Напътствия за идентифициране на отворени данни в институциите

Кратко ръководсво

1 Терминология

Отворени данни е термин, който комбинира няколко характеристики – публичност, структура и надеждност. Всяка от тях е ключова за това информацията, публикувана от институции или частни организации, да бъде достъпна, подходяща за автоматичен анализ и свободна за разпространение.

Публичността се изразява в това данните да бъдат на разположение или на сайта на ведомството, или в централизиран портал. Съпътстващият лиценз трябва да е такъв, че да не ограничава обработката и разпространението на анализите базирани на тези данни.

Структурираността на данните е изключително важна. Те трябва да са представени в единен отворен формат (XML, CSV, RDF, SQL и прочие), който да може да се обработва автоматично от анализиращ софтуер. Надеждността на данните означава, че в тях не трябва да има структурни или логически грешки и те да се обновяват редовно или в реално време.

2 Стъпки в идентифицирането на данни

Първите стъпки при идентифициране на данни подходящи за "отваряне" в една държавна институция са отговорите на следните три въпроса:

1. *Какви бази данни и документи се поддържат?*
2. *Кои са обществено достъпите данни сред тях?*
3. *Как може да се изкарат данните в структуриран (табличен или дървовиден) вид от тези бази данни и документи?*

2.1 Анализ

Първата задача е да се опишат структурите от данни, които са налични. Това може да бъдат релационни или XML бази данни, списъци в текстови файлове, електронни таблици или документи в свободен текст. Важно е да се определи до колко тези данни са актуални, кой отговаря за тях и кои информационни системи имат достъп до тях (*с кой софтуер се отварят и обработват*). Пример за това са бази данни с информация за фирми, стенограми от заседания или бюджетни справки.

2.2 Публична спрямо защитена информация

От споменатият опис може да се направят изводи за това кои данни са обществено достъпни. При наличието на лични данни или служебна информация в някои масиви от данни, те могат да бъдат заличавани и така данните да станат безопасни за публикуване, като запазят аналитичната си стойност. Това може да стане по два начин - заличаване на колони от таблиците с цел премахване на чувствителна информация или обобщение и категоризиране на записите. При последният метод се групират данните по различни категории - възраст,

време, местоположение, тип документ и прочие. При избиране на правилна детайлност на това групиране, е възможно данните да бъдат анонимни като запазят полезността си. Пример за това може да са данни за дължимите суми на граждани към НАП групирани по населено място и месец.

2.3 Структура на данните

След като са идентифицирани данните подлежащи на отваряне, трябва да се определи структурата им. Отворените данни не са само публично достъпни, но и такива, които могат да бъдат автоматично обработванеи. Има много подходящи формати за тази цел и е важно да се избере такъв, който ще е най-близко до оригиалът, за да не се загубва информация. Като правило е добре свързаните данни да се публикуват в XML и SQL формат. При таблиците често CSV е най-подходящ.

Друг важен момент е как по принцип се отварят и обработват въпросните данни. По-новите системи, които използват отворени стандарти, често имат възможност за автоматично извеждане направо в удобен формат. В други случаи обаче, данните трябва да се изкарват директно от базата данни. В тези случаи е важно да се разчита на документация, която обяснява структурата, в която се пази информацията, за да се представи правилно в отворения си формат. В тези случаи е важно да не се загуби логиката на данните (*примерно коя стойност какво означава*), защото има риск резултата да съдържа грешки. Най-накрая, има системи, които са затворени като код и данни и които привидно е невъзможно да се свържат с други системи. Тези трябва да документирани като такива, защото ще представляват пречка в интеграцията в рамките на електронното управление.

2.4 Автоматизация

Последната фаза на "отварянето" на данните е автоматизирането. Когато открием как да изкараме данните от съответните бази данни и системи за обработка на документи, трябва това да се прави периодично и автоматично. Така ще се гарантира надеждността и ще се намали натоварването върху администрацията. Това може да стане чрез прости програми с ограничен достъп за четене до определени ресурси в системата. Периодично те могат да изкарват данните в определения отворен формат и да ги публикуват на порталът за отворени данни на правителството. При някои системи това може да става и в реално време чрез използване на информационни услуги.

3 Документация

Резултатът предварителният анализ ще предлага данни, които са подходящи за отваряне. Нужно е да се опишат следните характеристики:

- Какви са данните и какво описват?
- Къде се намира базата данни или системата за обработка на документи, кой я поддържа и използва?
- Има ли чувствителна информация в данните и възможно ли е тя да се премахне?
- Известна и документирана ли е структурата на данните?
- Възможно ли е автоматично извеждане на данните от въпросните системи и/или трансформация в отворен формат?