Meeting Open Data Halfway: on Semi-Open Data Paradigm

Mortaza S. Bargh
Research and Documentation Centre
Ministry of Justice and Security
The Hague
The Netherlands
m.shoae.bargh@minvenj.nl

Sunil Choenni
Research and Documentation Centre[1]
Rotterdam Uni. Of Applied Sciences[2]
[1]The Hague, [2]Rotterdam
The Netherlands
r.choenni@{[1]minvenj.nl, [2]hr.nl}

Ronald Meijer
Research and Documentation Centre
Ministry of Justice and Security
The Hague
The Netherlands
r.f.meijer@minvenj.nl

ABSTRACT
Some organizations have faced serious obstacles for disseminating their data according to the Open Data requirements and characteristics, e.g., for everybody and any use. Often this is the case when the data is of low quality, has (potentially) sensitive information, or has non-interoperable data format and semantics. Not being able to (completely) satisfy the Open Data requirements may have made such organizations to appear incompliant with the ideals and objectives of Open Data, despite their full commitments and efforts for data opening. In this contribution we propose a new paradigm – called Semi-Open Data paradigm – in order to frame, acknowledge, and encourage such initiatives and efforts that strive along the Open Data objectives but do not comply with Open Data requirements completely due to some practical constraints. For the proposed Semi-Open Data paradigm we further present an assessment method to measure and categorize Semi-Open Data initiatives objectively. This method offers a better way to assess and reward the extent of organizations' efforts to meet the Open Data characteristics than the current method that checks whether all Open Data requirements are met or not (i.e., by making a binary decision).

CCS Concepts
• **Social and professional topics~Governmental regulations** • Applied computing~E-government

Keywords
Measurement Method; Open Data; Open Data Impediments; Open Data Objectives; Semi-open Data Paradigm

1. INTRODUCTION
Open Data has been an initiative for scientific and governmental institutions to, among others, gain the public trust, achieve transparency, stimulate innovations and deliver economic growth. Nevertheless some organizations have faced serious obstacles and impediments for disseminating their data according to Open Data requirements and characteristics, for example, for everybody, in a timely way, with primacy and permanence, as raw as possible, with appropriate metadata, complete, without costs, license free and reusable, interoperable, and machine readable.

Not being able to (completely) fulfill the Open Data requirements may have made such organizations to appear incompliant with the ideals and objectives of Open Data, i.e., being resistant against transparency, innovations, and economic growth, in the current era of information super highway, even though these organizations are (fully) committed to and put lots of efforts in (the process of) data opening. This negative image can be costly for public organizations as they do not gain or even may lose the trust of the public and may relinquish the benefits of well-informed societies and citizens.

Opening information is potentially subject to many challenges and dangers like privacy breaches, data misinterpretations, and misleading of data consumers and the public [14][20]. These challenges, in turn, may inflict various negative consequences and costs on individuals, the public order and the society. Therefore, organizations in possession of sensitive and private information are hesitant to open such data and information. In practice, however, there are evidences that a large number of such organizations disseminate their data to some degrees such that the disseminated data cannot be strictly classified as Open Data. This is because they cannot comply with all requirements of Open Data. For example, the disclosed data is aggregated information, not offered via publicly accessible Internet portals, and/or offered to specific groups of data consumers. Although such data dissemination efforts play an undeniable role in achieving (part of) the aforementioned objectives of transparency, innovation stimulation, economic growth, etc.; they are not regarded as Open Data initiatives and therefore do not contribute to the Open Data image of such organizations.

In this contribution we aim at finding a way to frame, acknowledge and count the efforts and initiatives that strive along Open Data objectives but do not comply with Open Data requirements completely due to some practical constraints and reasons. Specifically the main research question we address is: *How can we frame the data dissemination efforts of those organizations whose data dissemination efforts do not adhere to all Open Data requirements?* Hereto we shall elucidate the impediments of data opening in some public institutions and propose a new data dissemination paradigm of Semi-Open Data. Finally we shall sketch the principles of an assessment mechanism to measure and categorize Semi-Open Data initiatives objectively.

This contribution actually provides a vision for acknowledging of and giving credits to those data dissemination efforts that cannot be categorize as Open Data initiatives but *partially* serve the Open Data objectives. This is an explorative study that reflects upon the experiences of the authors and others in opening the criminal and judicial data of a judicial research center. To this end, additionally, we have carried out a literature study to elucidate the Open Data objectives, definitions and characteristics as proposed and used by various (governmental) organizations.

The organization of the paper is as follows. Section 2 provides an overview of the current Open Data vision, including its definition, objectives, and characteristics. Section 3 reflects on the outcomes of Open Data initiatives in Dutch governmental organizations.

ICEGOV '15-16, March 01-03, 2016, Montevideo, Uruguay
© 2016 ACM. ISBN 978-1-4503-3640-6/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2910019.2910037

Section 4 presents our motivations and vision for Semi-Open Data paradigm and Section 5 embeds the proposed vision in the related work. Finally Section 6 presents our conclusions and elaborates on some future research directions.

2. OPEN DATA

This section serves as the related work on the history and the current concept of Open Data. In addition to laying down a common ground, this section together with the subsequent section presents those concepts that we will use to substantiate our Semi-Open Data paradigm in Section 4.

2.1 Definition

Open Data refers to the data that can be freely used, re-used and redistributed by everyone for any purpose. Specifically, according to the Open Definition [5], Open Data is defined as: "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)" [5]. The Open Definition explains the principles of its definition as follows:

- "Availability and Access: the data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
- Reuse and Redistribution: the data must be provided under terms that permit reuse and redistribution, including the intermixing with other datasets.
- Universal Participation: everyone must be able to use, reuse and redistribute – there should be no discrimination against fields of endeavor or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed" [6].

2.2 Objectives

Governments and governmental organizations, individuals (e.g., citizens, journalists, and scientists), and businesses (e.g., companies and enterprises) have embraced the idea of Open Data in order to achieve various objectives as enlisted below.

- Governments are interested in Open Data to provide transparency into government operations, meet regulatory compliances, increase public participation and collaboration (thus strengthening democracy), anticipate on economic developments, and initiate innovations.
- Individuals can make better personal decisions (e.g., they can use neighborhood statistics in buying houses and implanting trees and plants in polluted areas), journalists can obtain evidences to support their journalistic articles, and scientists can use the data for reviewing scientific results and for deriving new insights and knowledge.
- Businesses, enterprises, and entrepreneurs can use the data for their innovations (e.g., to creative recommender apps based on weather conditions) and for supporting their strategic business decisions.

These Open Data objectives are genuine idealities that should aspire and drive the policies, services, products, and activities of individuals, governments, and businesses. Every attempt to achieve (some of) these objectives must be recognized, acknowledged, and encouraged including our here-proposed Semi-Open Data set of solutions or paradigm. Due to importance of these objectives in demonstrating the relevancy of our vision on Semi-Open Data paradigm, we enlist the objectives of Open Data for future reference below (mainly selected and summarized from [3][25][7]).

- Improving *transparency*: Access to data by citizens enables them to monitor the deeds of a democratic government.
- Increasing *accountability and compliance*: Access to data by the public and taxpayers increases governments' accountability and democratic reforms.
- Supporting *participatory governance*: Access to data by citizens enables them to engage (i.e., to influence and contribute proactively) in the process of governance.
- Supporting *innovation*: Access to data by private sector can lead to creating new services and products.
- Improving *efficiency and cost-effectiveness*: Access to (scientific) data for public interest purposes by businesses can improve existing commercial applications and products. Proactive disclosure of information can also reduce the burden on governments for handling of data requests.
- Supporting *informed decisions*: Access to data by consumers, policymakers and businesses can teach them about and help them with their daily and strategic decisions.
- Supporting for *research*: Access to data by scientists can support them with evidence-based primary research.

2.3 Characteristics

In order to achieve the objectives of Open Data there are a number of requirements defined for data opening by various governments and organizations. These requirements determine the characteristics of Open Data and can be divided in five categories of data access and usage control, ways of data access, data formats, data contents, and post release supports. In the following we enlist these requirements from four sources (Canadian government [2][28], US government [12], Dutch government [BZK: an internal draft report], and organization Open Data Research Network (ODRN) from South Africa [3]).

From the viewpoint of *data access and usage control*, the data should be for the public in being accessible for everybody, without any discrimination, and without any need for identification or usage justification. This access is, of course, restricted to the extent that is permitted by law (i.e., subject to privacy, confidentiality, security, or other valid restrictions). The released data should be license free and reusable for data recipients in order to enable (commercial) innovations and (commercial) reuse of the data.

From the viewpoint of the *ways of data access*, ease of access is a common criterion that requires the data access to be without any registration of data recipients and without any technological restrictions (e.g., requiring browser-oriented technologies). Data should be accessible via an API (Application Programming Interface) for automatic data processing and be accessible via a web portal (e.g., in the Netherlands via www.Opendata.nl). The data access should be delivered free of charge (or, according to some, with a low expense covering cost).

From the viewpoint of *data format*, the data should be machine-readable and process-able, which requires using common file formats suitable for machine processing, e.g., CSV (Comma-Separated Values) and XML (Extensible Markup Language), and standard data formats for data storage and processing.

Concerning the *data contents and types*, one should disseminate data

- Timely (i.e., as quickly as possible),
- As is (i.e., as raw as possible),
- In a complete/bulk form (i.e., without removing parts of it, unless there are privacy issues or information sensitivity),
- In its primacy (i.e., the data is published by its primary sources) to enable a proper control of data collection and storage,
- With metadata and sufficient information to enable understanding the strengths, weaknesses, analytical limitations, security and privacy requirements, ways of data processing, etc. (for example, with additional descriptions of the purpose of data collection, the population of interest, the characteristics of the sample, and the method of data collection), and
- With (persistent) URIs (Unified Resource Identifiers) to enable locating data objects.

From the viewpoint of *post release support* and data management, permanence requires that the data remain online with appropriate version tracking, archiving, and history logging over time.

Table 2 provides an overview of the abovementioned characteristics, and per every characteristic the table indicates which source (related to a government/organization) includes the characteristic in its list of Open Data characteristics/requirements. According to [12], derived or aggregate data may also be considered as Open Data if data dissemination includes also the corresponding primary data. It is worthwhile to mention that sensitive information falls out scope of Open Data in all Open Data lists that we have investigated, see the last row in Table 1. In the Netherlands, for example, the exceptions of Open Data include those datasets that contain privacy sensitive data, national security data, and business sensitive data.

Table 1: Characteristics of Open Data in four countries.

Open Data characteristics	From [2]	From [12]	From [BZK]	From [3]
For the public				
License free and reusable				
Ease of access				
Accessible via a web portal				
Free of charge				
Machine-readable				
Standard data formats				
Timely				
As it is				
Complete				
Primacy				
With metadata				
With URLs				
Permanence				
No sensitive information				

3. REFLECTION ON OPEN DATA

This section presents some statistics on Open Data initiatives and an overview of the obstacles on opening of data.

3.1 Success Cases

According to report [1], recently published by the Court of Audit of the Netherlands [4], there have been about 3200 datasets openly accessible from web portal of the Dutch government (i.e., www.Opendata.nl) as of 27 March 2015. Almost half of these datasets belong to various departments of Dutch central government and the rest of the datasets largely belong to Dutch regional governments (like municipalities and provinces). For the former group of datasets, more specifically those datasets made open by ministries of Dutch central government, Table 2 summarizes the number of datasets made open per ministry from [1]. The forerunner in opening datasets is the Ministry of Infrastructure and Environment by far, followed by the Ministry of Economics affairs. Most of the released data is concerned with geo-data [1].

Table 2: Number of Open Data datasets per ministry of The Netherlands, copied (and translated) from [1].

Dutch ministries	**Total number**
Ministry of Infrastructure and Environment	1225
Ministry of Economic Affairs	177
Divers	118
Ministry of Education, Culture and Science	96
Ministry of the Interior and Kingdom Relations	25
Ministry of Health, Welfare and Sport	18
Ministry of General Affairs	8
Ministry of Finance	5
Ministry of Defense	4
Ministry of Foreign Affairs / Ministry of Foreign Trade and Development	2
Ministry of Social Affairs and Employment	0
Ministry of Security and Justice	0

As indicated in Table 2 Ministry of Social Affairs and Employment and Ministry of Security and Justice have not opened any dataset via the official portal data.overheid.nl. One declaration, mentioned in [1], is that the data dissemination channels of these ministries are different from data.overheid.nl. We believe, however, that also (and mainly) information privacy and sensitivity account for such a low number of opened datasets.

3.2 Obstacles

Experience shows that there are a number of challenges and obstacles for publishing and using open datasets. As a result, some organizations have been unable yet to adhere to Open Data requirements and disseminate their datasets as expected, see for example the statistics of some Dutch ministries in bottom rows of Table 2.

In report [1] the Court of Audit of The Netherlands mentions a number of reasons behind the abovementioned low (or null) number of opened datasets. We categorize here these rationales (i.e., obstacles/ impediments on the way of Open Data) from [1].

High costs (and no infrastructure): The opening of the data requires an (initial) investment and/or there is no infrastructure to adequately mange the disseminated data. For example, data controllers have mentioned arguments like: Not knowing where the data is (i.e., need to discover data sources), having no capacity to manage data, being infeasible for our website to handle the data opening process, being technically infeasible according to our IT staff/supplier who says it cannot be done (technically), and asking too much money by IT suppliers.

Low added value: Some use the cost-benefit argument and undermine and doubt on the added value of data opening. For example, data controllers have mentioned: What the business case is, what the end/limit of data opening is, whether it will be interesting/useful for someone, it has not been done before, or there is no time and means for it.

Information sensitivity and authorization: This is a common argument used by data controllers. For example, the controllers perceive their data confidential, commercially valuable, and (potentially) privacy sensitive. They have also argued that: They cannot acknowledge/deny collecting the information, their bosses are opposed, the data is owned by or under control of another (unknown) party (who does not allow opening the data), they cannot accept the responsibility of data reuse, they are unsure whether laws and regulations allow sharing of data, and they have no authority to release it.

Uncertainty about data usage: After releasing the data it is difficult, if not to say impossible, to make sure that the data is used in accordance with the terms and conditions that the data is collected and opened. For example, data controllers have mentioned rationales like: They suspect that people will use the data wrongly, misuse the data, become overloaded, derive wrong conclusions, link the data with other data leading to privacy and confidentiality breaches; think that only they understand the data (due to, e.g., having the domain knowledge to interpret the data appropriately); suspect that data opening would lead to useless discussions, angering people, disturbing the market, etc.; and are worried about the backslashes of data opening, e.g., receiving unjustifiable opinion and view, on the organization.

Low quality data: The data in possession of the organization is imperfect and has low quality. For example, data controllers have mentioned: The data is not in digital form, not in useful form, not errorless, incomplete, too old, too detailed, low quality (we think), and too big. They were also concerned that if the data is downloaded and used later, it becomes too old and low quality.

Too restrictive criteria: Finally some arguments refer to the fact that data opening is possible but not according to (current) requirements of Open Data. For example, data controllers have argued that: They can open their data with some modification/ adaption (e.g., with 90% edition), their data can be found (e.g., it is online) but cannot be published, or is in a PDF (Portable Document Format) format.

Note that our intention for the categorization above is to group related rationales meaningfully/semantically and the categories are not meant to be orthogonal/independent. In fact one can correlate, for example, high costs and low added value arguments. Interested readers are referred to [15][16][20][29] that have also enlisted a number of similar impediments on opening of data by (governmental) organizations.

In order to achieve the objectives of Open Data, we can adopt various solution directions like: take strategic and managerial decisions and policies to dedicate enough (financial) resources, change the culture and attitude within organizations, deploy information sensitivity and privacy protection techniques; adopt audit, accountability and governance measures and procedures, and realize data usage monitoring and data usage control mechanisms. In this paper we argue that embracing a Semi-Open Data paradigm can be a solution direction that directly provides a remedy for reducing the mentioned obstacles, while helping to achieve some of (and move towards) the objectives of Open Data.

4. PROPOSED PARADIGM

For some organizations and their datasets it is hardly possible to quickly meet all conditions of Open Data completely and undeviatingly. Nevertheless, in practice, these organizations initiate and take pragmatic steps towards the ideals of Open Data. These initiatives are currently not considered as Open Data, which makes the efforts behind these initiatives unnoticed. This overlooking (or perhaps ignorance) is discouraging and demotivating for these initiatives unfortunately. In this section we argue to give enough credit to these initiatives and do not adopt a binary (black and white) stand on the matter.

4.1 Motivations

When the data of an organization is of low quality (due to having inconsistent, imprecise, uncertain, missing, and incomplete data objects), has private or business sensitive information potentially (e.g., when it is combined with other data), or has proprietary/ unstandardized/non-interoperable data format and semantics, it becomes difficult, if not to say impossible, for the organization to open the data as required by the characteristics of Open Data mentioned in Table 1.

On the one hand, opening the data as it is, to the public, can lead to various problems such as:

- Privacy and sensitive (business) information breaches, due to revealing (potentially) privacy sensitive information. This violates the basic principle of Open Data, namely: not to open privacy and sensitive (business) information!
- Misinterpretation and misleading; which violates the objectives of, for example, transparency and decision support.
- No or low economic growth, due to making it hard to link or integrate the data.

On the other hand, one can argue that let's make investment in such data by enhancing its quality, reducing its sensitivity, and improving its format to some acceptable levels. Hereto one needs also to create the appropriate metadata. This option, however, requires complex operations, inflicts extra costs on the organization, and/or makes only the 'processed' data open. These measures quite often violate the cost reduction purpose (and perhaps undermine the opening raw data characteristic of Open Data). In conclusion, one cannot prescribe the Open Data remedy for such low quality, sensitive, or legacy data.

In order to motivate the introduction and embracement of Semi-Open Data paradigm for opening low quality, sensitive, or legacy data we enlist three metaphoric solution scenarios to deal with the mentioned Open Data obstacles.

1. Do not open the data. This is the current situation for some organizations as sketched in Section 3.
2. Open the data as it is and with minimum efforts (e.g., by applying basic data anonymization) to a group of (entrusted) independent experts who have domain knowledge and can interpret/use data properly (like scientists). This is a relatively low cost solution (assuming that the basic infrastructure is in place), where misinterpretation and privacy issues are covered. Thus it eliminates the high costs, information sensitivity, and data usage uncertainty concerns and obstacles.
3. Improve the data quality, information sensitivity, and data format to an acceptable level, share the resulting data with the public, and ask for a grant/budget or a minor fee from data

recipients to cover the expenses. This solution eliminates most concerns/obstacles of Open Data, namely: high costs, low quality data, data usage uncertainty, and information sensitivity.

Obviously the first solution is insensible from the viewpoint of achieving the objective of Open Data. The second and third solutions do not comply with Open Data requirements and characteristics, in the sense that, for example, the data is not made open for everybody, as raw as possible, or free of charge. Nevertheless, the second and third solutions serve, to some degree, the objectives of Open Data, particularly transparency and compliance because independent domain experts (in the second solution) and the public (in the third solution) can learn about the organization by using the (high quality) data and can examine the organization's adherence to laws and regulations. Or highly qualified experts and entrepreneurs can use the released data to make innovative services and products, leading to economic growth. Therefore, we introduce the concept of Semi-Open Data to mark and recognize those initiatives that satisfy the characteristics of Open Data partially, while they are aligned with or serve the Open Data objectives.

4.2 Vision

One needs to acknowledge those solutions and initiatives that push the frontiers of information sharing towards the objectives and ideals of Open Data. In this way, not only do we encourage and promote such initiatives, but also we obtain a more realistic view on the landscape of Open Data ideals. Therefore we have coined the umbrella concept of Semi-Open Data to mark those solutions that do not adhere to all requirements of Open Data but surely help us to achieve some objectives of Open Data. Semi-Open Data paradigm include those data sharing solutions that aim at Open Data Objectives (like transparency, compliance, innovation, decision support, cost reduction, participation, and collaboration) but do not fulfill all conditions of Open Data as outlined in Subsection 2.3.

Instead of making a binary decision whether or not a data sharing initiative fulfills all conditions of Open Data, we advise to adopt a multi-dimensional multi-level measurement framework to measure and quantify those initiatives that push the frontiers towards Open Data ideals. To this end, we assume that the requirements outlined in Subsection 2.3 define the dimensions along which one can measure data sharing initiatives. For example along dimension 'for the public', one can define a number of ordinal levels starting from 'share with no one' to 'share with the public', corresponding to closed (or confidential) data and Open Data settings, respectively. Example intermittent levels could be: 'share data within a specific group', 'share data within a department of an organization', 'share data within an organization /ministry', and 'share data among a federation of organizations'. For the standard data format aspect (related to data interoperability and link-ability), one can think of the following levels: 'without any specific data format', 'with a data format of acceptable convertibility' (applicable within data space environments [17]), 'with a data format of high/precise convertibility' (applicable within data warehouse environments [18]), and 'with a standardized data format' (applicable within a database management system). Similarly one can define multiple ordinal levels per every dimension of 'free of charge', 'license free', 'ease of access', etc.

In practice, these dimensions and their ordinal levels need to be standardized locally within a setting or, more preferably, globally. Every data sharing initiative can then be mapped to a point in the multi-dimensional space. Figure 1 illustrates the concept by showing three data sharing initiatives as three points, denoted by Ini_1, Ini_2 and Ini_3, in a two dimensional space for measuring data opening levels.

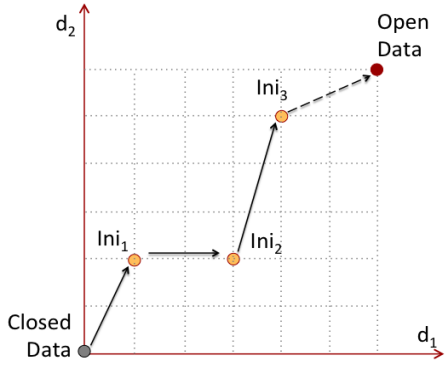

Figure 1: An illustration of a two dimensional space for measuring data opening initiatives.

By choosing appropriate interval units along every dimension, assigning appropriate number of units to the ordinal levels, and adopting an appropriate (distance) measure (like Euclidean distance in a Cartesian space (or similarly to [13][23] by using the multi-attribute utility theory) one can measure the distance of every data opening initiative to the Open Data point. Also one can monitor and learn from the progresses made in consecutive data opening initiatives (e.g., see the illustration shown by the arrows among initiatives Ini_1, Ini_2 and Ini_3 and their relative distances to the ideal point in every step). In our opinion *this is a better way to assess the extent of organizations' efforts to meet Open Data objectives than the current binary decision method*.

We note that Semi-Open Data paradigm basically advocates using a multilevel and multidimensional assessment mechanism instead of a binary one to measure the level of adherence to Open Data ideals. As suggested in Figure 1, organizations should continuously seek out the ways to approach the ideal Open Data point (by taking, for example, small steps in strategically chosen directions toward the ideal Open Data point). Care should be taken on, as also mentioned by one of the anonymous reviewers of this paper, Semi-Open Data paradigm only being seen as a means of providing "a perverse incentive for organizations to reduce their compliance with the full set of Open Data principles". To this end, one approach would be to perceive data opening as a continuous process to resolve an untamed or wicked problem [11].

4.3 Example Initiatives

In the following we provide two typical existing examples of Semi-Open Data initiatives. Our research center – i.e., the Research and Documentation Centre (abbreviated as WODC in Dutch) of the Ministry of Security and Justice of The Netherlands – systematically collects, stores and enhances the Dutch judicial information obtained from its internal and external partner organizations. The research center makes various reports and datasets freely accessible, reusable and redistributable. One of these available datasets is from the Dutch Recidivism Monitor project [26]. The project is a long-term research project within the center to conduct standardized measurements of recidivism amongst several groups of offenders. The measurements are based on the anonymized data from the Dutch Research and Policy Database for Judicial Documentation (abbreviated as OBJD in Dutch). This input data contains the current and historical penal documentation

of any natural and legal persons who came into contact with the judicial system [26].

The processed output data of the Dutch Recidivism Monitor project is made open via a web portal called Recidivism Prevalence Information System (REPRIS), which is an interactive web application publicly accessible through the WODC website [8]. The web interface allows users to select various parameters (like the offender group(s), the observation period(s), recidivism type(s), gender group, and age group) with dropdown lists. Based on the choices made, REPRIS returns the answer in the form of a table or graph with aggregated data. The data in the table can also be exported in CSV format in Excel. The recidivism data from REPRIS is accessible for everyone.

As such, REPRIS is a typical Semi-Open Data initiative that provides the public an access to 'processed' data. Another similar example is the electronic databank of Statistics Netherlands (so-called StatLine [9]). StatLine offers for all users a free of charge access to processed data by compiling their own tables and graphs.

5. RELATED WORK

In this section we review those publication records and works that we have found in our literature study on some concepts similar to the Semi-Open Data paradigm. Furthermore, we position and embed the envisioned paradigm among those existing ones found in the surveyed literature.

5.1 Similar Ideas

In weblog [27] the "cases where we only get partial information" are positioned in a class of issues called Semi-Open Data. According to [27], examples of Semi-Open data include the use of PDFs for data releases, providing maps without underlying data, or providing only tiny slices of stale data. The author gives an example about how incomplete Open Data causes misinterpretation of data. The example is about a report on comparing measles vaccination rates in the public and private schools in New York and the New York state. "So with Semi-Open Data, we are unable to make a fair comparison" [27]. As explained in Sections 2 and 3, Open Data requires disseminating complete data to have good data utility and in order not to lead to misinterpretations or wrong conclusions. We share the same viewpoint. Nevertheless, we don't believe that the point raised is inherent for Semi-Open Data (i.e., it could have happened even if the disseminated data was complete). Also making incomplete data open to the public to make wrong conclusions – as [27] added: "But that did not stop the media from trying. It seems that New York Magazine just went ahead and made the comparison any way" – is not a case of Semi-Open Data as we defined. According to our view on Semi-Open Data paradigm, incomplete or low quality data should be accompanied with sufficient metadata, interpreted by (domain) experts, and/or preprocessed to reach a higher quality, in order to lead to satisfactory outcomes/insights.

Considering a number of recent fraudulent practices in Dutch universities, a fact finding committee of The Royal Netherlands Academy of Arts and Sciences urged to raise the awareness of the importance of careful and proper handling of research data. This proper handling requires that scientific research outputs (the publications and the underlying data) should be open to validate research results, prevent fraudulent research, and facilitate data reuse [5]. This means opening scientific data to attain transparency, accountability, and innovations. Consequently, the School of Business and Economics of Maastricht University has considered a form of Semi-Open Data to regard research data as the most valuable asset for researchers. The "school … is thinking about a Semi-Open Data policy: raw, self-generated data are stored at the faculty level with restricted access and a careful description of the generation process, and made available for 10–15 years after generation (in case of problems). Manipulated data should be stored at a central university level, also well documented and in principle available for colleagues to use" [19]. Similar requirements, i.e., publishing the information relevant for reproducibility of experiments, are proposed in [22], and are already adopted by some journals and conferences (like Nature, Perspectives on Psychological Science developed, and conferences such as the ACM (Association for Computing Machinery) Special Interest Group on Management of Data [10]).

The way that Open Data is defined is important because it determines the datasets that can be classified as open [3]. The Open Data in Developing Countries (ODDC) [3], therefore, proposes a broader definition of open data where also the impact and context of Open Data are taken into consideration ("Open Data = Open definition + (impact/context)" [3]). For example, data granularity and timeliness can be dependent of whether the data is used for improving transparency/accountability or for creating innovative and economic impacts. Making Open Data definition dependent of impact/context is not backed by others to the best of our knowledge. Maybe this is because such a dependency complicates the matter of Open Data. We, nevertheless, acknowledge having this dependency, and believe that it can be better manifested according to the Semi-Open Data paradigm.

5.2 Embedding in Existing Paradigms

In [24] Swire investigates the rising and important question of optimum openness in security settings. He attempts to characterize the dueling approaches of disclosure and secrecy from the perspective of security. The investigation is based on answering the questions: Whether disclosing information helps or harms three types of actors: the attacker, the designer or the defender of a system. The answers to these inquiries have resulted in identification of four basic paradigms in [24], namely: the Open Source paradigm, the Military/Intelligence paradigm, the Information Sharing paradigm, and the Public Domain paradigm.

In the *Open Source* paradigm disclosure helps security (i.e., slogan of "there is no security through obscurity"). This paradigm is based on the following (implicit) assumptions [24]: the attackers will learn little or nothing from information disclosure (e.g., when everybody knows about the vulnerability of a common software, malicious attackers cannot gain significantly from knowing about the vulnerability), the system designers do learn from information disclosure (e.g., open source developers can try to fix the vulnerability), and system defenders can benefit from the disclosure (e.g., they can take protective actions).

In the *Military/Intelligence* paradigm, secrecy helps security (i.e., slogan of "loose lips sink ships"). This paradigm is based on the following (implicit) assumptions [24]: the attackers will learn a lot from information disclosure (e.g., publishing about software vulnerability of a (/an expensive and hard to modify) military defense system helps malicious attackers significantly), the system designers don't learn from information disclosure (because likely there are no helpful experts with domain expertise out there), and system defenders cannot benefit from the disclosure (because there are no other system defenders outside).

In the *Information Sharing* paradigm, some organizations (e.g., members of a federation) share information among themselves to improve security. This paradigm is based on the following

(implicit) assumptions [24]: the attackers shall learn a lot from information disclosure (e.g., publishing about a software vulnerability of a banking system helps malicious attackers significantly), the system designers learn from information disclosure (as those system developers, who are associated with an organization of the federation, can learn about the arising vulnerabilities at other organizations), and system defenders can benefit from the disclosure (e.g., they can take protective actions).

In the *Public Domain* paradigm, there are minor costs and benefits from information disclosure, as either the public already knows about it or the information concerns security vulnerabilities of minor impacts. This paradigm is based on the following (implicit) assumptions [24]: the attackers, system designers, and defenders shall learn little or nothing new from information disclosure (e.g., publishing about software vulnerability/bug of a game app that is related to the game's visual effects, and is known by everybody). Table 3 summarizes the benefits of information disclosure for the three actors in the four paradigms identified in [24].

Table 3: Benefits of knowing information for the three actors in 4 paradigms of [24], where L (Low) and H (High).

Actors vs. paradigms	Military	Info sharing	Open source	Public
Attackers	H	H	L	L
Designers	L	H	H	L
Defenders	L	H	H	L

An informed judgment about disclosing information can take into account other objectives than security, like privacy and accountability [24]. Similarly, within the context of Open Data and the current work, we observe that:

- The objectives of Open Data like transparency, compliance and accountability, innovation, participation and collaboration, and cost reduction, as mentioned in Section 2.1, should be considered instead of the security objective considered in [24],
- The same paradigms of [24] (i.e., Open Source, Military/ Intelligence, Information Sharing, and Public Domain) are applicable, prevalent and valid (except that, following the arguments of [21], we use the term 'Open Data' instead of the term 'Open Source' within our context),
- Criminals and data misusers, policymakers, and information users and managers can fulfill the roles of the attacker, the system designer and the defender in [24].

Based on these observations we adapt Table 3 for the Open Data context and present the result in Table 4.

Table 4: Introducing and positioning the semi-open data paradigm among those paradigms of [24], based on the benefits of knowing information for the three actors.

Actors vs. paradigms	Military	Info sharing	Semi-Open Data	Open Data	Public
Criminals and data misusers	H	H	L/H	L	L
Policymakers	L	H	H	H	L
Information managers/users	L	H	H	H	L

As can be seen in Table 4, the Semi-Open Data paradigm is positioned between the Information Sharing paradigm and the Open Data paradigms. The disclosed information is of high value for policymakers and other information users and managers who can make use of the disclosed information to make better public policies, make better decisions, innovations, and compliance audit and control. The criminals and data misusers can potentially misuse the disclosed information, but Semi-Open Data solutions, as we envision, should minimize such threats. This can be achieved by sharing data with a trusted group, with experts, in a preprocessed form, etc. This dual character of the disclosed data in the Semi-Open Data paradigm is, therefore, marked by L/H (Low/High) values, depending on whether an effective Semi-Open Data solution is applied to the disclosed data or not, respectively.

Another reason that we envision the Semi-Open Data between the Information Sharing and Open Data paradigms is our own experience within the WODC to disclose judicial information. The WODC typically operates according to this Semi-Open Data paradigm, where it:

- Obtains the required raw data from partner organizations, based on the Information Sharing paradigm,
- Processes the data to produce aggregated, reliable, valid or high quality information, and
- Shares (most of) the resulting data with the public, in according to the Open Data paradigm.

As illustrated in Figure 1, we actually regard the Semi-Open Data as a movement from the Information Sharing paradigm to the ideal of the Open Data paradigm, given the contending objectives and limitations on the disclosed data.

6. CONCLUSION

Based on the objectives and characteristics of Open Data we argued that there is a need to recognize a new class of data disclosure initiatives, coined as Semi-Open Data initiatives in this contribution. These initiatives aim at achieving and delivering (some of) the objectives of Open Data, but they cannot satisfy all requirements of Open Data. Therefore, these initiatives are not recognized and appreciated enough, which could be discouraging and disappointing for those organizations behind such initiatives.

The proposed Semi-Open Data paradigm embraces those data disclosing solutions that satisfy some requirements of Open Data often because the disclosed data has low quality, is potentially (privacy) sensitive, or has proprietary/unstandardized/non-interoperable data format and semantics. Those solutions that fall within the scope of Semi-Open Data paradigm allow disseminating the data in a restricted form and scope to achieve (some of) the objectives mentioned. We positioned the proposed Semi-Open Data paradigm between the existing Information Sharing and Open Data paradigms. Solutions within the Semi-Open Data paradigm can actually be regarded as attempts to move towards the Open Data paradigm in order to achieve all Open Data ideals.

Finally we proposed a method to measure the level of adherence of Semi-Open Data initiatives to Open Data characteristics and requirements. This method offers a better way to assess and reward the extent of organizations' efforts to meet the Open Data characteristics than the current binary decision method (i.e., our method determines how far an initiative fulfills all characteristics of Open Data).

It is for our future research to extend and formalize the proposed method for measuring the level of adjacency to the Open Data

setting. Furthermore we are interested to measure the trend of some existing Semi-Open Data initiatives, based on the measurement method proposed. Hereby we can identify the dimensions along which it is possible to guide data opening initiatives and move faster towards the desired Open Data point.

7. REFERENCES

[1] –. 2015. Trendrapport open data 2015" (in Dutch). *Technical Report*, Court of Audit of The Netherlands (31 Mar. 2015). http://www.rekenkamer.nl/Publicaties/Onderzoeksrapporten/Introducties/2015/03/Trendrapport_open_data_2015. Accessed: 29 Jun. 2015.

[2] –. 2015. Open data 101. *Government of Canada site*, (29 Jun. 2015). http://open.canada.ca/en/open-data-principles. Accessed: 29 Jun. 2015.

[3] –. Open data research network. http://www.opendataresearch.org/content/2013/566/early-insight-2-definitions-open-data. Accessed: 29 Jun. 2015.

[4] –. *Website of the Court of Audit of The Netherlands*. http://www.courtofaudit.nl/english. Accessed: 29 Jun. 2015.

[5] –. The open definition. http://opendefinition.org. Accessed: 29 Jun. 2015.

[6] –. Open data handbook. http://opendatahandbook.org/guide/en/what-is-open-data/. Accessed: 29 Jun. 2015.

[7] –. Welcome to open government data. http://opengovernmentdata.org/#sthash.kjB8Ex3v.dpuf. Accessed: 29 Jun. 2015.

[8] –, REPRIS portal. https://WODC-repris.nl/Repris.html. Accessed: 1 Jul. 2015.

[9] –, StatLine portal. http://statline.cbs.nl/Statweb/?LA=en. Accessed: 1 Jul. 2015.

[10] Alberts, B., et al. 2015. Self-correction in science at work, *Science*, (26 Jun. 2015). Vol. 348, No. 6242, 1420-1422.

[11] Bargh, M. S., Choenni S. and Meijer, R. 2015. Privacy and information sharing in a judicial setting: A Wicked Problem. *In Proceedings of the 6th Annual International Conference on Digital Government Research (DG.O'15)*, Phoenix, USA.

[12] Burwell, S.M., Roekel, van S., Part, T. and Mancini, D.J. 2013. Open data policy – managing information as an Asset. White House Memo on Open Data Policy (9 May 2013). https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf. Accessed: 29 Jun. 2015.

[13] Carlson, L., et al. 2012. Resilience theory and applications. *Technical Report, Argonne National Laboratory, Decision and Information Sciences Division, ANL/DIS-12-1*, IL, USA.

[14] Choenni, S., Bargh, M. S., Roepan, C. and Meijer, R. 2015. Privacy and security in data collection by citizens. In *Gil-Garcia, J. R, Pardo. T. A., & Nam, T. (ed.), Smarter as the New Urban Agenda: A Comprehensive View of the 21st Century City*, LNCS, Springer.

[15] Choenni, S., Dijk, J. van and Leeuw, F., 2010. Preserving privacy whilst integrating data: Applied to criminal justice. *Information Polity*, 15(1-2), 125–138.

[16] Conradie, P. and Choenni, S. 2014. On the barriers for local government releasing open data. *Government Information Quarterly*, 31(XX), 10–17, doi:10.1016/j.giq.2014.01.003

[17] Dijk van, J., Kalidien, S. and Choenni, S. 2013. Development, implementation and use of a judicial data space system. *Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*, Oct. 22-25, Seoul, Republic of Korea. DOI= 10.1145/2591888.2591917.

[18] Franklin, M., Halevy, A. and Maier, D. 2005. From databases to data spaces: A new abstraction for information management. *ACM SIGMOD Record*, 34(4), 27-33.

[19] Hoogen, van den H. 2013. Taking care of research data: Outcome of a seminar on managing data integrity in research. *Library Connect Newsletter, Issue: Research Data Management*, (18 Mar. 2013). Vol. 1, No. 1.

[20] Kalidien, S. N., Choenni R. and Meijer, R. F. 2010. Crime statistics online: potentials and challenges. *Proceedings of the 11th Annual International Conference on Digital Government Research (DG.O'10),* Puebla, Mexico.

[21] Lindman, J. 2014. Similarities of open data and open source: Impacts on business. *Journal of Theoretical and Applied Electronic Commerce Research*, vol.9, n.3, 46-70.

[22] Nosek B.A., et al. 2015. Promoting an open research culture. *Science*, (26 Jun. 2015). Vol. 348, No. 6242 1422-1425, DOI: 10.1126/science.aab2374.

[23] Petit F.T., et al. 2013. Resilience measurement index: An indicator of critical infrastructure resilience. *Technical Report, Argonne National Laboratory, ANL/DIS-13-01*. IL. http://www.ipd.anl.gov/anlpubs/2013/07/76797.pdf. Accessed: 1 Jul. 2015.

[24] Swire, P. 2004. A model for when disclosure helps security: What is different about computer and network security? *J. on Telecommunications and High Technology Law*, Vol. 2.

[25] Verhulst, S., Noveck Simone, B., Robyn, C., Brown, K. and Paz, C. 2014. Released: The open data era in health and social care. http://thegovlab.org/nhs/. Accessed: 29 Jun. 2015

[26] Wartna, B.S.J., Blom, M. and Tollenaar, N. 2011. The Dutch recidivism monitor (4th edition). *Technical Report, WODC, Memorandum 2011-3a*. http://english.WODC.nl/onderzoeksdatabase/1964b-brochure-the-dutch-recidivism-monitor.aspx?cp=45&cs=6800. Accessed on 1 Jul. 2015.

[27] Wellington, B. 2015. Measles reporting and the dangers of semi-open data. *I Quant NY*, (26 Feb. 2015). http://iquantny.tumblr.com/post/112113837909/measles-reporting-and-the-dangers-of-semi-open. Accessed: 29 Jun. 2015.

[28] Wonderlich, J. –. 2010. Ten principles for opening up government information. *Website of Sunlight Foundation*. http://sunlightfoundation.com/policy/documents/ten-open-data-principles/. Accessed: 29 Jun. 2015.

[29] Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. and Sheikh Alibaks, R. 2012. Socio-technical impediments of open data. *Electronic Journal of eGovernment*, vol.10, nr. 2.