RDF Resources of Saudi Open Data

Khalid Aloufi
Department of Computer Science
College of Computer Science and Engineering
Taibah University
Email: koufi@taibahu.edu.sa

***Abstract*—Different organizations and governments are working toward publishing their data on the web to achieve a variety of goals. Published data facilitate transparency and cooperation. Currently, most of the data available on the web are not easily accessible, which limits their application. The Semantic Web technology standards are used to access data with a specific representation: Resource Description Framework (RDF). Some published data require transformation to the Semantic Web standard formats to ensure that they are accessible. This work investigates Saudi open data and provides an algorithm to generate RDF resources. Additionally, some recommendations and analyses are discussed.**

***Index Terms*—Semantic Web, Linked Open Data (LOD), Open Data.**

I. Introduction

The world wide web (WWW) is the most important source of information today. online information covers all types of topics ranging from scientific and social to news. Most information is not structured and not identified by metadata that convey its meaning. Thus, data-mining techniques have been developed to classify and understand this information.

To help search engines and agents in general find the desired information according to its meaning, the WWW community has developed methods of tagging information. By providing such a description of the information, a layer of abstraction will be added to the information used by agents. For this purpose, the W3C recommends the Semantic Web [1]. The Semantic Web is a web framework used to publish data about information available on the web [1]. This information is published in defined formats and linked together to create more meaningful and connected data.

The published data are thus easily accessible, increasing the opportunity to apply the data in new ways. This is currently a hot topic among researchers in the field of web technologies [2]. More awareness of this issue is evidenced by the increasing amounts of open data published by various organizations and governments. One example is Saudi open data.

This work investigates the Saudi open data published as part of the Saudi open portal [3] [4]. The Saudi open portal provides data in a variety of resource formats but not in RDF format. This work presents a methodology to create RDF versions of available resources.

This paper is organized as follows. After the Introduction, a summery of the Semantic Web and linked data is provided. Then, Saudi open data are introduced with detailed statistics. Next, a methodology to create RDF resources using existing datasets is proposed, followed by the results of the study and some recommendations and analyses. Finally, the last section presents the conclusions and recommends some future investigations.

II. Semantic Web and Linked Data

This section summarize the Semantic Web and linked open data (LOD). Most information on the web is accessible through search engines or indexing. However, search engines do not access all available data. Indeed, most data available on the web are not easily accessible, limiting the range of potential applications. Additionally, some data are behind a firewall or protected by other security measures.

To increase the data connectivity, LOD are organized on the web according to defined rules [5], for example, through the use of HTTP URIs that provide detailed information and links to other URIs. The Semantic Web is a data framework for publishing data on the web in a way that is transparent [1]. Because of the increased awareness of LOD, the amount of LOD is increasing yearly [6]. Based on the Semantic Web requirements, the information resources on the web are represented in specific formats.

The data are represented using the Resource Description Framework (RDF) [7]. RDF is the main format for data representation in the Semantic Web and is recommended by W3C. Semantic Web technology standards are used to access data with a specific representation. RDF is composed of Subject, Predicate and Object, where Subject represents the entity identifier, Predicate represents the attribute name, and Object represents the attribute value. RDF graphs consist of the connected nodes of Subjects and Objects where Predicates represent the edges. The RDF format can be queried by SPARQL, which is the W3C-recommended query language [8].

Currently, organizations and governments are publishing their data on the web to ensure transparency and cooperation with stakeholders. Various vocabularies are used to organize and add definition and structure to the data. Data Catalog Vocabulary (DCAT) is one such vocabulary and is recommended by W3C for RDF [9]. DCAT defines the connectivities between published catalogs on the web. When datasets are tagged using DCAT, the datasets become more available and accessible to agents and search engines.

When information is published using DCAT and can be accessed, it becomes available for agents to use. Furthermore, when information or resources are available in RDF format, agents can easily connect to and query the data. Indeed, the richer the associated metadata, the more useful and accessible the information becomes. The next section presents an example of using the Semantic Web and LOD.

III. Saudi Open Data

In this section, the Saudi open portal is presented, with emphasis on open data. For more information, please visit the main Saudi portal website [3]. The Saudi portal provide various services to the community, and one of the main services is the provision of open data. Saudi Open Data are accessible via the following link to the Saudi Open Data website: http://data.gov.sa [4].

The datasets of the Saudi Open Data website are available at http://data.gov.sa/dataset. The simple RDF data graph of the data in these resources is shown in Figure 1. The open data website is divided into some sections. These sections are filtered, as in the tables showing the group I, publisher II, resource format III and license IV filters. These divisions are then subdivided into datasets, which are presented in different resource formats, as shown in Figure 2. There are 319 a total of datasets and 6,744 user-to-user resources. [4].

The available resources are published in different formats, including XML, xlx, xlsx and jpg. The header of the page identifies the prefixes for regular XML and HTML. The page then presents the title of the dataset. Some data are defined by the metadata term, which include a value for each property. These terms are maintained by the Dublin Core Metadata Initiative [10] and include the following:

- property="dc:group"
- property="dcterms:modified"
- property="dcterms:issued"
- property="dcterms:identifier"
- property="dc:license"
- property="dc:publisher"

The data are represented in a table of Fields and Values. For example, the Field Group is identified by property="dc:group". Then, property="dcat:distribution" indicates the section of the resources. Resources can be accessed by direct-download URL. Additionally, links to other pages are provided to access detailed information about the resource. The data are maintained by the DKAN Datastore API [11]. Each resource format is used to present a different type of data. According to our investigation, no data are available in multiple resources. Additional investigations will be required to measure the accuracy of the data and ensure that no duplication was present. Some of the data in the resources have not been updated since 2012, which is an important piece of evidence relating to how open data should be handled and will be discussed further below. Three methods are used to access resource-identifying resources with the property="dcat:Distribution":

- direct links, text and images can be used for resource-information pages identified by property="dcat:accessURL", property="dc:format", and data-format="xls".
- direct links to the resource-information page described as "Explore Data".
- and direct-download URLs.

Saudi open data use the Semantic Web standards in some parts of information tagging. However, the resources are not published in the recommended format. The following section presents an algorithm able access one format and publish a RDF version of the accessed data.

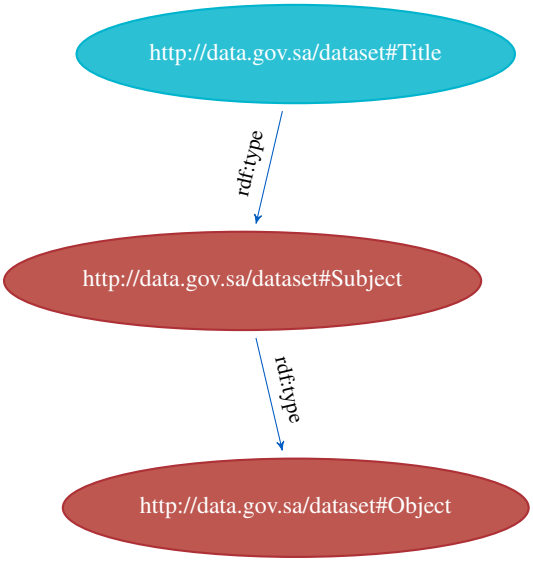

Fig. 1. http://data.gov.sa/dataset data graph

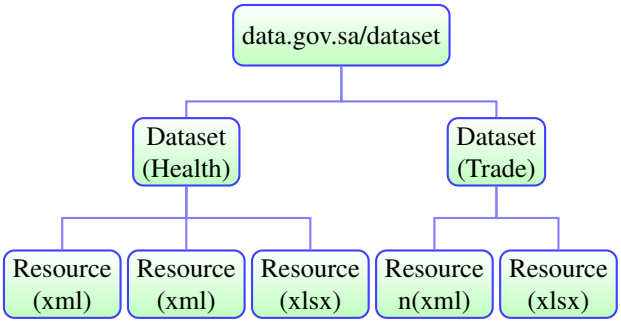

Fig. 2. Datasets and Resources

TABLE I
Filtered by group

Datasets	Resources
Social Services	(52)
Health	(51)
Accounts of Financial Monetary Affairs and Industry	(29)
Transportation and Communications	(29)
Agriculture and Fishing	(27)
Population and Housing	(20)
Education and Training	(16)
Energy and Water	(14)
Labor Market	(14)
Trade (internal and external)	(14)
Arab Gulf Cooperation Council (GCC)	(12)
Industry	(12)
Weather Conditions	(12)
Prices and Indices	(8)
Social Insurance	(7)

A. Algorithm

As described earlier, Saudi open data are presented in different formats. However, not using the RDF format limits the ability of Semantic Web applications to access the data. Furthermore, the information is not connected and is not linked to any other resources in the LOD graphs. This paper investigates one resource format used for Saudi

TABLE II
FILTERED BY PUBLISHER

Datasets	Resources
Central Department of Statistics and Information	(52)
Ministry of Health	(47)
Ministry of Agriculture	(21)
Information Sector - Secretariat General of the GCC	(12)
Presidency of Meteorology and Environment	(11)
Central Department of Statistics and Information - National Income Statistics	(9)
Ministry of Labor - Ministry's Agency for Planning and Development - Information Center	(9)
Ministry of Social Affairs	(9)
Saudi Arabian Monetary Agency	(9)
Ministry of Commerce and Industry	(8)
Ministry of Municipal and Rural Affairs	(8)
Saudi Post	(8)
Annual Statistical Report of General Organization for Social Insurance	(6)
Central Department of Statistics and Information, Annual Economic Survey of Establishments	(6)
General Organization for Technical and Vocational training	(5)
Ministry of Civil Service	(5)
Ministry of Petroleum and Minerals	(5)
Ministry of Culture and Information	(4)
Ministry of Water and Electricity	(4)
Saudi Arabian Airline Organization	(4)
Administration Of General Budget	(3)
Central Dept. of Statistics and Information - Social Statistics	(3)
Department of Statistics and Information, Economic Survey 2012	(3)
General Organization for Desalination	(3)
Ministry of Justice	(3)
Port Authority	(3)
Saudi Government Railroad Organization	(3)
Cement Companies	(2)
Communication and Information Technology Commission (CITC)	(2)
Customs Department	(2)
General Directorate of Civil Defense "info.stat.dept"	(2)
General Presidency for Youth	(2)
Human Resources Development Fund	(2)
Ministry Of Finance (Budget Division)	(2)
Ministry of Higher Education	(2)
Ministry of Interior	(2)
Ministry of Interior - General Directorate of Traffic	(2)
Ministry of Islamic Affairs and Endowments and Da'awa and Guidance	(2)
Saudi Arabian Basic Industries Company (SABIC)	(2)
Saudi Commission for Tourism and Antiquities	(2)
Saudi Geological Survey	(2)
Saudi Industrial Development Fund	(2)
Saudi Red Crescent Authority	(2)
Board Of Grievance	(1)
General Organization for Silos and Mills	(1)
Institute of Public Administration	(1)
Ministry of Finance	(1)
Public Pension Agency	(1)
Saudi Arabian Agricultural Bank	(1)
The National Gypsum Co.	(1)

TABLE III
FILTERED BY RESOURCE FORMAT

Datasets	Resources
xls	(287)
xlsx	(226)
jpg	(35)
xml	(5)
xlb	(2)
jpeg	(1)

TABLE IV
FILTERED BY LICENSE

Datasets	Resources
not specified	(317)
other/open	(2)

open data: the xls format. The following algorithm is able to access data in the xls format. It then reads the data in the file and replicates the data in RDF format. As mentioned previously, some information contained in such files also consists of DCAT metadata, which are available in the source code of the page in the main website.

This algorithm could be extended to compare these metadata and add them to the divisions used to categorize RDF resources. The algorithm is interacts with the Saudi open data as shown in Algorithm 1.

Algorithm 1 RDF Resources of Saudi Open Data

Input: http://data.gov.sa/dataset

Output: RDF Resources connect to the website of the open data

1: obtain the URIs according to the DCAT vocabulary
2: access the data included in resources in different formats
3: apply information restructuring according the format of the resource (xls in this paper)
4: apply the data transformation to generate the RDF format
5: apply the following rules for the table in the xls file:
- Subject represents the row header
- Predicate represents the column header
- Object represents the cell value.

6: publish the data resources in RDF using the DCAT vocabulary

The algorithm generate a complete RDF file from the xls file. The following section shows the result of the algorithm. The algorithm will have to be extended to process the other resource formats used for Saudi open data. The following section presents the result of applying the described algorithm.

B. Generated RDF

The RDF file generated from the algorithm defined above is composed of all the rows in the xls table. In this paper, the algorithm was applied to a single file. The same algorithm could be applied to additional resources in different formats.

Figure 3 shows the prefixes generated using the data in the header of the RDF file. Figure 4 presents an image of the original table of the xls file. The algorithm was applied to the dataset "Economic Indicators for Banking and Insurance Activity and Establishment Size, 2001 A.D." The resource information can be found at [12]. Figures 5, 6, 7, 8, 9 and 10 give the RDF descriptions of total Number of employees, Employees, Compensation, Non-Saudis, Expenditure, Saudis, and Revenues, respectively.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="http://data.gov.sa/dataset#    الجملة    (Total)"
    xmlns:j.1="http://data.gov.sa/dataset#9-1"
    xmlns:j.2="http://data.gov.sa/dataset#99-50"
    xmlns:j.3="http://data.gov.sa/dataset#الجملة (Total)"
    xmlns:j.4="http://data.gov.sa/dataset#100+"
    xmlns:dataURI="http://data.gov.sa/dataset#"
    xmlns:j.5="http://data.gov.sa/dataset#49-10" >
```

Fig. 3. RDf Prefixes

الجملة (Total)	الجملة (Total)	100+	99-50	49-10	9-1	المؤشـــر
Total No. of Employees	38,659	27,499	730	5,304	4,681	اجمالي المشتغلين
Saudi	20,782	15,668	309	2,449	2,356	سعوديون
Non-Saudi	17,877	12,276	421	2,855	2,325	غير سعوديين
Employees compensation (Million of Riyals)	7,569	5,630	190	717	1,032	تعويضات المشتغلين(مليون ريال)
Revenues (Million of Riyals)	55,609	32,107	104	9,522	13,876	الايرادات (مليون ريال)
Expenditurs (Million of Riyals)	37,305	21,076	858	6,696	8,675	النفقات (مليون ريال)

Fig. 4. Economic Indicators for Banking and Insurance Activity and Establishment Size, 2001 A.D.

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#اجمالي المشتغلين">
  <j.3:_>Total No. of Employees</j.3:_>
  <j.0:_>38659.0</j.0:_>
  <j.4:_>27499.0</j.4:_>
  <j.2:_>730.0</j.2:_>
  <j.5:_>5304.0</j.5:_>
  <j.1:_>4681.0</j.1:_>
  <dataURI:اجمالي المشتغلين<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 5. Total No. of Employees

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#تعويضات المشتغلين( مليون ريال)">
  <j.3:_>Employees compensation (Million of Riyals)</j.3:_>
  <j.0:_>7569.0</j.0:_>
  <j.4:_>5630.0</j.4:_>
  <j.2:_>190.0</j.2:_>
  <j.5:_>717.0</j.5:_>
  <j.1:_>1032.0</j.1:_>
  <dataURI:تعويضات المشتغلين( مليون ريال)<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 6. Employee compensation (Million Riyals)

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#غير سعوديين">
  <j.3:_>Non-Saudi</j.3:_>
  <j.0:_>17877.0</j.0:_>
  <j.4:_>12276.0</j.4:_>
  <j.2:_>421.0</j.2:_>
  <j.5:_>2855.0</j.5:_>
  <j.1:_>2325.0</j.1:_>
  <dataURI:غير سعوديين<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 7. Non-Saudis

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#النفقات ( مليون ريال)">
  <j.3:_>Expenditurs (Million of Riyals)</j.3:_>
  <j.0:_>37305.0</j.0:_>
  <j.4:_>21076.0</j.4:_>
  <j.2:_>858.0</j.2:_>
  <j.5:_>6696.0</j.5:_>
  <j.1:_>8675.0</j.1:_>
  <dataURI:النفقات ( مليون ريال)<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 8. Expenditure (Million Riyals)

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#سعوديون">
  <j.3:_>Saudi</j.3:_>
  <j.0:_>20782.0</j.0:_>
  <j.4:_>15668.0</j.4:_>
  <j.2:_>309.0</j.2:_>
  <j.5:_>2449.0</j.5:_>
  <j.1:_>2356.0</j.1:_>
  <dataURI:سعوديون<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 9. Saudis

```
<rdf:Description rdf:about="http://data.gov.sa/dataset#الايرادات ( مليون ريال)">
  <j.3:_>Revenues (Million of Riyals)</j.3:_>
  <j.0:_>55609.0</j.0:_>
  <j.4:_>32107.0</j.4:_>
  <j.2:_>104.0</j.2:_>
  <j.5:_>9522.0</j.5:_>
  <j.1:_>13876.0</j.1:_>
  <dataURI:الايرادات ( مليون ريال)<_المؤشـــر</dataURI:_المؤشـــر>
</rdf:Description>
```

Fig. 10. Revenues (Million Riyals)

```
PREFIX sa:<http://data.gov.sa/dataset#>
PREFIX rdf:<http://www.w3.org
    /1999/02/22-rdf-syntax-ns#>
                SELECT ?s ?p ?o
                WHERE {
                ?s ?p ?o.
```

Fig. 11. RDF SPARQL QUERY

The following section applies SPARQL queries to the generated RDF file.

C. RDF SPARQL Query

After producing the RDF resources, the included data become available for querying. The data are queried by SPARQL. Figure 11 shows a query of all the data, which returned all the data in the RDF file. For comparison, Figure 12 returns the data from columns ”49-10_”, as shown in Figure 13.

Having the SPARQL capable of inquiring the RDF file means the RDF resource is in correct RDF format. Recommendation are presented in the next section.

IV. Recommendations

By applying the W3C recommendations to a portion of the Saudi open data, it became evident that this procedure should be extended to all Saudi open data. Saudi open data are rich in information relevant to different sectors but are not used by many commercial entities, communities or researchers to develop services. The data were created using

```
PREFIX sa:<http://data.gov.sa/dataset#>
PREFIX rdf:<http://www.w3.org
    /1999/02/22-rdf-syntax-ns#>
PREFIX xsd:<http://www.w3.org/2001/
    XMLSchema#>
                SELECT ?s ?o
                WHERE {
                ?s "sa:49-10_" ?o.
```

Fig. 12. RDF SPARQL QUERY for a specific column

```
-------------------------------------------------------------------------
| s                                                          | o        |
=========================================================================
| <http://data.gov.sa/dataset#6696.0" |          "النفقات ( مليون ريال)> |
| sa:10-49" |                                          "المؤشـــر  |
| sa:2449.0" |                                          "سعوديون |
| <http://data.gov.sa/dataset#9522.0" |       "الايرادات ( مليون ريال)> |
| <http://data.gov.sa/dataset#717.0" | <"تعويضات المشتغلين( مليون ريال)  |
| <http://data.gov.sa/dataset#5304.0" |              "اجمالي المشتغلين> |
| <http://data.gov.sa/dataset#2855.0" |                 "غير سعوديين> |
-------------------------------------------------------------------------
```

Fig. 13. Query Results

various commonly used formats and published on the web site using an API [11]. However, as a result, an API is also required for processing all of the resources, regardless of whether they are presented in RDF or another format. The available resources are not published in formats that facilitate programmers accessing and processing the data because the formats are not defined. Defined formats and API help increase the accessibility of data by agents for these data in all formats used in the open data portal.

The data in the Saudi open portal are ready for human investigation but are not suitable for smart agents. The data are required to be in the format recommended by the World Wide Web Consortium (W3C) [13]: RDF. The data in the Saudi portal are added using an API, as recommended by the W3C. However, the data are contributed from wide range of publishers. Hence, they should be published using an agent that connects the various publishers to the Saudi open data. These publishers thus need a methodology to publish the same datasets in different formats and thereby increase the range of applications that could utilize the reported data. Additionally, most data available on the website are static, and no active dynamic data, such as stock data or weather data, are included.

V. Conclusion and future work

The Semantic Web and linked data are W3C standards required to publish the data in the web used by a variety of organizations. Open data are assumed to facilitate additional services for the community and to enhance the value of the available data. In conclusion, Saudi open data are very good example of published data. However, substantial research and development will be needed to make the data accessible to agents. Additionally, those who benefit from access to these data should develop services for the community.

In the future, the algorithm will be extended to generate RDF resources from data presented in different formats. Additionally, the best methodologies to process different data resource formats to generate RDF files will be investigated, the data available on the website will be statistically analyzed, and a general framework for publishing LOD in open portals will be developed.

Acknowledgment

The author would like to thank Taibah University for supporting this research.

References

[1] Semantic Web, " Semantic Web - W3C - World Wide Web Consortium)" w3.org, 2015. [Online]. Available: http://www.w3.org/standards/semanticweb/. [Accessed: Jul. 30, 2015].

[2] Wouter Beek and Laurens Rietveld and Hamid R. Bazoobandi and Jan Wielemaker and Stefan Schlobach, "LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data," in The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I, 2014, pp. 213–228.

[3] SAUDI National e-Government Portal, "National Portal", " saudi.gov.sa, 2015. [Online]. Available: http://www.saudi.gov.sa/wps/portal/. [Accessed: Jul. 30, 2015].

[4] SAUDI National e-Government Portal, "Saudi Government Open Data portal," data.gov.sa, Jan. 31, 2001. [Online]. Available: http://data.gov.sa/. [Accessed: Jul. 30, 2015].

[5] Tim Berners-Lee, "Linked Data," w3.org, Jun. 18, 2009. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed: Jul. 30, 2015].

[6] Linking Open Data W3C SWEO Community Project, "SweoIG/-TaskForces/CommunityProjects/LinkingOpenData," w3.org, Jan. 31, 2001. [Online]. Available: http://www.w3.org/wiki/SweoIG/ TaskForces/CommunityProjects/LinkingOpenData. [Accessed: Jul. 30, 2015].

[7] RDF Working Group, "RDF," w3.org, Feb. 25, 2014. [Online]. Available: http://www.w3.org/RDF/. [Accessed: Jul. 30, 2015].

[8] Eric Prud'hommeaux, W3C ¡eric@w3.org¿, Andy Seaborne, Hewlett-Packard Laboratories, Bristol ¡andy.seaborne@hp.com¿, "SPARQL Query Language for RDF," w3.org, Jan. 15, 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/. [Accessed: Jul. 30, 2015].

[9] Fadi Maali, DERI, NUI Galway John Erickson, Tetherless World Constellation (RPI) , "Data Catalog Vocabulary (DCAT)" w3.org, Jan. 16, 2014. [Online]. Available: http://www.w3.org/TR/vocab-dcat/. [Accessed: Jul. 30, 2015].

[10] DCMI Usage Board, "DCMI Metadata Terms" http://dublincore.org, Jun. 14, 2012. [Online]. Available: http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms. [Accessed: Jul. 30, 2015].

[11] nucivic, "DKAN The all in one open data platform," nucivic.com, 2015. [Online]. Available: http://nucivic.com/dkan/. [Accessed: Jul. 30, 2015].

[12] Central Department of Statistics an Information, Annual Economic Survey of Establishments, "Accounts Financial Monetary Affairs and Industry," data.gov.sa, Jan. 31, 2001. [Online]. Available: http://www.data.gov.sa/dataset/economic-indecators. [Accessed: Jul. 30, 2015].

[13] W3C, " World Wide Web Consortium (W3C)" w3.org, 2015. [Online]. Available: http://www.w3.org/. [Accessed: Jul. 30, 2015].