ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФИЛОСОФСКИ ФАКУЛТЕТ
КНИГА БИБЛИОТЕЧНО-ИНФОРМАЦИОННИ НАУКИ
Том 7, 2015

ANNUAIRE DE L'UNIVERSITE DE SOFIA „ST. KLIMENT OHRIDSKI“
FACULTE DE PHILOSOPHIE
LIVRE DES SCIENCES DE L'INFORMATION ET DES BIBLIOTHEQUES
Tome 7, 2015

БИБЛИОГРАФСКИ ДАННИ В СЕМАНТИЧНИЯ УЕБ

МИЛЕНА МИЛАНОВА

Milena Milanova. BIBLIOGRAPHIC DATA ON THE SEMANTIC WEB

Bibliographic data produced by different cultural organizations are essential part of the web. Many of these data aren't available for indexing and retrieval by searching engines. The new approach of Link Data on the web is a possibility to change this state. Haw this is possible and what are the ways to put the bibliographic data on the semantic web are the main questions which answer try to find this paper. Bulgarian libraries must be part of the new processes in world of chancing data on the web.

През последното десетилетие все по-често говорим за семантичен уеб и то все по-често в контекста на данните, създавани от библиотеки, библиографски агенции, архиви, музеи, издателства. Провокацията е свързана с развитието на ИКТ технологиите по отношение на създаване на разбираемо за машините съдържание в уеб. През 2001 г. Tim Berners-Lee, James Hendler и Ora Lassila[1] представят идеята за свързването на данните в уеб пространството, по начин, който да позволи тяхното манипулиране от компютрите според съдържанието им. За да функционира семантичният уеб компютрите трябва да имат достъп до структурирани колекции от информация и набор от правила, които да позволяват „автоматизираното разсъждаване“.

[1] **Berners-Lee**, Tim, **Hendler**, James, **Lassila**, Ora. The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: *The Scientific American* [online] 17 May 2001 [Viewed on 15.07.2015]. Available from: http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html

Два са основните пътища, чрез които се реализира семантичния уеб: свързване на данните в него или това са свързаните данни (Linked Data); създаване и добавяне в уеб на установени предварително набори от данни в определен контекст. Постепенното осъществяване и развитие на идеята е видимо в облака на отворените свързани данни (Linked Open Data Cloud, LOD).

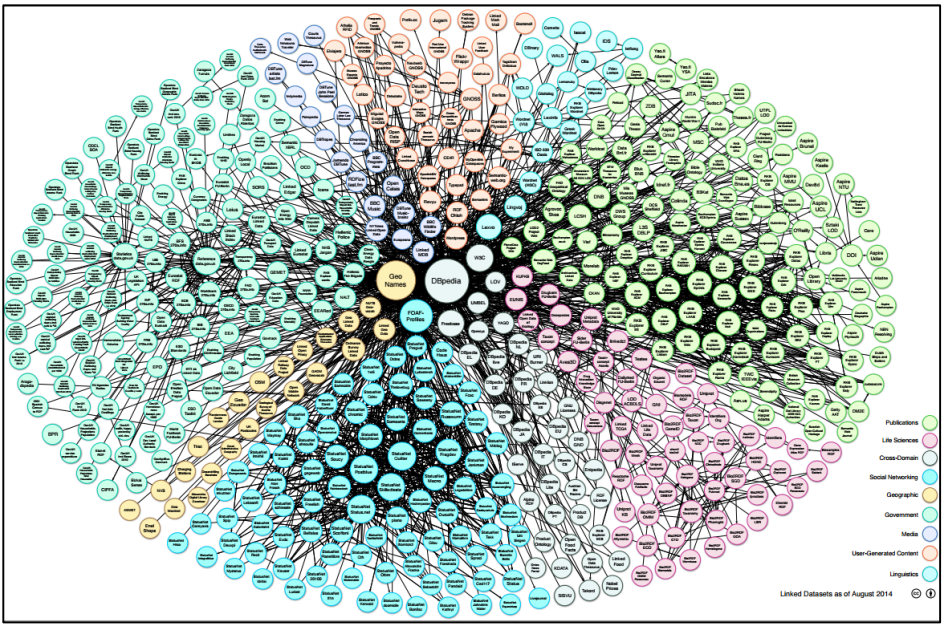

Схема 1. Диаграма на Облак на отворени свързани данни (LOD Cloud) към април 2014 г.[2]

Библиотеките са особено заинтересовани от тези процеси защото библиотечните каталози и данните в тях, така както ги познаваме, са високо структурирани, изработени са от специалисти и в този смисъл са подготвени да бъдат част от процеса по създаване на свързани данни, което ги поставя като основен участник в реализирането на семантичния уеб[3]. На прага сме, а и дори можем да твърдим, че вече се осъществява нова радикална промяна в света на библиотечните и библиографските данни, промяна сравнима с тази от 70-те години на миналото столетие,

[2] **The Linking Open Data cloud diagram** [online]. [Viewed 12.07.2015]. Available from: http://lod-cloud.net/

[3] **Coyle.** Karen. *Linked Data Tools. Connecting on the*. Chicago, ALA, 2012. Library Technology Report, Vol. 48, N 4. p. 10.

когато за първи път се реализира машинната обработка и създаването на автоматизираните библиотечни каталози. От този период насетне ИКТ технологиите все по-активно навлизат в различните библиотечни процесите и се превръщат във все по-важен инструмент за осъществяването на активни комуникационни връзки между света на библиографската информация и потребителите[4].

Технологиите, които позволяват да се реализира идеята на семантичния уеб са маркиращият език XML (eXtensible Markup Language) и структура за описание на ресурс RDF (Resource Description Framework)[5].

XML е маркиращ език, създаден, за да описва данните в уеб. Той е независим от специфичен хардуер или софтуер и позволява данните да бъдат управлявани. XML предлага гъвкава рамка, синтаксис за описване на документална структура и метаданни за съхраняване и предаване на документи[6]. Създаден е за да опрости съхранението и споделянето на данни. Описани със синтаксиса на XML всички данни са достъпни за ползване и интерпретиране от всякакъв вид устройства (компютри, гласови машини, мобилни устройства и др.). Синтаксисът е много прост и лесен за употреба. Изисква информацията да бъде заградена с етикети (например <catalog>), които трябва да бъдат отварящ и затварящ и се влагат един в друг, така че се получава йерархична организация на информацията. Езикът указва само как да бъдат използвани етикетите или определя единствено какви са данните, без да се посочва информация за тяхното съдържание или употреба. Създават се схеми XML (XML DTD, XML Schema) или допълнителни разширения, които разписват синтактични правила, определящи кои етикети и къде могат да бъдат прилагани и използвани, както и тяхното значение или семантика. Схемата, на която отговаря един XML документ се посочва в началото и следва определения синтаксис[7].

---

[4] **Guerrini,** Mauro**, Possemato,** Tizian. Linked data: a new alphabet for the semantic wed. In: *JLIS*, Vol. 4, N 1, (Gennanio/January) 2013, p. 77–78.

[5] **Харизанова,** Оля. Новите измерения на World Wide Web и предизвикателствата пред библиотеките. В: *Годишник на Софийския университет "Св. Климент Охридски"* [онлайн], Философски факултет, Книга Библиотечно-информационни науки, 2, 2010, с. 135–136. [Прегледан на 15.07.2015]. Достъпен от: https://research.uni-sofia.bg/bitstream/10506/316/1/OHarizanova_godishnik.pdf

[6] **Миланова,** Милена. *Българската каталогизация в глобалното информа-ционно пространство на XXI век.* [онлайн] *Анализи, стратегии, перспективи*. Дисертация за присъждане на образователна и научна степен „Доктор" / Научен ръководител доц. д-р Татяна Янакиева. София, 2008, с. 95. [Прегледан на 15.07.2015]. Достъпен от: https://research.uni-sofia.bg/handle/10506/1089

[7] **W3School.com** [online] http://www.w3schools.com/default.asp

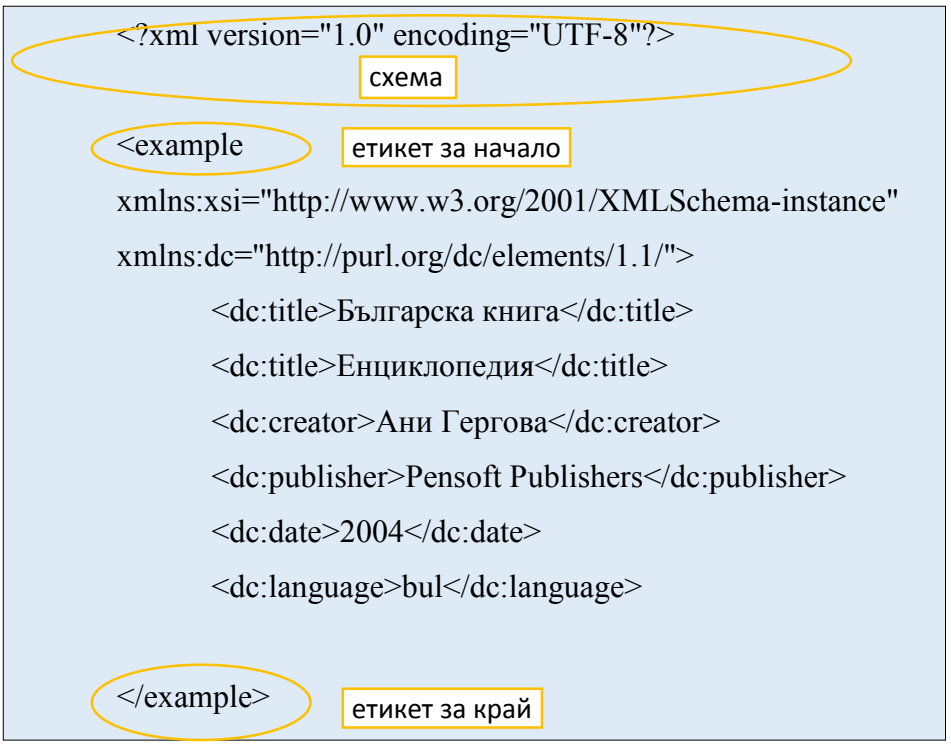

Схема 2. Пример за описание на документ с XML синтаксис.

Семантиката на един уеб документ е възможно да бъде представена чрез модела RDF. Това е среда, структура или стандарт за описание на ресурсите в уеб пространството. RDF характеризира семантично ресурсите и връзките между тях[8] или лексиката на RDF описва връзките. RDF е създаден от консорциума W3C и е разработен като модел от данни за представяне на метаданни, които се отнасят до уеб страниците и тяхното съдържание, използвайки синтаксиса на езици или формати (например, RDF/XML), правейки ги не само машинночетими, а също така и машиноразбираеми. Формално RDF е модел от данни, представящ ресурси, техните свойства и стойностите на тези свойства. RDF създава структура, която позволява свойствата на един ресурс да бъдат описани чрез посочване на друг ресурс[9].

[8] **Guerrini,** Mauro**, Possemato,** Tizian. Linked data: a new alphabet for the semantic wed. In: *JLIS*, Vol. 4, N 1, (Gennanio/January) 2013, p. 80

[9] **Миланова,** Милена. *Българската каталогизация в глобалното информационно пространство на XXI век* [онлайн]. *Анализи, стратегии, перспективи*. Дисертация за присъждане на образователна и научна степен „Доктор”. Научен

Синтаксисът на RDF изисква прилагане на концепцията на уникалните идентификатори (Uniform Resource Identifier URI*) на ресурсите в интернет, което позволява тяхното отличаване при използва-нето им в разнообразен контекст, така че той да може да бъде търсен, използван, свързван и изобщо манипулиран от различни системи[10]. RDF описва всеки един ресурс чрез неговия уникален идентификатор като посочва свойствата и стойността на тези свойства.

При съставянето на описанието е особено важно пространството на имената (namespace), което използва RDF, тъй като всеки елемент се предшества от етикет, асоцииращ го към определено пространство на имената, което не е създадено за целта на описанието, а използва наименования от различни стандарти, схеми с метаданни, модели. По този начин се постигат две цели: към името на елемента се присъединява съществуващата дефиниция; елементи от различни схеми могат да бъдат използвани заедно. RDF използва пространство на имената, осигурено от XML. Имената са дефинирани веднъж, когато се отнасят до URI, който осигурява името и свързва към началния идентификатор, използван, за да анотира всяко име в RDF спецификацията, дефинирайки оригинала на отделното име. Чрез осигуряването на кодиращ синтаксис и XML-пространство на имената, RDF прави метаданните недвусмислени. Постига се смесване на набори от елементи в дадено описание на метаданни без да има опасност от изпадане в противоречие при употребата на имената на елементите. Един елемент може да се основава на част от дадено пространство на имената (например Dublin Core), като няма опасност да започне да си противоречи с елемент със същото име от друго пространство на имената. RDF е в състояние да обхване семантично описателните стандарти за голям набор от елементи от метаданни, създадени от голям брой потребителски общности и да ги съвместява.

ръководител доц. д-р Татяна Янакиева. София, 2008, с. 107–108. [Прегледан на 15.07.2015]. Достъпен от: https://research.uni-sofia.bg/handle/10506/1089

* URI представлява поредица от символи, които позволяват да се идентифицира името на ресурс. Идентификатора, позволява взаимодействие с ресурс в уеб чрез използването на специфичен протокол. Най-популярната форма на URI е URL (Uniform Resource Locator) познат като уеб адреса на даден ресурс.

[10] **Guerrini,** Mauro**, Possemato,** Tizian. Linked data: a new alphabet for the semantic wed. In: *JLIS*, Vol. 4, N 1, (Gennanio/January) 2013, p. 78

Бихме могли да твърдим, че уеб е глобална мрежа от свързани изявления. Връзките между тях са качествени, а не случайни. Всяко едно изявление може да бъде сведено до описание на модел, включващ в себе си ресурс (уеб-сайт, изображение, човек, град, абстрактни понятия и др.), свойство и стойност, което е в основата на средата RDF и се отнася до „обектно-свързан“ модел, основаващ се на идеята за създаване на декларация за уеб-ресурси под формата на израз: субект–предикат*–обект или RDF триплет (RDF triple). За означаването на елементите във всеки

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
xmlns:dc= "http://purl.org/dc/elements/1.1/">

<rdf:Description rdf:about="http://www.w3schools.com">
        <dc:description>W3Schools - Free tutorials</dc:descrip-
        tion>
        <dc:publisher>Refsnes Data as</dc:publisher>
        <dc:date>2008-09-01</dc:date>
        <dc:type>Web Development</dc:type>
        <dc:format>text/html</dc:format>
        <dc:language>en</dc:language>
</rdf:Description>
```

Схема 3. Пример за използване на пространството на имената на стандарта

* В логиката това, което се изказва, съобщава, твърди или отрича в съждението на неговия субект.

триплет се използват уникалните идентификатори. Изявлението се изразява в RDF под формата на граф*, в който върховете са обекта и субекта, а чрез свързващи дъги се изразява отношението.

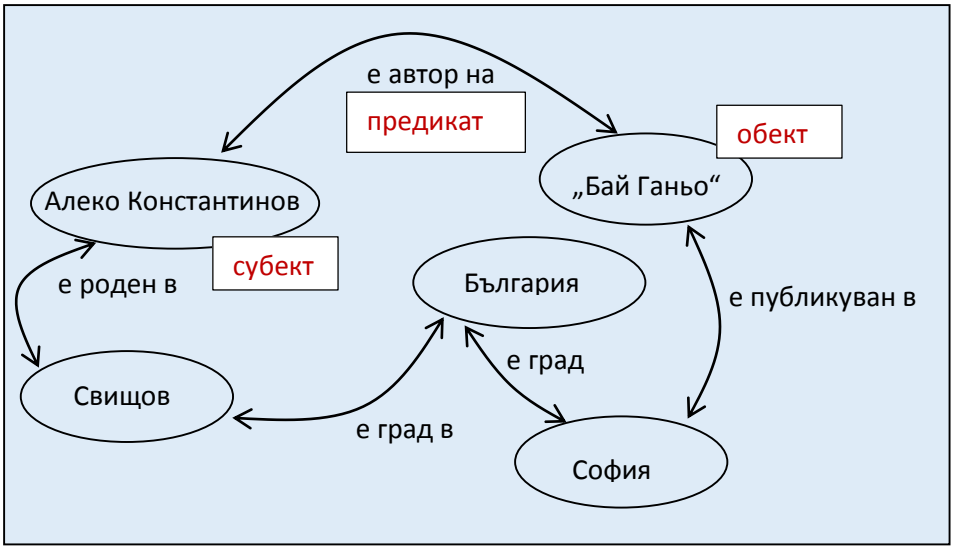

Схема 4. Пример за изразяване на свързани RDF триплети.

В семантичния уеб изявлението ще бъде:

<идентификатор за *Алеко Константинов*>
<идентификатор за *„е автор на“*>
<идентификатор за *„Бай Ганьо“*>

При използването на реалните кодове изявлението ще изглежда по следния начин:

<http://viaf.org/viaf/22183765>
<http://rdvocab.info/roles/author>
<http://worldcat.org/entity/work/id/3237700>[11].

Всяко твърдение изразено чрез RDF триплет, може да бъде генератор на нова информация. Това води до нарастване на твърденията (набори от данни) в различни области на приложение, които се преплитат постоянно. Все повече твърдения или данни се представят в уеб и са сво-

* Абстрактна структура, която представя връзките между отделните елементи на дадено множество. В компютърните науки е абстрактна структура от данни.

[11] **Coyle,** Karen**.** *Understanding the Semantic Web: Bibliographic Data and Metadata.* Chicago, ALA, 2010. Library Technology Report, Vol. 46, N 1. p. 26

бодно достъпни, обогатяват се и се превръщат в категоризирана информация (словници, речникови състави в определени области). В света на библиографската информация този принцип е приложен в изследванията на ИФЛА за функционалност на записите*, което доведе до промени в основните нормативни документи* в библиотеките, свързани с описателното и съдържателното представяне на ресурсите, обект на библиотечните и библиографските системи. Всъщност трябва да отчетем, че всички участници в процеса на създаване, опазване и управление на културното наследство на света са засегнати от този процес и все по-активно работят в посока съвместяване на усилията по представяне на обектите, които притежават в уеб пространството и тяхното включване в семантичния уеб[12].

За функционирането на този механизъм е необходима технологична инфраструктура, която да се използва за уникалното идентифициране на обектите и в която приложните програми да могат да разпознават тези обекти, да извършват асоциации и еквивалентност между тях или да бъде представена семантиката, значението на твърдението. За целта са необходими именно словници, речникови състави, в които да се представят понятия, имащи еднозначен смисъл във всякакъв контекст, таксономия (йерархически организиран списък с термини) и онтология, която да укаже термините, според определена концепция и отношенията между тях в конкретна предметна област и определен контекст[13].

Примери за речници и онтологии, познати в библиотечната сфера, са FOAF и SKOS. Онтологията FOAF[14] или Friend Of A Friend се използва за описание на лица, техните дейности, връзки с други хора или неща. Тя е много полезна за структурирането на контролни файлове.

---

* Functional Requirements for Bibliographic Records (FRBR), Functional Requirements for Authority Data (FRAD), Functional Requirements for Subject Authority Data (FRSAD). Изследванията се основават на „обектно-свързан“ модел на обектите, атрибутите и връзките в библиографските и контролните записи.

* Международни принципи на каталогизацията (ICP), Международен стандрат за библиографско описание (ISBD), Описание на ресурс и достъп до него (RDA)

[12] Вж. **Миланова,** Милена. Каталогизацията в дигиталната епоха. В: *Годишник на Софийски университет „Св. Климент Охридски“* [онлайн]. Философски факултет. Книга Библиотечно-информационни науки, 6, 2014, с. 23–32 ; **Миланова,** Милена. Дейности по каталогизация в международен контекст. В: *ББИА онлайн*, III, 2013, № 1, с. 22–24

[13] **Guerrini,** Mauro**, Possemato,** Tizian. Linked data: a new alphabet for the semantic wed. In: *JLIS*, Vol. 4, N 1, (Gennanio/January) 2013, p. 83–84

[14] **FOAF** [онлайн]. Available from: http://www.foaf-project.org/

SKOS[15] или Simple Knowledge Organization Sysytem е семейство от формални езици, създадени за представяне на тезауруси, класификационни схеми, таксономии, предметни рубрикатори и всякакъв тип контролирани речници. Това е една от първите структури реализирани на основата на RDF. ИФЛА поддържа проекти за публикуване на стандартите си в RDF чрез създаването на речников състав и онтологии за FRBR, FRAD, FRSAD и ISBD. Тези речникови състави са публикувани в регистъра за отворените метаданни Open Metadata Registry[16]. Там също така е представен и речниковият състав на англо-американските правила за каталогизация „Описание на ресурс и достъп до него“ (Resource Description and Access, RDA[17]).

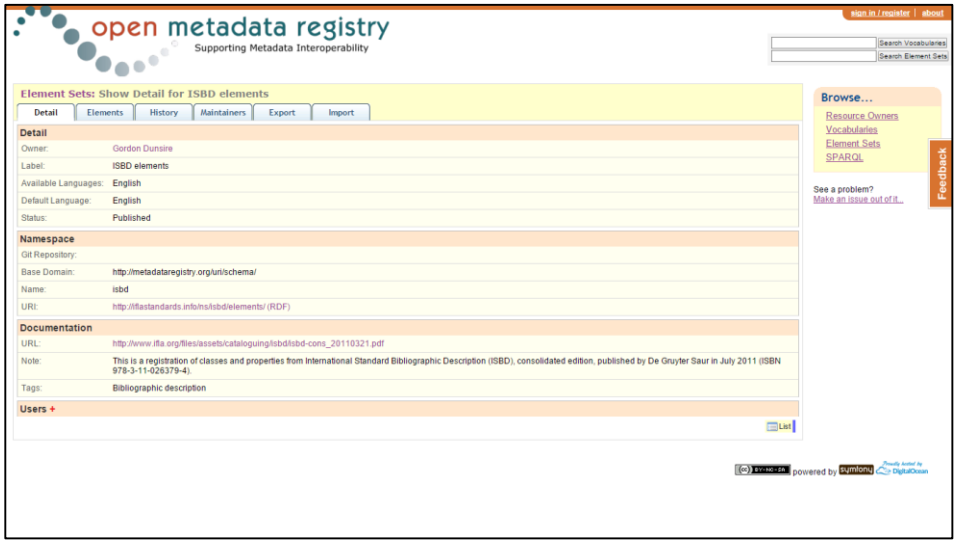

Схема 5. Наборът от данни на ISBD, публикуван в OMR.

[15] **SKOS Simple Knowledge Organization System - Home Page** [online]. Available from: http://www.w3.org/2004/02/skos/ ; Вж. и **Дипчикова,** Александра**, Тотоманова,** Антоанета**.** Промените в системите за организация на знанието – предизвикателство към българската библиотечна общност. В: *Потребностите на информационното общество и библиотеките: хармония или конфликт?*. Доклади от XXI нац. год. конф. на ББИА, Благоевград, 9–10 юни 2011. София, 2012, с. 68–78. Достъпно и от: http://www.lib.bg/publish/BBIA/sbornik_konf_2011.pdf

[16] **Open Metadata Registry** [online]. Available from: http://metadataregistry.org/
Място, създадено от W3C за подкрепа и използване на контролирани речници и където се преставят и съхраняват онтологии от различни области.

[17] **RDA toolkit** [online]. Available from: https://access.rdatoolkit.org/

През 2006 г. Тим Бърнърс-Лий (Tim Berners-Lee) представя идеята за свързаните данни, която показва начина за изграждане на семантичния уеб[18]. Той определя четири правила за създаване на свързани данни: „използвай URI като наименование на нещата; използвай HTTP URI, така че наименованията да могат да бъдат търсени от хора; когато някой търси URI, осигури достатъчно полезна информация, чрез използването на стандарти (RDF, SPARQL*); включи връзки към други URI, така че да могат да се разкриват повече неща.“ (прев. е мой – ММ). Акцентът е върху използването именно на уникалните идентификатори, които са ключовият елемент за реализиране на идеята за свързани данни.

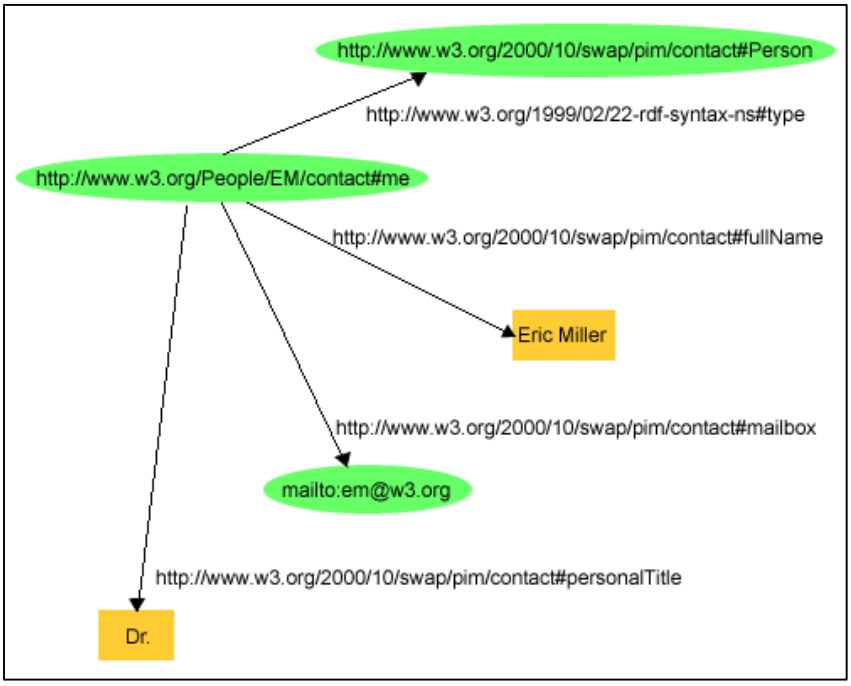

Схема 6. Граф RDF описващ лице[19].

В примера по-долу Eric Miller, като индивид, се идентифицира с уникален идентификатор http://www.w3.org/People/EM/contact#me; това,

[18] **Berners-Lee**, Tim. *Linked Data* [online]. 2006-07-27, Last change: 2009/06/18 18:24:33. Available from: http://www.w3.org/DesignIssues/LinkedData.html

* Команден език разработен за свързаните данни.

[19] **RDF Primer** [online]. Available from: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/#figure4

че той е човек се идентифицира с http://www.w3.org/2000/10/swap/pim/contact#Person; различни негови свойства, качества, се идентифицират с http://www.w3.org/2000/10/swap/pim/contact#mailbox; стойността на тези свойства е например пощенския адрес mailto:em@w3.org.

Същият граф може да се изрази с помощта на език RDF/XML:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf=http://www.w3.org/1999/02/22-rdf-syntax-ns#
         xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">

  <contact:Person
rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Eric Miller</contact:fullName>
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>

</rdf:RDF>
```

Схема 7. RDF/XML[20].

Създаването на свързани данни означава да се изрази значението на информацията, правейки я възможна за споделяне сред различни приложения и използваема от приложения, различни от тези, за които тя е била първоначално създадена. Това означава, че свързаните данни са подход, който дава добра отправна точка за постигане на стратегическата цел – библиотечните данни да са достъпни за търсене от всякакви инструменти[21]. Свързаните данни позволяват в уеб да се публикуват данни във формат, който осигурява възможност те да бъдат четени, интерпретирани и най-вече разбирани от машините и чието значение е експлицитно

[20] Пак там.

[21] **Guerrini,** Mauro**, Possemato,** Tizian. Linked data: a new alphabet for the semantic wed. In: *JLIS*, Vol. 4, N 1, (Gennanio/January) 2013, p. 77–78

дефинирано с помощта на низ (поредица) от думи и маркери. Създава се мрежа от свързани данни, които принадлежат на първосъздателя на основния текст, свързват се с други външни набори от данни, които са извън създателя на основния текст, в контекста на постоянно нарастващи връзки. След това те се представят в облак с отворени свързани данни (LOD)[22].

Традиционният библиографски запис е съставен от стойностите на множество атрибути или характеристики, асоциирани с библиографски обект – например заглавие или физически описание. Възможно е това описание да се разложи в запис от отделни RDF триплети. Добавяйки ги към пространството на семантичния уеб и установявайки еквивалентност между различните идентификатори за отделните библиографски обекти, ще се постигне и повиши ползата им и чрез допълването на триплети и от други общности като издатели, книжари, онлайн енциклопедии, социални сайтове, архиви, музеи[23]. Анализирайки облака на отворените свързани данни може да се види, че част от него се състои от данни, създавани и имащи отношение към библиотечното пространство.

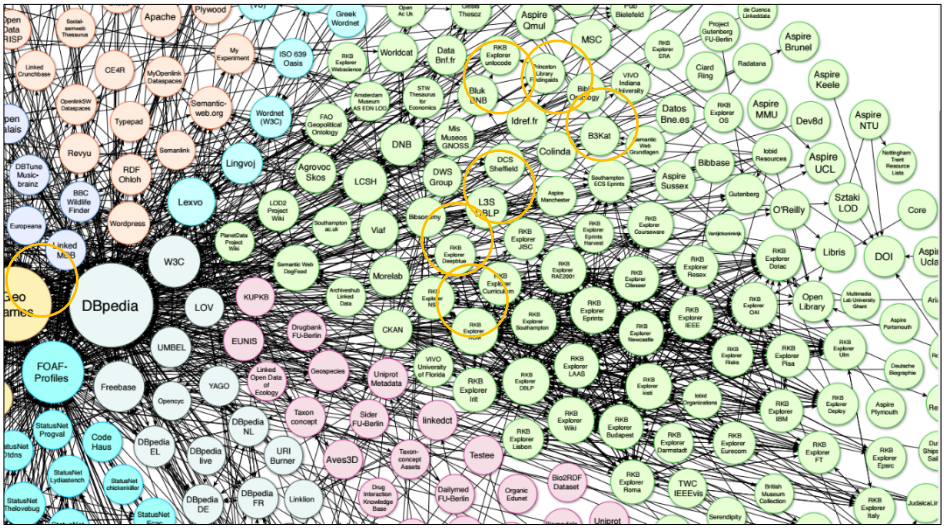

Схема 8. Примери за различни речникови набори в LOD[24].

[22] Пак там, p. 67

[23] **Dunsire,** Gordon**, Willer,** Mirna. Standard library metadata models and structures for the Semantic Web. In: *Library Hi Tech News.* Emerald Group Publishing Limited, 2011, № 3, p. 7–8

[24] **The Linking Open Data cloud diagram** [online]. [Viewed 12.07.2015]. Available from: http://lod-cloud.net/versions/2014-08-30/lod-cloud.svg

Непрекъснатото добавяне на данни в LOD обогатява възможностите за търсене и свързване на информация. С представянето на данните от библиотечните каталози в уеб пространството то ще се обогати с множество интересна и сигурна информация, която до този момент не е била достъпна. Възможността да се свързват данни и да се извлича смисъл от тях, изостря още повече проблема с произхода на тези данни. Библиотеките, традиционно създават информация, която е проверена и на която може да се разчита. Така например VIAF (The Virtual International Authority File) представлява набор от контролирани речници (национални контролни файлове) на имената на лица и колективни органи, създавани и поддържани от националните библиографски агенции на различни страни. Имената от един речников състав си кореспондират с имената от други речници и всички тези набори от имена и връзки са достъпни като свързани данни[25].

В библиотечното пространство се разработват различни модели и проекти, които представят библиотечните данни като свързани данни. През 2011 г. Библиотеката на Конгреса на САЩ, Библиотеката на университета Станфорд и консорциума W3C представят работата си по свои проекти, които целят да направят лесно достъпни и отворени библиотечните библиографски данни[26]. Европеана също създава модел за представяне на данните в LOD. Тя преработва своя модел на метаданни, като го прави да отговаря на изискванията за създаване на свързани данни[27]. Британската библиотека реализира свой модел на свързани данни. Тя използва различни съществуващи речникови състави и онтологии като VIAF, предметния рубрикатор на Библиотеката на Конгреса (Library of Congress Subject Headings, LCSH), GeoNames (база от данни с географски наименования), кодовете на страните и езиците в MARC, Dewey.info (основни класове от класификационната схема на Дюи) и др. През 2013 г. този модел е приет за национална информационна инфраструктура на Великобритания[28].

---

[25] **Dunsire,** Gordon**, Willer,** Mirna. Standard library metadata models and structures for the Semantic Web. In: *Library Hi Tech News.* Emerald Group Publishing Limited, 2011, № 3, p. 9

[26] Това са проектите: BIBFRAME (http://www.loc.gov/bibframe/); Stanford Linked Data Project (https://lib.stanford.edu/stanford-linked-data); W3C Library Linked Data Incubator Group (http://www.w3.org/2005/Incubator/lld/)

[27] Първоначалният модел е The Europeana Semantic Elements specification, който е преработен на Europeana Data Model (EDM).

[28] **Linked Open BNB**. In: *British Libraray* [online]. [Viewed on 24.07.2015]. Available from: http://www.bl.uk/bibliographic/datafree.html; Path: Home; Collection Metadata; Data Services; Free data

Организации като ИФЛА, Комитетът по развитие на RDA, Британската библиотека, Библиотеката на Конгреса на САЩ, OCLC са водещите организации, които подкрепят и развиват идеята за представянето и участието на библиотеките в изграждането на свързаните данни и семантичния уеб. Пред българската библиотечна общност стои предизвикателството да стане част от тези процеси. Необходимо е усилено да работим по стандартизирането на библиотечните и библиографските ни данни, да изграждаме своите бази от данни, следвайки правилата и изискванията на водещите библиотечни организации и не на последно място да подготвим, споделим и представим своите бази от данни в общото уеб пространство.