Понятието „машиночитаемост” и концепцията open data

Юри Тодоров

С въвеждането на Закона за електронното управление у нас в практическо действие държавните институции се задължават да разработват и внедряват стратегии, с които да се разширява и улеснява достъпа до информацията, която те създават, събират и съхраняват. Лесен и често прилаган способ е публикуването на сведения в мрежата, което действие от само по себе си демонстрира неразбиране и пренебрегване на значението за достъпния характер на информацията. В този смисъл понятието „**машиночитаемост**” придобива ново, актуално значение. В смисъла на нормативните разпоредби е да направим тази информация четима не само за човека, но и за машините, т.е. компютрите. Степента на която информацията е достъпна за обработване от компютър способства за решаване на приоритетите, зададени в електронното управление в смисъла на отвореността (open government, open data[1]). Много често тази степен на откритост има пряко ограничаващо въздействие за гражданите и други заинтересувани страни върху начините на употреба на предоставената информация.

Усилията за предоставяне на сведения относно управлението на страната в исторически план се съсредоточават в публикуването на статична информация по различни теми и области на обслужване на гражданите в смисъла на електронното правителство. Предполагаемият потребител тук е човек, който чете, разпечатва, попълва формуляри. Същински достъп до информационният ресурс, с който работи институцията, до регистри, реални електронно създадени документи и дигитални архиви, почти не се предлага. Основно намерение на стратегията е дигитализацията и използването на цифрови формати [2]

[1] Отворено управление, отворени данни и формати – платформено независим начин за представяне на информация, който е машинно четим, предоставя се на широката общественост без ограничения и който благоприятства употребата на тази информация.

[2] Цифровият формат е стандарт, по който информацията може да се съхранява и възпроизвежда от компютър.

В приложен аспект тези практики се налагат в голяма степен чрез господстващите технологии, които са базирани на хипертекстове (**HTML**) и факсимилета от документи (**PDF**). Самите технологии ограничават използваемостта на предлаганата чрез тях информация. Машиночитаемите формати[3] за представяне на информационния ресурс включват възможността за реален достъп до този ресурс, и те отдавна се прилагат в системите за управление на съдържанието, например. Най-популярната от тези технологии е може би **XML** (eXtensible Markup Language[4] - език, който задава правилата за кодиране на документи по начин, който е достъпен както за човека така и за компютъра, т.е. машиночитаем формат). Други технологии с широки възможности са **RDF** (Resource Description Framework – език за представяне на данни и информация в тяхната свързаност, RDF е отворен стандарт) и **JSON** (JavaScript Object Notation [5] – машиночитаем формат за представяне на елементарни структури от данни и асоциативни полета, също отворен формат по RFC 4627 [6]).

Става ясно, че понятието за „машиночитаемост" се различава по значение и обхват от понятието „цифрово достъпен" или възпроизводим с цифрови технологии, по отношение на работата с документи в по-широк смисъл. Процесът на преобразуване в цифров вид на информация представена като текст, графика, таблица и т.н. я прави достъпна за цифрово възпроизвеждане, но не и за обработване от компютъра най-вече в нейната субстанция и същност, а не само по отношение на нейната външност и форма [7].

Какви фактори определят една технология като налагаща се? Най-напред това е относителната **постоянност и устойчивост**, обстоятелства които правят възможно широкото разпространение и възприемане на техноло-

---

[3] Всички машиночитаеми формати са и цифрови формати, но не всички цифрови формати са четими от компютър.
[4] Разширим маркиращ език.
[5] Скриптов език за описание на обекти.
[6] подробно описание на: http://www.ietf.org/rfc/rfc4627.txt
[7] Елементарен пример за това е страницата на едно списание, която след сканиране се представя като дигитализирано копие. Само по себе си то е графично съответствие на външността, за съдържанието на списанието можем да съдим по щрих кода, който е представен в действителен машиночитаем формат.

гията, основаваща се на един повсеместно приет модел. От друга страна, особено важно е **лесното използване и употреба**, без да се правят промени в съдържанието на обекта. Технологията трябва да е приложима навсякъде и да е **често срещана**. Приложението на такава технология трябва да е оправдано от икономическа гледна точка и дори да помага за повишаване на **ефективността на работа**. Особено важно качество е **разширяемостта** и възможността за развитие както и за приложение в разнообразни области.

Подобни фактори излизат на преден план, когато става дума за машиночитаемостта на данните за електронното правителство, откритото управление и отворените архиви. Очакванията по отношение на лесната употреба, ефективност и разширяемост трябва да създадат условия за внедряване на технологиите. Очевидно, необходими са усилия в насока на устойчивостта на форматите и тяхната стандартизация като в същото време се очаква лесно приложение и внедряване. Икономическата ефективност от приложението на машиночитаемия формат за данни трябва също да е доказана. Когато всички посочени условия са изпълнени, машиночитаемата информация ще измести сегашните дигитални, електронни формати, предлагайки по-широки възможности.

С нарастване на задълженията по отношение на отвореността в управлението усилията в насока на разполагаемостта, достъпността до информацията следва да се концентрират в насока на създаване на машиночитаеми версии на информационни обекти, а не само на статични копия за информационни фрагменти под формата на документи в обичайните формати (pdf, html, jpeg). Една таблица, показана във формата pdf, например, може да бъде разчетена и съдържанието и да стане визуално достъпно, но не е възможно да се вникне в данните, които са послужили за създаването на таблицата. Тази информация остава скрита и недостъпна, в противовес на принципите за откритост. Усилията по отношение на отварянето на информационните ресурси трябва да гарантират на потребителите получаването на достъп до фундаменталните сведения, а не само да крайни резултати в статичен вид. В този смисъл машиночитаемите формати помагат на управлението да преодо-

лее пропастта между „документите”, неподвижни и замръзнали в техните статични формати и динамичните данни, които да са достъпни за обработка.

Машиночитаемост на документи

Обичайният формат за цифрово преобразувани документи е **PDF**. Той предлага редица предимства при публикуване на документите. Например, достъпност на документа в интернет, предоставяне на копия за използване. Този начин на публикуване не дава отговор на редица въпроси, относно реквизитите на документите. Едно средство за преодоляване на този недостатък са наборите от **метаданни**, които изглежда да са панацеята на електронното управление. В метаданните са включени допълнителните сведения, които позволяват проследяването на автора, редакциите върху документа и елементите в еволюцията на неговото съдържание. Без съпровождащи метаданни документите не са достъпни за информационни запитвания. Без съмнение, представянето на документи в pdf формат представлява крачка напред в посока на машинната четимост, когато е съпроводен с метаданни.

Форматът **XML** е разработен за различни цели, между които включване на възможността за непосредствено представяне на метаданните към един документ. Практическото приложение на този формат в оптимален вариант води до автоматично извличане на информацията от документи, която да може да се търси, анализира и обработва непосредствено. Много важно свойство на формата е предлагането на механизми за проследяване на версиите на един документ и еволюцията при неговото създаване. Така историята на документа става видна за други потребители. Това е качество, което е в смисъла на откритостта, проследяемостта и прозрачността при взимане на решения.

Получаването на релевантни метаданни е особено важно за информационните системи, когато при поставено запитване трябва да се намери определена информация. Форматът xml сам по себе си не произвежда метаданни, но представлява добро средство за предоставянето им за употреба, когато такива са на лице. Генерирането на метаданни обаче остава задължение на собствениците и авторите на документи.

Машинна четимост на данни

Машиночитаемостта оказва непосредствено влияние върху използваемостта на информацията. Огромните бази от данни сами по себе си не са използваеми за потлебителите, докато не се обработи по някакъв начин – чрез анализиране, онагледяване или обобщение. За да се разкрие напълно потенциалът на отвореното управление, държавното управление трябва да предоставя информация във формат, който позволява тя да се обработва. Предлагането на информация от такъв тип ще я освободи за проучване на начините за нейното създаване, за намиране на отговорите на различни проблеми. Така информацията става наистина производствена сила и предлага потенциала за генериране на нещо ново.

Най-елементарният машиночитаем формат е **CSV** (comma separated variables[8]) и той се поддържа от всички продукти, работещи с бази данни и таблици. Това е текстов формат, който обединява таблични данни в обикновен текст, който е лесно достъпен за компютъра. Този формат по своята същност не произвежда метаданни, но има много стандарти за описание на метаданни, които са съвместими с csv. Основният проблем тук е откриването на общите елементи между различни набори от данни. Така, ако в един набор от данни срещаме името „Бургас“, а в друг – „Бс“, трябва отделно да се генерира информация за връзката между двете. Това обстоятелство се има предвид с понятието „свързани данни“. Новият формат **RDF** (resource description framework [9]) е създаден с намерението за непосредствено обвързване на набори от данни с техните описания и за свързване на общи термини и понятия между различни набори от данни. С това форматът става все по-интересен за съвременни приложение в смисъла на отвореното управление и архиви.

Съвременното схващане за отвореност (open government) залага на обединяването на съдържания (**content syndication**). Това са документи и метаданни, които позволяват непрекъснато автоматично проследяване на ин-

[8] Променливи, отделени със запетая.
[9] Схема за описание на ресурси.

формацията. Така със средствата на информационната система всички свързани данни се сглобяват на търсещия компютър. Общоприетите формати за синдикация[10] сега се използват за общо ползване на машиночитаеми документи, но те самите не са решения на проблема за четимостта. Те са едно временно решение, което днес е познато и се използва по-често.

Огромната лавина от цифрово представена информация поставя класическите архиви и традиционните фондообразуватели пред нови предизвикателства [11]. Те рефлектират от промените в начина на работа на обществените правителствени и неправителствени организации, но също така и на тези от частния сектор. Организациите в днешно време са принудени да взимат мерки за запазването и достъпността на информационните ресурси, които произвеждат, защото цифровата информация може да се загуби или разруши лесно. Технологичният прогрес от своя страна налага непрекъсната и високоскоростна еволюция на техническите системи в тяхната цялост. Промените се отразяват непосредствено върху формати и стандарти, които са задължени да се усъвършенстват в кореспондиращи темпове. Затова трайното запазване, съхранение и използване на основния информационен ресурс винаги се нуждае от допълваща информация, която е жизнено важна за възпроизвеждането и достъпа до оригинала. Дейностите по дългосрочното съхранение на документи са задължение не само на архивите, но и на фондообразувателите и преследват целия жизнен цикъл на документите.

[10] Това са форматите RSS, Atom и JSON. RSS – really simple syndication (формат за публикуване на информация, която се актуализира много често); Atom е формат, основаващ се на XML, използва се при публикуване на новини и обвързва документа с метаданните.

[11] В отговор на тези предизвикателства е разработен Референтния модел за отворена информационна система на архивите (OAIS – Open Archival Information System). ISO 14721:2012
Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57284