Point id="khPoint541_copy0">
                <coordinates>
                        38.06284424434902,44.56842733252498,0
                </coordinates>
        </Point>
</Placemark>
</kml>
```

GeoJSON та TopoJSON

GeoJSON - відкритий формат, призначений для зберігання графічних структур даних, заснований на форматі JSON.

Формат може зберігати примітивні типи для опису географічних об'єктів, такі як: точки (адреси та місця розташування), лінії (вулиці, шосе, кордони), полігони (країни, штати, ділянки землі). Також можуть зберігатися так звані мультітипи, які представляють собою об'єднання декількох примітивних типів.

Формат GeoJSON відрізняється від інших стандартів ГІС тим, що він був написаний і підтримується не певною організацією зі стандартизації, а за допомогою робочої групи розробників.

Подальшим розвитком GeoJSON є TopoJSON, розширення GeoJSON, яке кодує геопросторову топологію, і, як правило, забезпечує менший розмір файлів. Сьогодні ці формати підтримуються більшістю ГІС, а також безпосередньо застосовуються при публікації інтерактивних картографічних даних в веб, зокрема засобами бібліотек OpenLayers та Leaflet.

```
{ "type": "FeatureCollection",
        "features": [
          { "type": "Feature",
            "geometry": {"type": "Point", "coordinates": [102.0, 0.5]},
            "properties": {"prop0": "value0"}
            },
          { "type": "Feature",
            "geometry": {
              "type": "LineString",
              "coordinates": [
                [102.0, 0.0], [103.0, 1.0], [104.0, 0.0], [105.0, 1.0]
                ]
              },
            "properties": {
              "prop0": "value0",
              "prop1": 0.0
              }
            },
          { "type": "Feature",
             "geometry": {
               "type": "Polygon",
               "coordinates": [
                 [ [100.0, 0.0], [101.0, 0.0], [101.0, 1.0],
                   [100.0, 1.0], [100.0, 0.0] ]
                 ]
             },
             "properties": {
               "prop0": "value0",
               "prop1": {"this": "that"}
               }
```

```
31                    }
32                ]
33            }
```