Methodology for Assessment of Open Data

ROUMEN TRIFONOV, GALYA PAVLOVA
Faculty of Computer Systems and Technology
Technical University of Sofia
8 Kliment Ohridski Bul., Sofia 1000
RADOSLAV YOSHINOV
Laboratory of Telematics
Bulgarian Academy of Sciences,
Akad.G.Bonchev St, bl. 8, 1113 София
BOYAN JEKOV
Faculty of Information Sciences
University of Library Studies and Information Technologies
119 Tszrigradsko shoes Bul., Sofia
BULGARIA
r_trifonov@tu-sofia.bg, yoshinov@cc.bas.bg, boyan.jekov@gmail.com, raicheva@tu-sofia.bg

Abstract: - The Open Data goal is similar to the other open data movements – Open Sources, Open Access etc. Although the Open Data philosophy is defined long time ago, the term gains popularity with the Internet and World Wide Web - WWW rise and mostly with the Open Data initiatives Data.gov and Data.gov.uk. The curiosity of the world enforces the Big Data to become Open and then to connect the available open data in linked. In this article is presented a comprehensive review of established methodologies for assess the data quality. It is proposed a multistep integrated approach for quality assessment of Open data as the methodology for its evaluation.

Key-Words: - Open data, linked data, open knowledge, open data cloud, methodology, assessment

1 Introduction

Open Data presents an ideology according to which a certain data and content should be freely accessible to be used by all without restrictions such as copyrights, patents and other control mechanisms. The Open Data goal is similar to the other open data movements – Open Sources, Open Access etc. Although the Open Data philosophy is defined long time ago, the term gains popularity with the Internet and World Wide Web - WWW rise and mostly with the Open Data initiatives Data.gov and Data.gov.uk.

Formal definition of Open Data provided by the organization Open Definition: "Some of the data or content is considered" open "/ free, if everyone is free to use, reusing and relays - subject are most, the requirement to determine and sharing. "Most often, Open data includes non-textual materials such as maps, genomes connections (improved map of neural connections in the brain), chemical formulas, mathematical and scientific formulas, medical data, biodata and biodiversity data. The problem is that these data are of commercial value or are part of the creative and scientific activities of individuals. Access to them is controlled by private or public organizations, through restrictions, licenses, copyrights, patents or fees for use and reuse. The Open data movement aims to make the data freely available to all. As a result “Linking Open Data community project” has started in 2007. The goal of this project is to expand the World Wide Web by publishing free datasets in RDF form or linking different datasets with RDF connections. [1] Fig. 1 shows a diagram until now of linked datasets. In 2010, all datasets (private and public) are incorporated in the project "Linked Open Cloud Data" and maintained by Richard Cyganiak and Anja Jentzsch. The diagram is presented in Fig. 2. The rapid increasing with new datasets of the project "Linked Open Cloud Data" is shown on the diagram from 2014 presented in Fig. 3. Unlike "Linking Open Project Data Community ", "Linked Open Data Cloud" is based metadata collected and organized by individual persons and organizations that are not present currently in "Linking Open Project Community Data".

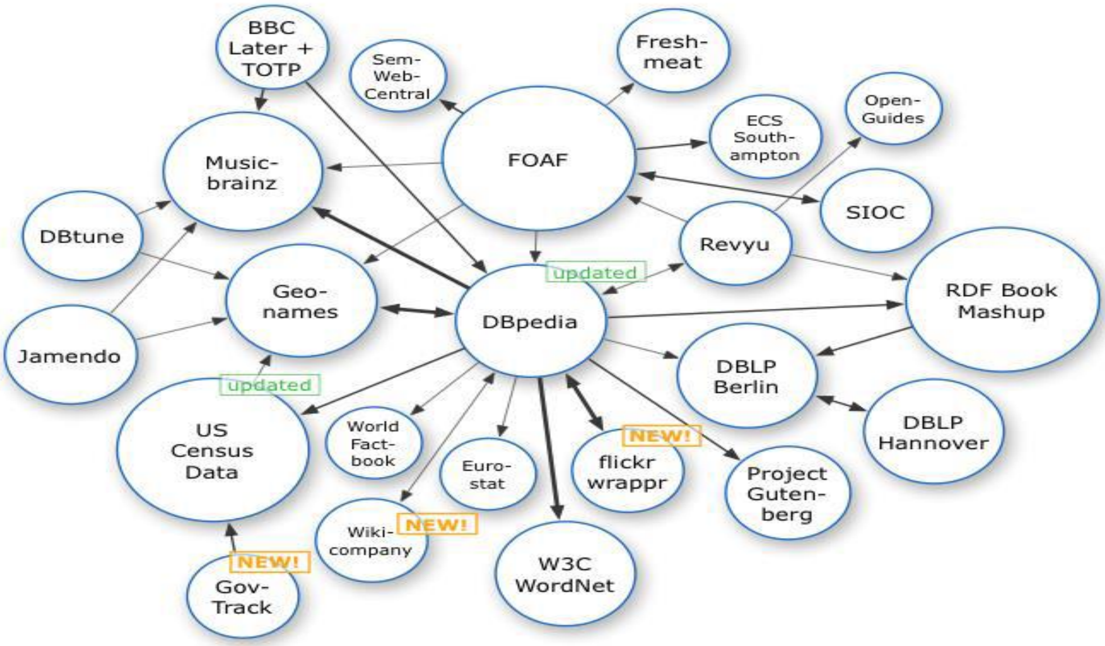

Fig. 1. Updated diagram of Datasets include in Open Data space

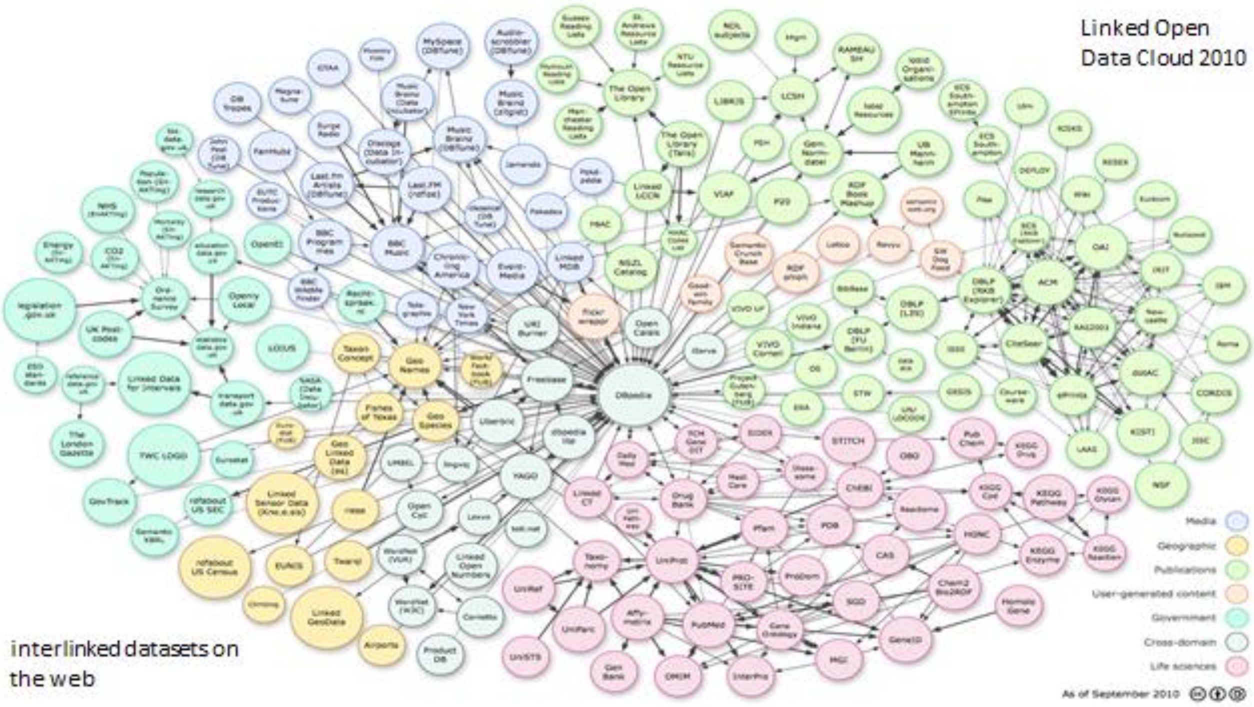

Fig. 2. Updated diagram of Linked Open Data Cloud 2010

2 Open Data

2.1 Definitions for Open data and Open knowledge

Wikipedia: Open data is the idea that some data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

The Open Definition: “Open data and content can be freely used, modified, reused, redistributed and shared by anyone for any purpose”. Elements - Availability and Access; Re-use and Redistribution; Universal Participation.

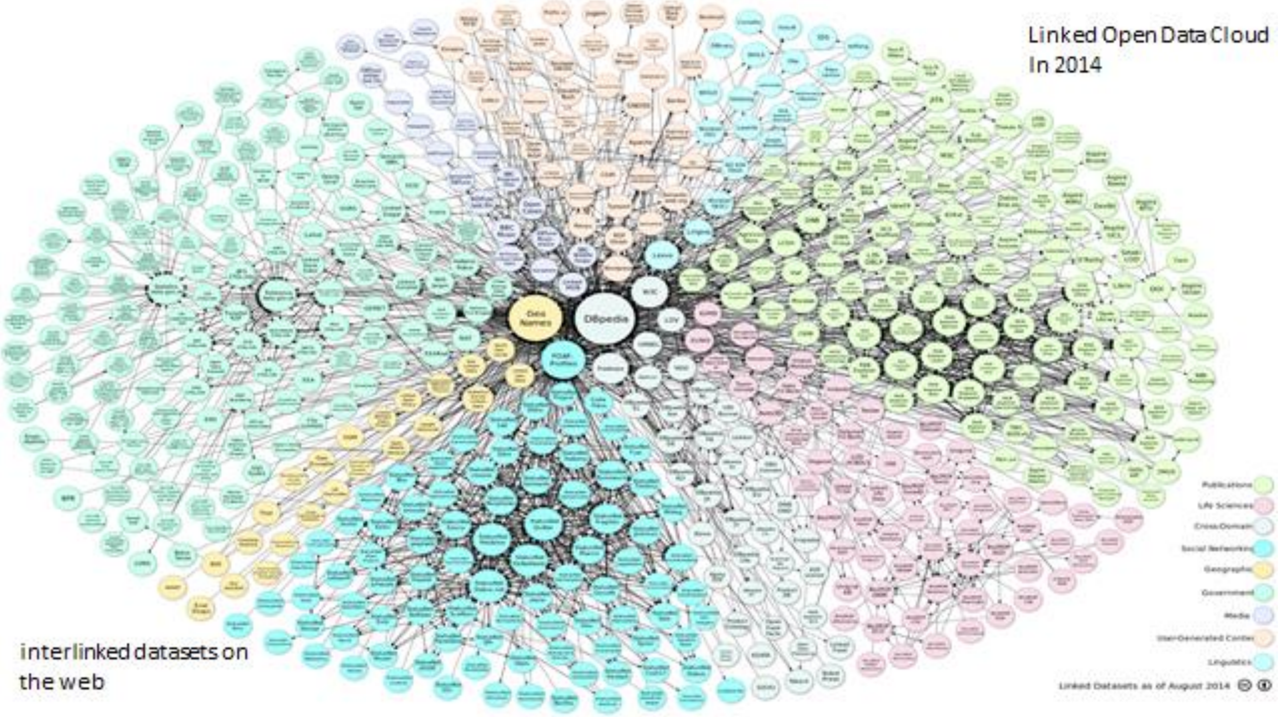

Fig. 3. Updated diagram of Linked Open Data Cloud 2014

Open government data is data produced or commissioned by government or government controlled entities, which is open as defined in the Open Definition. WHY? Transparency; Releasing social and commercial value; for Participatory Governance.

Open knowledge: Knowledge is open if anyone is free to access, use, modify, and share it, promoting a robust commons in which anyone may participate, and interoperability is maximized.

2.2 Linked Open Data (LOD) cloud basics

Linked Open Data (LOD) cloud has emerged as one of the largest collections of interlinked various datasets on the web, covering a broad set of domains from life sciences to media and government data, usually accessed via data portals.

Data portals expose metadata via various models, providing the minimum amount of information that conveys to the inquirer the nature and content of its resources.

- General information (e.g. title, description, ID etc.), required for classification and enhancing dataset discoverability.
- Access' information (e.g. resource name, URL, license title, format, size).
- Ownership information for the dataset (e.g. organization details, maintainer details, author).
- Provenance information (e.g. creation and update dates, version information, version number).

Data portals are datasets access points providing tools to facilitate data publishing, sharing, searching and visualization using Metadata provisioning to attach metadata needed to effectively understand and use datasets.

Linked Data visualization techniques aim to provide graphical representations of some information of interest within a dataset, which is usually in a form readable by machines such as RDF/XML or Turtle (fig. 4).

Resource Description Framework (RDF) is an XML-based language for presenting the information, describing Web resources in the World network and a standard of W3C is semantic network architectures. Information could be different (content description, author, title and other matadata) and is not represented explicitly in the web sites. The graphs are used to define schemes and information.

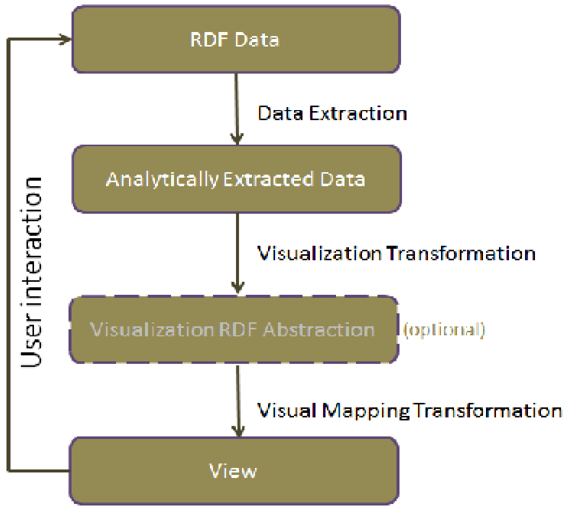
RDF Data
Data Extraction
Analytically Extracted Data
Visualization Transformation
Visualization RDF Abstraction
(optional)
Visual Mapping Transformation
View
User interaction

Fig. 4. Diagram of Linked Data visualization

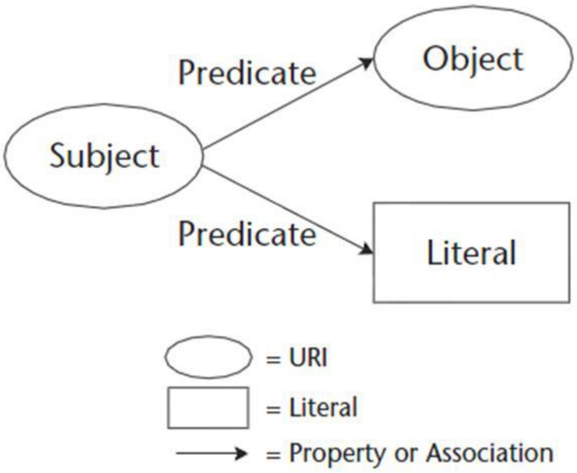

Fig. 5. Structure of the RDF model

The RDF model is used for presenting the data, which belongs to the Open Data space. This model consists of resources and properties, by which the access from one resource to another one is provided. The structure of the RDF model is shown on fig. 5.

The RDF model uses structures called metaphors through which people can model the primitives and the correlation between them easier. Metaphors for the corresponding areas are listed in Table 1.

Table 1. Correspondences between metaphors in different structures

Metaphor	**Part 1**	**Part 2**	**Part 3**
Spoken language	Subject	Predicate	Object
Object orientated	Class	Property	Value
Graph	Node	Peak	Node
Web link	Source	Link	Destination
Data bases	Table	Connection	Table

2.3 Open Data in Science

Data obtained as an observations result and results of scientific activities, accessible to all for analysis and reuse, underlie the formation of global data centers-World Data Centers of the International Council for Science. The presence of the Internet makes it easy to publish and data access. In 2004, research ministers from all Member States of the Organization for Economic Co-operation and Development (OECD) signed statement that all records from data collected with public funds be publicly available. So in 2007 the OECD published "Principles and practices of research data access from publicly funded discoveries" as 'soft law' recommendation.

One of the biggest sources of scientific Open Data are:

- data.uni-muenster.de - containing scientific artifacts from the University of Munster, Germany. Available from 2011;
- linkedscience.org/data - scientific "open" sets of data structured in related data (Linked Open Data). Available from 2011.

2.4 Open Data in Government

Several Governments create web sites which disseminate partially the data they have. More than 200 local, regional and national catalogs Open Data are available from the Project "Data Catalog ". The more – important ones are presented in Table 2.

Table 2.Governments'Open Data websites

Access Portal	Country	Available since
dados.gov.br	Brazil	October 2012
dados.gov.pt	Portugal	October 2012
data.belgium.be	Belgium	Beta version, April 2011
data.gc.ca	Canada	March 2011
data.gouv.fr	France	December 2011
data.gov	USA	May 2009
data.gov.au	Australia	March 2011
data.gov.in	India	November 2012
data.gov.it	Italy	October 2011
data.gov.ma	Marcco	April 2011
data.gov.uk	United Kingdom	September 2009
data.govt.nz	New Zealand	November 2009
data.gv.at	Austria	October 2011
data.norge.no	Norway	April 2010
data.overheid.nl	Netherlands	October 2011
date.gov.md	Moldova	April 2012
daten-deutschland.de	Germany	February 2013
datos.gob.cl	Chile	September 2011
datos.gob.es	Spain	October 2011
datos.gub.uy	Uruguay	November 2011
datosabiertos.gob.go.cr	Costa Rica	October 2012
geodata.gov.gr	Greek	July 2010
opendata.ee	Estonia	September 2012
open-data.europa.eu	European Commission	April 2011
opendata.go.ke	Kenya	July 2011
opengovdata.ru	Russia, private initiative	November 2010
paloalto.opendata.junar.com	Palo Alto	August 2012
rotterdamopendata.nl	Rotterdam	August 2012
satupemerintah.net	Indonesia	November 2012

In addition other government levels create Open Data Portal websites. USA Portal, for instance consist of 31 States, 13 cities and more than 150 agencies and sub-agencies.

2.5. Open Data in Other Fields and Movements

Except already above mentioned fields Open Data takes part in commercial projects such as DBpedia, as a part of structured data in Wikipedia.

There are many Open linked datasets published by private persons or Non-Government Organizations. The most well-known are:

- DBpedia – contains big part of Wikipedia articles, as they are available in English [2];
- FOAF – Friend of a Friend – a dictionary constructed in RDF standard describing people relationships [3];
- GoPubMed – a millions of biomedical publications. A search engine included [4];
- NextBio - scientific experimental data. Includes search engine. Only registered researchers can add data [5].

In addition to Open Data projects, Open Data movements are known:

- Open Access – aims to make scientific articles freely available in Internet;

- Open Content – to put the resources, people guided, in an audience mode, such as prose, movie, pictures etc.;
- Open Knowledge – aligned to Open Data Ideology, contains: scientific, historical, geography information, books, music, films, government and other administrative information;
- Open Notebook Science – scientific data information including failed experiments and raw data;
- Open Source - affect the licenses which may spread computer programs.

3. Assessment of Open Data

This section goal is to present the criteria and determinants of the open data quality, as exhibited some best practices for publication.

The scientific publications examine quality of data and metadata"[6] suitable for use in operations, decision-making and planning" [7]. In other words: "High quality data are accurate, accessible, complete, conformable, reliable, recyclable, appropriate and timely" [7].

The development and standardization of Semantic Web technologies have led to unprecedented volume of data published on the internet as linked Open Data (LOD). Although the collection and publication of such huge amounts of data is certainly a step in the right direction, the data is useful when you have quality attributes. Data sets cover a wide range of areas. However, the data from the Internet showed major differences in quality. For example, data extracted from semi-structured or even unstructured sources often contain non-conformities, as well as false and incomplete information. Already there are many methodologies for assessing data quality, all targeted at different aspects of this task, face a number approaches [8], [9]. However, the quality of data in the data network includes a number of new aspects such as consistency through links to external datasets , presentation of data quality or compliance to indirect information.

3.1 Aspects of assessing open data

Aspects of assessing open data are:

- Technical assessment of datasets;
- Assessment and ranking of Open Government Data (OGD) initiatives;
- Providing quantitative metrics of open data outcomes and impacts;
- Providing qualitative judgements on performance of an open data initiative;
- Developing qualitative case studies about open data use and impacts;

Data quality is usually perceived as a multidimensional structure with the popular definition of "aptitude for use" [10]. Data quality may depend on various factors such as accuracy, timeliness, completeness, relevance, objectivity, reliability, consistency, firmness, availability, and the ability to control [11].

From semantics point of view has different concepts of data quality. Semantic metadata is key concept taken into consideration in quality assessment of datasets [12]. On the other hand the concept of connectivity quality of is another important aspect of the open linked data which is introduced when performes automatic recognition [13].

Bizerte and Kiganyak [14] define problems with the data quality as selecting a web-based information design systems that integrate information from different suppliers. According to Mendez and others [15] the data quality issue is related to their value as a result of the conflict (contradiction) between different data sources.

3.2 Different perspectives for measuring and assessing aspects of open data activities

Perspectives for measuring and assessment aspects of open data activities include:

- Benchmark and compare the use of open data between different countries;
- Compare the use of open data in specific sectors within and between countries;
- General learning from the use of open data in other countries/sectors;
- Support the day to day management of open data initiatives;
- Improve the quality, reliability and quantity of open datasets, and establishing standards and guidelines for the collection and availability of open data;
- Understand the commercial or social impacts of open data, and assess the Return on Investment;
- Prioritize the availability of certain open data;
- Support critical research that can improve policy and practice;
- Support advocacy for more and better open data;

Policy makers, activists, researchers and other users measure and assess open data in aspects of:

- legal and regulatory environment;
- organizational context;
- political will & leadership;
- technical capacity;
- the wider social environment, in terms of civil society and political freedoms;
- the commercial environment and capacity of firms

The data quality assessment methodology is defined as a process that involves measuring the qualitative dimensions that are relevant to the user and comparing the evaluation results of quality requirements. After analyzing the multitude of selected approaches a core set of different quality indicators can be identified and after implementation to evaluate the linked data. The indicators related to each dimension have also been identified and accounted for. The dimensions are shown in fig. 6.

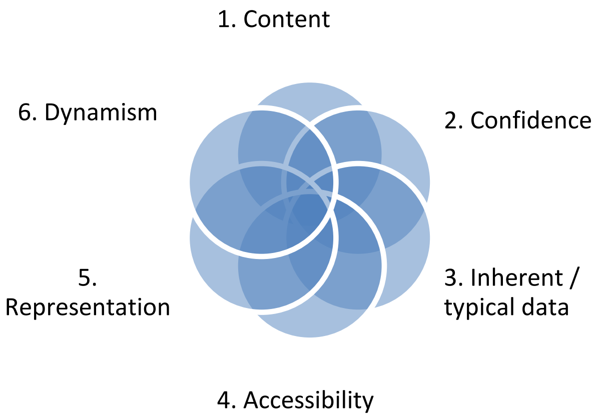

Fig.6. Qualitative dimensions of quality assessment methodology

These groups are not fully independent. They may partially overlap. Also, the dimensions are not independent of each other and there is a correlation between the dimensions of the same group or between groups.

The data amount refers to the amount and data volume connected to a particular task. Data amount provided by a data source affects the usability and should be sufficient to meet the goals and objectives. The amount of data can be measured by bytes or bonds present in the data. This amount should be an appropriate volume of data for a specific task, which is adequate scope and level of detail.

The significance according to Bizerte [16] explains as "the extent to which information is relevant and useful for the task." The applicability relates to the provision of information and in accordance with the specific task and the specific users needs. The applicability is highly dependent on the context and can be measured by means of available meta-data to judge whether the content is relevant to its partial settlement. Moreover, the recovery of the documents can be performed using a combination of hyperlinks analysis methods and data recovery.

3.3 Content

The content is dimension which strongly depends on the task context, as well as subjective preferences of the data sure. There are three criteria: completeness, data amount and significant which are part of this group and together with a detailed list of their respective indices. Completeness refers to the extent present in the defined set of data with all the necessary information. Completeness can be measured by detecting the number of classes, values and relationships that are present in the data to the original data set. It should be noted that in this case, users can gain a circle of assumptions. Especially in linked data the completeness is a prime magnitude to integrate data sets from multiple sources, where one of the goals is to increase completeness.

3.4 Confidence

Confidence is this dimension that focuses on the reliability of the data set. There are five indicators that are part of this group: origin, verifiability, reliability, reputation and licensing which display together with their respective indices. The data integration of two different datasets will occur if both use the same scheme to represent the data. If integration leads to values duplication, this causes extensional tightness. This can be a reason for conflicting values and can be solved by merging duplicate records and merge common properties.

In the scientific literature there are many definitions that emphasize different views to define indicators origin. This chapter considers the origin refers to contextual metadata. It focuses on how to submit, manage and use information on the source origin. The origin helps to assess the credibility of the data and their authenticity and ensure reproducibility. The origin can be measured by analyzing the metadata related to the source. The resulting information can be used to evaluate the reliability, accuracy of the data source.

Verifiability described as "the extent and ease which information can be checked for correctness with" [16]. Verifiability criterion is used as a means by which the user is provided and can be used for examining the data for correctness. Verifiability refers to the correctness degree a data user can assess the data sets and therefore their credibility. Verifiability can be measured by an objective third party, if indicates source data set or the an electronic signature availability.

Verifiability is an important dimension when one set of data sources include low credibility or reputation. This indicator allows data users to decide whether to accept the information provided. One way to check the associated data is to provide basic information about their origin with the dataset using existing dictionaries by SIOC, Dublin Core. Another possibility is the use of electronic signatures.

Reputation is the result of direct experience and/or others recommendations. They offer tracking reputation through a centralized authority or decentralized voting. Reputation is a judgment made by the user to determine the reliability of the source, which is associated primarily with a person, organization, group of people or professional community. Reputation is usually the result of a real value between 0 (low) and 1 (high). There are various options for determining reputation that can be classified in their use. The main method is through study or examination of other members who can help determine the source reputation.

Credibility refers to the fidelity and reliability information degree and can be designated as trusted because it is a subjective measure of consumer attitudes about the data authenticity. Credibility is defined as the degree to which the information is assumed to be true, genuine and reliable and is measured by checking whether the authors of the information are in trusted providers list. In related data reliability can be measured by subjective analysis of the information or the origin of dataset.

In order to enable users to use the data in clear and legal conditions every RDF document should contain a license. Licensing is determined by authorizing users to reuse data set under certain conditions. Licensing may be checked by information relating to this dataset , which clearly states reuse permition.

3.5 Inherent/typical data

Inherent or typical are these data , which are independent of the context. This data dimension focuses on whether the information correctly and accurately represents reality and if the information is logical and consecutive. Accuracy can be defined as a degree to which these data are correct, it means the extent to which it correctly and without error is a fact. In particular, we are talking about semantic precision, which refers to the accuracy of any value to the actual value in the real world, what the accuracy of meaning is.

Objectivity is defined as the degree of objectively data interpretation and use. This indicator is highly dependent on the type of information and therefore is classified as a subjective dimension. Objectivity can not be quantified, but indirectly by authenticity checking of the information source, whether the data are neutral or publisher has a personal impact on the data provided. Furthermore, it can be measured by checking whether the independent sources may confirm a certain fact.

The dimensions accuracy and objectivity guide to the representation of real data correctness as well. Thus, objectivity overlaps with the concept of accuracy, but differs from it, as the concept of precision does not depend on consumer preferences. On the other hand objectivity is affect by the user’s information type. Objectivity is related to the dimension of verifiability. Although the relationship dimension isn’t directly linked to other dimensions, it is included in this group because it is independent of the user.

The documents validity consists of two aspects that affect the documents usability: valid use of basic dictionaries and valid document syntax [16] (syntactic accuracy).

Compatibility means a knowledge base without (logical/formal) contradictions regarding specific examples present mechanisms’ knowledge and conclusions. Tidiness refers to persons and schemes reduction. It can be classified as: A) temporary tightness (scheme level), which refers to attributes reduction and B) extended firmness (data level), which refers to sites reduction.

Example: When a user searches for flights between Paris and New York, instead of flights returning information starting from Paris, France, the search engine returns flights between Paris in Texas and New York. This kind of semantic labeling inaccuracies and classification can lead to wrong results.

3.6 Accessibility

Dimensions belonging to this category include aspects related to the way data is made available. There are four indicators that are part of this group: availability, performance, security and response time. Availability refers to the availability degree of information, or easy and quick information recovery

[17]. On the other hand , the availability is expressed by the regular working of all access methods.

Security can be defined as a data access restriction degree, respectively illegal conduct and misappropriation protection. It refers to transmitted securely information degree between users and information source.

3.7 Representation / representative

Representative dimensions capture aspects related to data design, such as brevity, consistency, comprehensibility and interoperability. The brevity refers to data representation in compact, well-formatted form, clear and complete. Consistency is the information structure and form degree which require to a pre- return information. As the linked data include data aggregation of multiple sources, it allows the definition expansion, which implies not only compatibility with previous data but also other sources data. Comprehensibility refers to the data clearly understanding ease. Interoperability refers to data technical aspects and if the presented information uses appropriate tools to meet the user technical capabilities.

3.8 Dynamism

An important data aspect is their time defectiveness. Basic dimensions related to the dynamic dataset, according to some scientific publications are convertibility defectiveness and accuracy. Convertibility refers to the updating information speed according to the real world changes and needs. Defectiveness refers to the time period which the data remains valid. Accuracy refers to the actually data use time. An important aspect here is the information actuality according to the tangible time period so that it can be applicable and relevant.

4 Conclusion

In this article is: presented a comprehensive review of established methodologies for assess the data quality; marked the differences between these approaches in different dimensions; derived qualitative dimensions of quality assessment methods; linked different approaches with relevant indicators; defined tools appropriate for each approach, classified by data type, automation ranking and required level of potential users' operational skills.

It is proposed a multistep integrated approach for data quality assessment, which includes:

1. Analysis of requirements
2. Quality of checklist data
3. Statistics and analysis of low level
4. Aggregate and higher indicators of level
5. Comparison
6. Interpretation.

The findings made by the Open Data Assessment are essential in choosing the approbation technology for dynamic reference model of the Bulgarian electronic government [18].

References:

[1] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann and Soren Auer Quality Assessment Methodologies for Linked Open Data, *Semantic Web,* 1, 2012, 1–5 1 IOS Pres

[2] DBpedia available at: http://wiki.dbpedia.org/

[3] Friend of a Friend available at: http://www.foaf-project.org/

[4] GoPubMed available at: http://www.gopubmed.com/web/gopubmed/www/GoPubMed/Search/index.

[5] Illumine Nextbio Research available at: https://www.nextbio.com/b/authentication/login.nb

[6] ANSI American National Standards Institute (2004) Understanding Metadata (available at: http://www.niso.org/publications/press/UnderstandingMetadata.pdf)

[7] Juran, Joseph M. and A. Blanton Godfrey, *Juran's Quality Handbook,* Fifth Edition, McGraw-Hill, 1999.

[8] Pipino, L., Lee, Y. W., AND Wang, R. Y. Data quality assessment. *Communications of the ACM 45, 4* , 2002.

[9] BATINI, C., AND SCANNAPIECO, M. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[10] Juran, J. *The Quality Control Handbook,* McGraw-Hill, New York, 1974.

[11] Wang, R. Y., and Strong, D. M. Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems,* 1996, 12, 4, pp. 5-33.

[12] Lei, Y., Nikolov, A., Uren, V., and Motta, E. Detecting quality problems in semantic metadata without the presence of a gold standard. *In Workshop on "Evaluation of*

Ontologies for the Web " (EON) at the WWW, 2007, pp. 51-60.

[13] Gueret, C., Groth, P., Stadler, C., and Lehmann, J. Assessing linked data mappings using network measures. *In ESWC,* 2012.

[14] Bizer, C., and Cyganiak, R. Quality-driven information filtering using the wiqa policy framework. *Web Semantics,* 2009, 7, 1-10.

[15] Mendes, P., Muhleisen, H., and bizer, C. Sieve: Linked data quality assessment and fusion. In LWDM, 2012

[16] Bizer, C. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. PhD thesis, Freie Universität Berlin, 2007

[17] Flemming, A. Quality characteristics of linked data publishing datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010.

[18] Roumen Trifonov, Radoslav Yoshinov, Some Security Issues of the Governmental Cloud, *15th International Conference on ACE'16*, Mallorca, Spain, August 19-21, 2016, ISBN: 978-1-61804-327-6