Utilizing Open Data: A Primer for Public Procurement Research

Csaba Csáki Clifford P. McCue Eric Prier

Abstract:

Numerous open data initiatives by governments around the globe ostensibly promote better transparency and accountability, yet questions have arisen regarding the immediate usability of these datasets. This research reports on an attempt to utilize purchasing data published under the open data program of the European Union, which provides all expenditure data over certain thresholds from 33 European countries. However, the data and its informational quality as it has been published in CSV format leaves holes in trying to close that accountability gap across countries. This case study offers a recursive model which clearly conceptualizes the quality of data and information, and the research serves as a functional primer warning for users of the experientially-based issues of utilizing this and other open data. Key findings illuminate potential issues when working with open data and provide eight specific caveats on how to navigate the open data initiatives by governments.

Keywords:

open data; data quality; information quality; public procurement; purchasing; transparency; accountability; corruption prevention; European directives; TED.

1. Introduction

The concept and associated practices commonly known as 'open government data' have been around for well over a decade (Blakemore & Craglia, 2006), and its availability emanates from the "*right to information*" (Chun, et al., 2010; Organization for Economic Co-operation and Development, 2008). In terms of government practices, data provided by governments (open data) leads to usable information that generates particularized knowledge that promotes political, social, and economic transformation (Verhulst & Young, 2016). Recently, open data initiatives have fallen more broadly under the umbrella of Electronic Government (Chun, et al., 2010; Davies, 2013; Jaeger, 2003).1 Electronic Government (e-Gov) is often contextualized as the use of information technology to enhance the efficiency, effectiveness, transparency, and accountability of governments (see Jaeger, 2003; Janssen, 2011; Kraemer & King, 2003; Norris & Lloyd, 2006; World Bank, 2012).

Generally, open data refers to government initiatives that make both raw data and information in the public sphere available to be used and repurposed. While researchers of open data often emphasize their potential advantages (Chun, et al., 2010) open data initiatives are not without limitations (Zuiderwijk, et al. 2012; Martin, et al. 2013). For example, maintaining national security or protecting the privacy of citizen data limits the availability of certain types of data for public consumption. It is important to remember, however, that data are simply raw observable

1 While the literature distinguishes e-government from e-governance (see for example, Marche and McNiven, 2003), the current research focuses on open data and doesn't address this debate

facts or figures and only when data are contextualized to make them usefully meaningful are data transformed into "information."

But recent observations attest to the fact that while Sir Tim Berners-Lee argues that free open data are "*a great way to put power in the hands of citizens*" (Information Age, 2015), the World Wide Web Foundation (2015) reports that fewer than 8% of countries provided data on government budgets and spending, public sector contracts, and company ownership under open formats and open license agreements. This is hardly consistent with Anti-Corruption Open Data Principles (2015) advocating that open data needs to be available online; machine-readable in bulk so that it can be downloaded as one dataset and easily analyzed; free of charge; and open-licensed so that anyone has permission to use and reuse the data. However, examination of open government initiatives such as the European Union Open Data Portal (https://data.europa.eu/euodp/home) and other programs reveal substantial issues involving poor data quality (World Wide Web Foundation, 2017), and this has resulted in much research providing frameworks of quality dimensions or recommendations about open data measurements (Frank and Walker, 2016; Zaveri, et al., 2012). Given that most data quality (DQ) literature focuses on technology-related characteristics (see for example Rula and Zaveri, 2014) that are supply-side oriented, a dearth of studies address the user or demand-side of the open data equation (Frank and Walker, 2016) – the subject of the current article.

The Tenders Electronic Daily (aka TED2), the public procurement open data portal of the European Union provides the basis for this exploratory case study. The TED data is considered "open" in a strict sense (Prier et al., 2018; Davies, 2013), and the focus herein is to examine the quality of the TED data from the point of view of an end-user as the user-experience relates to the ostensible openness promised by e-government. The premise herein is simple: bad data leads to bad information, and bad information often leads to poor decisions which can be extremely costly. Therefore, governments not only have a responsibility of making public data freely available, but it must also ensure that the data provided is free of defect and easily usable. Only when data is open and free of defect can end-users make knowledgeable decisions which in turn, should lead to better governance.

The article is organized as follows. First the relationship between open data and usable information is explicated followed by a brief overview of open data quality frameworks that guide this case study. The next section looks at the TED dataset and its context – followed by some methodological groundwork. The core part of the paper identifies experiential challenges of the TED open data, and the final section provides conclusions and recommendations that may be used to enhance the quality of the TED dataset.

2. Open Data and Information Quality

Governmental webportals have become a key interface between citizens and governments in nearly all societies (Norris and Lloyd, 2006; OECD, 2008). While well-designed online services are able to open up government processes and strengthen the link between citizens and various policy and administrative actors (Chun et al.,2010), in all democratic systems it is transparency that

2 For a complete explanation of the TED initiative of the EU, please see http://data.europa.eu/euodp/en/data/dataset/ted-1.

anchors the relationship between integrity and accountability arising from government conduct. This implies that accountability requires providing answers and remaining responsible to others who have a legitimate claim to demand an account (Bovens et al., 2014). Meeting these goals assumes a requisite level of openness whereby non-government actors (the public) have mechanisms to know what governmental actors are doing. Thus, data about governmental behavior may be used to hold actors of the public sphere accountable for their actions (or inactions).

Increasingly data generated in public policy domains are being captured, digitized, and stored, and data availability can result in two outcomes. First, transparency goals are perceived to be enhanced through improved accessibility, which in turn can promote transactional efficiencies and better planning on the one hand, and greater accountability on the other (Leipold, 2007). Second, clarity in public expenditures used to fulfill public sector objectives, obligations, and activities in the pursuit of desired policy outcomes (Prier and McCue, 2009) can enhance better planning and delivery, as well as promote greater business access and enhance competition. However, even when governments provide data accessibility in an open environment, it should be available in a concise, useable and meaningful manner (Frank and Walker, 2016). This suggests that users of open data must be confident that the data is free of defects and that they are able to utilize the data to make informed decisions whether in the public or private sectors. If open data has defects, such as the data is incomplete, invalid, or not compliant with procedural rules, a data quality (DQ) problem becomes evident. When an end-user utilizes defective data to make decisions, the result is an information quality (IQ) problem, and Figure 1 helps to explain this situation.

Figure 1 - Conceptual Relationships Linking Data, Information, and Decisions

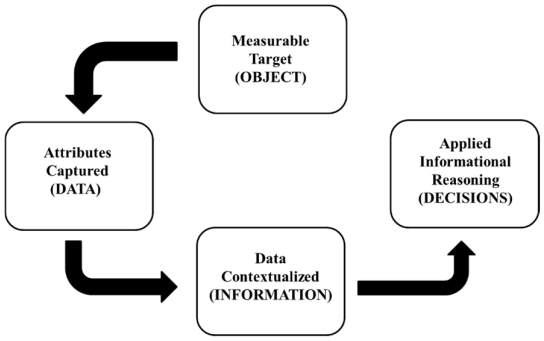

Figure 1 depicts the conceptual relationship between data, information, and decisions adapted to the public procurement decision making situation (see Shannon, 1948; also Liew, 2007). Beginning with the object to be represented or measured, the figure portrays the link between that object and its attributes that may be captured and stored as data. Consequently, discrete objective facts embody the useful features (object attributes) about empirical phenomena that become 'data' consisting of observable representations of a targeted phenomenon or event. Moreover, when each data attribute complies with the rules relating to that piece of data – high levels of data quality are obtained.

Data becomes information when users take and organize the raw data – giving it context – in ways that generate meaning and at which point the data become information. The final linkage in Figure 1 reveals that informed decisions require transforming information to create value for the open data end-user. Thus, data leads to information that undergirds decisions through purposive

application of cognitive reasoning that includes intellectual deliberation as to what, how, and why to apply information that results in effective decisions.

As an example, consider the measurable object to be an actual purchase that takes place on December 1, 2019. When an invoice is created and it registers a purchase date (the data value) as 01-12-2019, all appropriate data is presented to users in a concise and meaningful manner and a high level of information quality can be achieved. However, given the nature of the linkages exhibited in Figure 1, data problems, if they exist, may be inherent in the observations of the object and may be often related to the accuracy and validity of the data attributes. For instance, if a date attribute contains '13' in the month field, this is clearly a DQ problem, and this can then lead to information problems that may or may not become evident when the data is presented for use in a specific context (such as deciding about bidding deadlines, for example). Therefore, DQ problems often lead to IQ issues that may or may not systematically impact informed decisions. In addition, data measurement error issues can also result in IQ problems. For example, a poorly formatted form – while containing all data attributes – can lead the user to misunderstand the appropriate meaning of the data (an object) because the captured attributes are not stored in the appropriate fields. So DQ issues – whether systematic or random, frequently occasion IQ issues thereby making informed decisions problematic.

When a datum complies with all the rules associated with the attribute, transforming DQ into useable and meaningful IQ helps procurement officials to make better decisions. Of course, each linkage in Figure 1 is the result of context generated within and by that coupled association. This implies that context is not confined to the "Information" box but actually anchors the whole recursive process: object selection is done within a context, and so is data capture, yet when the data is stored

in a database or presented in a CSV file, it tends to be stripped of context. All of this suggests that in general, the open data end-user has essentially two options leading to the information box: either attempt to reconstruct the meaning of the object and thus also of the data or create new meaning of the data and hence of the object. These alternative purposes of data manipulation govern the conceptual linkages described above, and it strongly suggests that data in and of itself has no intrinsic value, i.e. absent specified purposes, open data has no intrinsic value proposition.

A logical predicate of good procurement decisions is their basis in appropriate data and information, yet the meanings of DQ and IQ can be elusive and challenging concepts, especially in the context of digitized government data. Scholars use the terms in different research contexts often without establishing clear definitions or only focusing on a narrow aspect of practical application (Wormell, 1990), and when coupled with the evolution of technology, dimensions and frameworks of assessing these issues have changed over time (Glogowska, 2016). Adding potential database (DB) issues (Levitin and Redman, 1995) surrounding timeliness of software updates and DB system reliability, accessibility, usability and security (Fox et al., 1995) multiplies the complexity.

The appreciation of various characteristics associated with the numbers, definitions, and measurability of DQ and IQ has recently emerged (Scannapieco and Catarci, 2002). For instance, the machine readability approach (Erickson et al., 2013) is concerned with linking, finding, relating and reading information typically using automated processes (Rula and Zaveri, 2014), and characteristics typically considered include number of formats, traceability, automated

tracking, use of standards, trustworthiness, authenticity or provenance. Or consider the ambiguous relative construct of "*fit-for-use*" (Wang and Strong, 1996) whereby data or information considered appropriate in one setting may not display acceptable attributes in another (Tayi and Ballou, 1998) thus encumbering measurability and operationalization (Frank and Walker, 2016). Indeed, the problems of data and informational intersubjectivity is not new (see Strong et al., 1997), and they often exude from whether they are related to the data or information itself, its manipulability, or its user intentions, among others (see Emamjome et al., 2013; Klobas, 1995; Naumann and Rolker, 2000; Olaisen, 1990). In summary, it becomes apparent that there are compounding complications from using numerous dimensions to address data and information quality, but adopting a user-centric perspective shows that content perceived as excellent quality by some users might be perceptively considered poor quality by others (Chai et al., 2009). However, it must be noted here, that quality of the data as stored, accessed and manipulated can substantially differ from the quality of the information that the data may offer.

3. The Case of Tenders Electronic Daily in the European Union

The Context

Making public procurement decisions understandable motivates the open data initiative of the European Union (EU). The EU data portal offers a single point of access (https://data.europa.eu/) to a growing range of data covering EU bodies and member states. By providing free access to data, the EC aims to promote transparency and through that accountability. A key component of this initiative is the Tenders Electronic Daily (TED) dataset comprised of public procurement data originally published and accessible as part of the TED public procurement website (http://ted.europa.eu/). The obligation to tender and thus become part of the TED dataset depends on several things, two of which are 1) the type of contracting authority (government or agency) and 2) the value of the planned purchase depending on the object and type of contract such as for goods, services, or works. In order to treat all businesses across Europe fairly, EU directives establish minimum public procurement rules and requirements. To appreciate the scope of the activities captured in this data, TED publishes over half a million awards per year worth about 420 billion Euro per year.

The Source

The current study utilizes data from the TED open data website where bulk European public procurement data is published (https://data.europa.eu/euodp/en/data/dataset/ted-CSV). Data come from the official online version of Supplement 32 to the Official Journal of the European Union, which publishes all public procurements made in EU member states that fall above minimum threshold amounts stipulated in the EU regulation for procurement. Other than the twenty-eight EU members, five affiliated countries also publish tender and award notices in the TED Journal to gain access to the EU market – these are Iceland (IS), Liechtenstein (LI), The Former Yugoslav Republic of Macedonia (MK), Norway (NO) and Switzerland (CH). Data in the Journal are collected from standardized public procurement forms as required by the corresponding EU Directive (Directives 2014/18 and 17) and their Annexes. At the time of download, the open data files stored information captured from the contract notices reported in standard forms #2, #4, #5, or #17. These forms announce information concerning a future purchase (i.e. call for tender). In addition, the data files also report contract award notice information on the outcomes of the procurement obtained from standard forms for public procurement #3, #6 or #18

(TED, 2016). Data in the TED Journal is entered through online forms, one notice at a time. The published open datasets also come with a user guide (TED, 2016) describing the fields in the available files.

The Actual Data

The TED open data is very complex because the CSV data files are embedded with three levels of procurement information: a) contract notices (CN); b) contract award notices (CAN); and c) contract awards (CA). While the process of public procurement is inherently complicated, for now it should suffice to state that one and occasionally two CNs lead to one CAN, but one CAN may lead to one or more CAs associated with it (this is because a CN may have a preliminary notice; while a single call may have several parts or lots with each leading to a separate contract being awarded – but published in one CAN notice). Each dataset is published in CSV format using UTF-8 coding and it contains data regarding the version of the XML schema definition (XSD) used by the Publications Office of the EU to publish the data. Calls (CNs) and corresponding awards (CANs and AWARDs) are presented in separate files each with its own data structure represented by a specific header row in the corresponding CSV file. Notices and awards each have both annual as well as cumulative (2009-2015) files. All data files were downloaded January 17, 2017. The total size of the fourteen different data files is approximately 2.13 GB consisting of over 4.5 million records. The datasets are accompanied by a codebook that serves as a guide: contract notice datasets (CN) have 54 fields, while award datasets (CAN/CA) have 50 fields.

4. Methodological Considerations

This research fills a substantial gap in the literature through a case study documenting issues experienced with actual use of the EU TED open data files. While statistical data challenges are reported in Prier et al., (2018), this study provides an experiential general primer on how to approach open data that is anchored in the theoretical literature. Readers can then generalize the applied findings of this public procurement open data by knowing what to expect in terms of operational results of utilizing open data and anticipating the challenges they might they face when attempting to utilize open data – especially for the first time. This helps to identify common issues in accessing open data preparing scholars to judge the status of a dataset before investing substantial effort to ready it for use.

Using the TED dataset as a single case, this study recounts the data-user experience by documenting the issues in each step of the data utilization process. What makes this case especially compelling is that this data is mandated by EU law and regulations and it is a result of iterative cycles of policy-making. One of the highest public sector ideals remains accountability, and this dataset is chosen exactly because it is intended to be an example of quality open data that is supposed to be, by its nature, transparent. While the study is organized in a segmented chronological path that follows a natural progressive timeline of the steps one normally takes to explore new data, the findings offer conclusions based on several key characteristics identified in the literature to judge open data quality.

A set of commonly-available software tools were utilized including MS Excel, MS Access (both from Office 2010 on Win7 OS), SPSS (v22.2), Oracle Database (11g Release 11.2.0.4) with SQLDeveloper interface (v4.1.5), MySQL Database (v5.7.14 on WAMP v3.0.6) with MySQL Workbench (v6.2.5) and the R open source statistical package (v3.3.1). Regarding the default

language setting of the MS Windows operating system (and through that the MS Office package) English (North American) and Hungarian were the languages of choice. Most of the statistical analysis of the original research had been completed in SPSS with some work done in Excel (using Power Pivot) to understand and manipulate the dataset in order to eliminate errors, discover operational issues, and to understand the nature of the data beyond mere reading and statistical summaries.

5. Case Study Results

Data quality is assessed on the following key dimensions identified in the literature: availability, accessibility, readability, technical qualities, data structure, content, usability (ease of use), traceability, and fit-for-purpose.

Availability (and awareness)

Theoretically, anyone interested in using open data for informed decisions should be able to locate it without be required to use "public records" requests to acquire the data. Some public procurement scholars have been aware of the availability of TED data, but prior to July 2016, bulk data had been only made available through periodic updates from volunteers associated with the OpenTED project (http://ted.openspending.org/#welcome). Since then, however, the European Commission itself has published machine-readable CSV bulk extracts of the TED data on its open data portal thereby making the OpenTED project superfluous and the current case study is confined to this EC data only.

Accessibility

One key point of the open data initiative is that the data provided is easily assessable. While there are annual data files available, there is also an integrated file covering seven years available at https://data.europa.eu/euodp/en/data/dataset/ted-CSV. The files are posted in CSV format and individual file sizes span from 60MB to 300MB (except for the integrated files which are .5 and 1.6GB). None of these posed any issues during download: with a normal Internet connection it only took fifteen minutes of work to download and sort the 22 files. The official TED Journal on the other hand offers individual notices as well as daily digests (in zipped xml format) – the size of which is typically 150MBs per monthly data.

Readability

Even though the format is standard CSV, initial opening of the first file using Excel resulted in unstructured lines with no segmentation (i.e. each line was rendered into one cell instead of recognizing the columns): the language setting of the MS Windows OS impacts how Excel reads data, namely, the Regional and Language settings determine the default field separator. Using English as a default enables Excel to read the data correctly and properly separate the fields. Substantively this means that when using other languages such as in this case, Hungarian, the Excel default separator may have to be reconfigured.

But even when the lines were properly segmented into fields, some of the text was scrambled. In fact, reading the file into SPSS or Access – and later adding it to an Oracle and a MySQL Database – often resulted in unreadable text with strange, meaningless characters. Since EU members may use any of the 28 official languages for their tender notice announcements, the problem may be appeared rooted in the encoding schema: the CSV files use UTF-8 which needed to be specifically

defined (and needs to be the 2-bytes version to cover all languages). In Excel the solution was to "import" the CSV instead of simply opening it to allow defining the encoding schema. But Access offered a slightly different hurdle: the character coding is not simply UTF-8, but it must also be language independent (i.e. should not be English or Hungarian, but "All" due to font mixing). This does not work in professional databases where one needs to use a special SQL setting before reading or manipulating the data. As a last point, although expected file sizes were reported at the download page, there was no ready documentation explaining how many lines of data should be correctly read into the CSV files. Hence users have no reliable information that precisely describes a properly imported file.

Technical qualities of the data

Reading the data also entails considerations about datatypes. While Excel has a limited capability to differentiate between a few datatypes such as Text, Date or Number, the CSV format does not carry such information (Excel would automatically assign a datatype though when the CSV file is opened – if nothing else it uses the "General" type as a default). On the other hand, many database or data management tools would offer a range of types, and this set might be quite sophisticated. It is noteworthy that each tool utilized in this project had its own special names and options – with Oracle having a different approach compared to Access or MySQL or even SPSS. In fact, Oracle is known for having a unique stance on datatypes – such as the lack of Boolean. Furthermore, each tool used herein had a different take on the "Date" type (which is understandably crucial in this investigation). All of the tools (Access, Oracle, and SPSS) offer automatic type recognition and also make suggestions regarding the potential maximum size or length of relevant types (such as integer or text).

Although it might sound like a minor concern, much effort was spent struggling with fields of "Date" type. This is due to the fact that there is no unique standard for storing date/time values, each tool offered different options which unnecessarily complicated what should be a simple conversion. For example, the Hungarian version of Access refused to accept the (given TED) English date format, e.g. it would not take "DD.MMM.YY" or "DD-MM-YY", instead, it would require "YYYY.MM.DD" or something similar. Oracle had similar issues while also accepting only a limited set of formats (and, interestingly, would not allow a field with only the year, such as 2015).

Structure

Text field length: Expect variations in field truncation because Access would truncate fields with longer size while Oracle would reject such records – all of which suggests that knowing the longest possible text field is important. Remember that choosing very large values for all fields results in larger database files requiring more storage and more memory to manipulate the data.

Multiple values in one field: A significant issue concerns occasional multiple values in one field where one column reported additional CPV codes and another column registered multiple winners. The former issue can be resolved with some text manipulation, but the latter presents a more sophisticated problem in separating out the individual data values. In public procurement more than one winner may occur in several cases: as a result of using a framework agreement; a dynamic purchasing system; in the case of contract separation into lots; and when the call notice has different parts. In the case of lots or parts, there should be one "contract award" with a unique

"CONTRACT_AWARD_ID" for each lot or part under the same CAN ID. However, for the other two cases, different authorities (in different countries) appear to have established different standards that report the resulting contracts.

Multiplied lines: In addition to all of these issues, every "periodic indicative notice without a call for tender" (a special CN typically used by utilities) is duplicated in the CSV file – which is apparently a mistake. In fact, it has an additional line for each separate lot – which is an even bigger potential source of informational distortion. What is problematic about these cases is that they should not directly lead to CANs yet some are specified (without an actual call with a different CN ID). There is no explanation for this situation in the codebook, and neither do the regulations give any indication of the need for duplicates. Furthermore, while some duplicated lines mirror each other, other lines show some empty fields in one line that are completed in its "duplicated" record (e.g. CPV code). As they both (or all) have the same date stamp, it is very suspicious whether they represent legitimately different actual CNs. The extra values in certain fields are not the result of a change or modification (i.e. the duplication is consistent for all such lines except for a few exceptions – e.g. for 2014 there are 10,050 such cases and even one triplicate).

These unexplained duplications particularly inure at the time of statistical analysis when data uncertainty may produce substantial statistical aberrations. While analysts might ignore cases or remove duplicates, few of the available guides, documents or informational explanations offered a clear resolution.

Content

Inappropriate values in specific fields: Some records report their form number as '2' when these notices should reflect use of Form #4 which indicates use of the wrong form or following outdated reporting regimes. There are also apparently intentional misrepresentations as well, because some contracting authorities entered values into fields that sometimes look suspicious. For instance, the estimated value of a contract is occasionally ad-hoc, such as €1234567 – instead of a proper calculation.

Cancellations: Intimately knowing the measurable objects of the data is imperative to understanding this dataset which requires a brief background in procurement operations. When a contracting authority amends the contractual condition, a modification to an existing contract notice leads to a new entry (called "Additional Information") which may or may not require a new contract notice identification number (CN ID). However, actual contract cancellations are only captured through "cancellation notices" that require a new form. However, this open data experience made clear that the method of and forms for reporting CN and CAN modifications and cancellations had changed over the period covered by the dataset given the new 2014 public procurement directives, and this has significant informational quality implications. This means that minor contract modifications required reporting only a modification that was "additional information" in the new form #14, while major changes including cancellations require using the full notice new form #2. Table 1 shows the distribution of records (forms) across the two CSV file types for the year 2015: contract notices (calls for competition) and contract award notices (actual awarded contracts). One can readily see that since there are no records of contract modifications generated from form #14, attempts at reconstructing the objects (contract processes) are nearly impossible given the data for this procurement procedure. Further, unsuccessful procedures under

the new directives should not be canceled but instead contracting authorities should use the relevant CAN form reporting "no award".

Problematically, all of these contingencies are not readily clear from documentation at the website, and the role of the algorithm in generating these apparently anomalous CSV files. In sum, these ambiguities produce continuity problems within the dataset because those countries which have not yet ratified the new directives into national law apparently still use the old forms.

Table 1. Standard Forms Generating Records of Calls for Competition and Contract Awards, 2015

	Percentages (Ns)		
	(1)	(2)	(3)
Contract Form Number (Descriptor)	CFC Records	CAN Records[a]	Directives
Form 1 (Prior information notice)	0.0% (1)		2014/24/EU
Form 2 (Contract Notice)	88.7% (173,250)		2014/24/EU
Form 3 (Contract Award Notice)		92.5% (497,635)	2014/24/EU
Form 4 (Periodic indicative notice-utilities)	0.2% (338)		2014/25/EU
Form 5 (Contract Notice-utilities)	9.4% (18,361)		2014/25/EU
Form 6 (Contract Award Notice-utilities)		6.5% (35,054)	2014/25/EU
Form 7 (Qualification system-utilities)	0.9% (1,674)		2014/25/EU
Form 10 (Public works concession)	0.1% (289)		2004/18/EC
Form 17 (Contract Notice-Defense or Security)	0.7% (1,447)		2009/81/EC
Form 18 (Contract Award Notice-Defense or Security)		0.4% (2,417)	2009/81/EC
Form 21 (Social and other specific services-public contracts)	0.0% (9)	0.5% (2,580)	2014/24/EU
Form 22 (Social and other specific services-utilities)		0.0% (15)	2014/25/EU
Form 23 (Social and other specific services-concessions)		0.0% (92)	2014/23/EU
Form 24 (Concession notice)	0.0% (5)		2014/23/EU
Form 25 (Concession award notice)		0.1% (390)	2014/23/EU
Total	100% (195,374)	100% (538,183)	5 Different

Source: Calculation by authors

[a] These do not include records that award contracts based on voluntary *ex ante* transparency (VEAT) notices

Content duplicates: As previously discussed, the 2014 Directives marked procedural changes in reporting Contract Notice cancellations and the empirical result has led to CSV files rife with two lines of records for cancellations: one for the modification of the original call and another for the cancellation (coding it as a modification as well). This is clearly a meaningless duplication of the CN (using the same CN ID) as there is no actual call here can potentially distort statistical calculations concerning CNs.

Missing values: According to TED documentation (TED, 2016), connecting contract notices with contract awards requires marrying up two CSV files through a special ID field called Future_CAN_ID. This field is often left empty which makes data quality checking problematic. For example, the consolidated 2009-2015 CN files showed that 66% (Total N=758,604) of this field was empty across that time period. Of course, there are often legitimate reasons for this situation including the fact that the call might have been cancelled (no winner was announced), or the procedure had not yet concluded at the time of publication. Unfortunately, it is also possible that the cancellation was not recorded or improperly coded, thus leaving the user of the data with the assumption that the call is still open.

All of this suggests that currently there is no means to establish this state of the data quality, unless one goes back to the source database and searches each individual notice in question, which is, of course time prohibitive. On the other hand, it is also possible that the cancellation generates a new CSV record (see above) when it officially should not (the problem may be rooted in the TED, where cancellations are recorded as "Additional Information" instead of leading to a new notice with unique CN ID or the use of the 'cancellation' field). Moreover, many Future_CAN_IDs point to documents that will be published in the future but the open data file of that year had not been made available at the time of download. Therefore, assuming that one would not individually search and download relevant notices from the TED using the interactive TED data website, certain types of analysis are either not possible; incomplete; or subject to both validity and reliability issues. This further suggests that although procurement accountability would often require an easy connection of calls to awards and their dates, the generated CSV files make this difficult in many ways – especially if one wants the data analysis process to be (semi)automated.

In addition, there are other potential problems associated with real missing values in numerous other fields. Of course, many of these missing data fields are concentrated in non-key or non-essential data elements such as national contract ID or the national code of the authority. But when needed, the lack of data values in these fields would cause problems in case of statistical analysis targeting those specific fields.

Traceability

Each CN record has a specific variable called "Future_CAN_ID" (and another "Future_CAN_ID_Estimated") which shows the ID number of the award notice resulting from the given call-for-competition notice. However, evidence suggests that this link is often generated inappropriately. Not simply may the Future_CAN_ID value be wrong, but during the generation of the annual CSV CN file new records were created that have no apparent meaning in the original TED. The reason behind these superfluous records is unknown, but manual investigation of several such cases lead to the conclusion that the CSV generating algorithm connects together otherwise unlinked CN and CAN items of the same contracting authority. In one case a Polish contracting authority had 14 CNs in 2014 with 16 CANs in the same or later years, but in the CN

CSV file there were cross-connected records totaling 153. On the other hand, a few of these CNs had additional information (i.e. modifications) using new CNs – but those CNs do not appear in the CSV file at all. The number or percentage of these false (or missing) lines is hard to calculate but it may be in the range of several percent annually. One potential reason for the confusion might be (again) that in 2014 new forms had been introduced.

The clear overall effect of this situation is the unreliability of traces from Contract Notices to Contract Award Notices. Without other means of confirmation there is essentially little confirmable traceability across the breadth of the data in the two CSV filetypes. Obviously, there are CNs without CANs, since not all calls will result in awards – some are revoked, while others may result in no award (no offers submitted or the evaluation was unsuccessful). In addition, the fresher a CN, the more likely the procurement process has not yet concluded. This is normal – as long as there is a cancellation or no award notice issued to record the fact. However, some of such notices also seemed to be missing in the CSV files, which further complicates the linking of records. To make an adequate assessment of the extent and impact of missing data requires a substantial investment of time in understanding the content of the procurement forms and the process of publishing notices – both of which can slightly differ among member countries. For example, the new form #2 should be used to report cancellations, but there appears to be no such field to enter the information, or to reference the notice that is being cancelled in that form.

Another question that arises when trying to understand how pieces of the dataset are linked is how those links have been generated and how the original data (in the TED) is being stored. This then also requires the researcher to know about the structure of the source storage e.g. whether it is stored in normalized tables; in the form of documents; or stored as the original pdf forms. Obviously, this drastically limits procurement field experts who then must also possess the skills of database experts.

Ease of use, usability

Similarly, ease of use appeared to depend on the same three matters as assessing the data: the format of the file (CSV), the tools required to work on the data (various tools had been tried), and the structure and content of the data in the files. The difficulties experienced due to language settings and date formats kept coming back during the analysis of the data whenever data had to be transferred between research sites using different language settings or needed to be loaded from one application to another. Although the data structure is described in the guide, a deep understanding of the meaning of various fields required extensive understanding not only of public procurement but also specific details of EU procedures. This was further complicated by the fact that the CSV fields often did not fully reflect either the fields in the TED nor the original forms contracting authorities required to use when submitting data related to calls and contract results. Even the guide did not explain the mapping between these three formats which further required additional effort in connecting the dots whenever a new research question was asked from the datasets. The fact that data is published in non-normalized form also required additional attention when making statistical calculations (due to multiplication of field values over numerous records).

Fit-for-purpose (value)

Considering the problematic dataset complexity and the requisite deep domain knowledge of European public procurement formalities, doing any kind of statistical analysis utilizing this open

data required a lot of careful preparation and attention including a lot of manual data cleaning. Moreover, analysts must prepare to consider the non-standardized way individual countries have reported data into the TED: some were at contracting authority level while other countries exercised control at the central government level, resulting in missing codes, or missing values or inconsistencies in the names of authorities – all impacting statistical analysis involving the affected data fields.

Perhaps the most obvious recurring issue that will affect any data analysis using this open dataset is the problem of missing values across numerous fields. In some cases, a simple visual browsing of the file was enough to see that certain records had missing values, while in other cases the statistics revealed a number of "no value" items. Whether the missing value was a result of the way CSV files were generated or that data had not been entered at all (the latter is most likely) is not known, but these data validation concerns can be difficult to detect because cancellation of calls appears to not always be reported consistently. Consider that over the period of 2009-2015 60% of the calls (452,078 of 754,378) had no reported outcome (i.e. had no award indicated and yet were not cancelled either). Given the centrality of a procurement outcome in an open dataset that is devoted to making government spending more transparent, a more complete documentable explanation by the authorities providing the data seems reasonable.

6. Discussion and Insights on Open Data

The technical details of the open data experience outlined above suggest that if one wishes to go beyond the analytical capabilities of Excel, technical issues may remain a hurdle to the user of the open data files and should probably be addressed. Overall, loading the data into management tools may require several preparation steps, such as assessing the types and sizes of fields as well as obtaining and applying the proper settings during the conversion. This case study offers a cautionary tale to those new to open data – and does so with a clear warning: one should be careful to spend time and effort preparing any project that intends to utilize large open data sets prior to making resource allocation decisions. While quality of actual datasets differs widely, this study documents at least three seemingly unrelated skills that are needed to appropriately utilize open data including those based in data management; data issues associated with software applications, as well as domain knowledge and expertise when attempting research that intends to rely on open data.

The result of this case study suggests the following eight generalized issues that end-users should consider when preparing to work with open data:

1) Finding the data: check for the data source to be authentic and whether the data is up to date and if it came with adequate and up-to-date description and sufficient documentation;

2) Downloading data sets: open data may come in many different formats and its size could be large (in the range of gigabytes) and is often composed of several files or parts;

3) Opening, loading and checking files: make sure that you have several tools available and that their settings fit the requirements of the data format – if something does not look right, try different language, coding and location settings;

4) Transforming the data: open data often looks different in software tools and transformation of different formats might be necessary – special support or expertise may be required to decide which tools fits best (don't stick with a tool just because that is the only one you know);

5) Assessing structure and content: even with available documentation, be careful – there might be errors, missing information, or the data structure might be so complex that considerable domain expertise might be required to understand both the meaning and the structure of the data;

6) Linking inside and outside the set: open data is rarely standalone and is often composed of several parts of related/connected datasets that may require referencing documentation of other sets (i.e. country codes, national abbreviations, etc.) - pay special attention and double check all such references for accuracy;

7) Manipulating the data for use: search out and find explanation for duplicates, missing values or even missing or omitted fields;

8) Interpreting and analyzing the data: depending of the issues uncovered during the earlier steps, the researcher might need to reconsider the questions that could be meaningfully answered from the dataset in actual use (which often differs from the intended use). Special attention should be paid to any generated statistical results which obviously depend on the records/fields/values.

7. Conclusions and Potential Future Directions

Open data are gaining increased attention in academic research, but data quality can vary dramatically. While a few frameworks have been put forward on how to assess open data quality and what measures to utilize, experiential studies investigating actual cases are lacking. This case study utilized procurement data from EU countries to demonstrate potential generalized hazards likely to be found in open data relevant to a variety of academic fields, and the experience recounted here provides useful insights for others planning to work with open government data for the first time.

This study explains a simple model that describes the conceptual relationships between a measurable target or object of inquiry; data; information; and policy decisions, and it lays the groundwork to more clearly think about myriad open data issues. This recursive model suggests how scholars can clearly conceptualize the quality of data and information and it is consistent with social phenomena that are often subject to this problematic lack of isomorphism (for example, see Bailey 1990, 13-47). This conceptual precision suggests that modeling social processes – in this case, procurement – involves varying levels of data itself which in turn influences how data and informational quality is conceived. For instance, filling out forms is actually data generation of the procurement process (the object). However, when the forms are transformed into flat CSV files, the forms can then also be considered to be the object which generates data in CSV format. Applying this logic reveals the consistently recursive nature of data generation even when decisional knowledge about the object is ultimately the goal of any data generation algorithm.

In sum, this article explains how the relevant literature examines specific data and informational quality issues that are often discussed in isolation from real-world experience. What makes this study different is filling the gap between theory and practice for researchers of open data by illuminating potential issues and providing applicable solutions. The eight general issues

described here go beyond offering differential measures of quality and instead, prepares researchers with warnings and tips on how to think about navigating the proliferating nature of public sector open data.

Acknowledgments

The authors wish to thank the staff of the European Commission for use of the TED CSV dataset (2009-2015).

References

- Bailey, J. E. and Pearson, S. W. 1983. Development of a tool for measuring and analyzing computer user satisfaction. *Management Science*, 29, 5: 530-545.
- Bailey, Kenneth D. 1990. *Social Entropy Theory*. State University of New York Press, Albany.
- Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys*, 41, 3, 16: 1–52.
- Blakemore, M. and Craglia, M. 2006. Access to Public-Sector Information in Europe: Policy, Rights, and Obligations. *The Information Society* , 22,1: 13-24.
- Bovens, M., Goodin, R.E. and Schillemans, T. (eds.) 2014. The *Oxford HanDBook of Public Accountability*. Oxford University Press, Oxford.
- Chai, K., Potdar, V. and Dillon, T. 2009. Content quality assessment related frameworks for social media. In *Procss of the Computational Science and its Applications Conference*, Springer, Berlin, 791-805.
- Chun, S.A., Shulman, S., Sandoval, R. and Hovy, E. 2010, Government 2.0: Making Connections between Citizens, *Data and Gov. Inf. Polity*, 15, 1-2: 1-9.
- Davies, T. 2013. *Open Data Barometer. 2013 Global Report.* Retrieved January 21, 2017, from
- http://www.cocoaconnect.org/publication/open-data-barometer-2013-global-report.
- Dedeke, A. 2000. A Conceptual Framework for Developing Quality Measures for Information Systems. In *Proceedings of the 5th International Conference on Information Quality*, 126-128.
- Emamjome, F. F., Rabaa'i, A. A., Gable, G. G. and Bandara, W. 2013. Information quality in social media: a conceptual model. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS 2013)*, Seoul, AIS Electronic Library (AISel).
- Erickson, J. S., Viswanathan, A., Shinavier, J., Shi, Y. and Hendler, J. A. 2013. Open Government Data: A Data Analytics Approach. *IEEE Intelligent Systems*, 28, 5: 19-23.
- Frank, M., and Walker, J. 2016. User centred methods for measuring the quality of open data. *The Journal of Community Informatics*, 12, 2: 47-68.
- Fox, C., Levitin, A., and Redman, T. C. 1995. *Data and data quality*: *Total Data Quality Management Research Program*, Sloan School of Management, MIT, Boston.
- Glogowska, D. J, 2016. *Information Quality Assessment in Social Internet Media*. Master Thesis, University College Cork, Ireland.
- Information Age 2015. *Sir Tim Berners-Lee calls on governments to honour their promises on open data. Online at* http://www.information-age.com/sir-tim-berners-lee-calls-governments-honour- their-promises-open-data-123458878/. Downloaded January 7, 2018.
- Jaeger, P. T. 2003. The endless wire: E-government as global phenomenon. *Government Information Quarterly*, 20: 323-331.

- Janssen, K. 2011. The Role of Public Sector Information in the European Market for Online Content: A Never-Ending Story or a New Beginning? *Info*: *the J. of Policy, Regulation and Strategy for Telecommunications, Inf. and Media*, 13, 6: 20-29.
- Kahn, B. K., Strong, D. M. and Wang, R. Y. 1997. A Model for Delivering Quality Information as Product and Services. In *Proceedings of the 1997 Conference on Information Quality*, Cambridge, MA, 80-94.
- Klobas, J. E. 1995. Beyond information quality: fitness for purpose and electronic information resource use. *Journal of Information Science*, 21, 2: 95-114.
- Kraemer, K.L., and King, J.L. 2003. *Information technology and administrative reform*: *Will the time after e-government be different?* Retrieved December 10, 2016 from http://www.crito.uci.edu.
- Leipold, K. 2007. *Electronic Government Procurement (e-GP) Opportunities & Challenges*. Talk given at the Congress to celebrate the fortieth annual session of UNCITRAL in Vienna, 9-12, July 2007. Retrieved February 26, 2012 from http://www.uncitral.org/pdf/ english/congress/Leipold.pdf.
- Levitin, A. and Redman, T. 1995. Quality dimensions of a conceptual view. *Information Processing & Management*, 31, 1: 81-88.
- Liew, Anthony. 2007. Understanding Data, Information, Knowledge and Their Inter-Relationships. *Journal of Knowledge Management Practice* 8(2): Accessed on 1-19-18 at http://www.tlainc.com/articl134.htm.
- Marche, S., & McNiven, J. D. (2003). E-government and e-governance: the future isn't what it used to be. Canadian Journal of Administrative Sciences, 20, 1: 74-86.
- Martin, S., Foulonneau, M., Turki, S., Ihadjadene, M., Paris, U. and Tudor, P.R.C.H. 2013. Risk Analysis to Overcome Barriers to Open Data. *Electronic Journal of e-Government*, 11, 1: 348–359.
- Naumann, F. and Rolker, C. 2000. Assessment methods for information quality criteria. In *Proceedings of the 5th International Conference on Information Quality*, Humboldt-Universität zu Berlin, Institut für Informatik, 148-162.
- Norris, F. D. and Lloyd, B. A. 2006. The scholarly literature on e-government: Characterizing a nascent field. *International Journal of Electronic Government Research*, 2, 4: 40-56.
- OECD [Organisation for Economic Co-operation and Development]. 2008. *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information* [C(2008)36]. Retrieved December 10, 2016, from http://www.oecd.org/internet/ieconomy/40826024.pdf.
- OECD [Organisation for Economic Co-operation and Development]. 2017. *Trust and Public Policy*: *How Better Governance Can Help Rebuild Public Trust, OECD Public Governance Reviews*, OECD Publishing, Paris.
- Olaisen, J. 1990. Information quality factors and the cognitive authority of electronic information, in Information quality: Definitions and dimensions. In *Proceedings of a NORDINFO Seminar, Royal School of Librarianship, Copenhagen, 1989,* Wormell, I. (ed.), Taylor Graham, London, 91- 121.

- Open Data Barometer (2016) accessed on 6-7-2018 at https://opendatabarometer.org/?_year=2016&indicator=ODB.
- Pignotti, E., Corsar, D. and Edwards, P. 2011. Provenance Principles for Open Data. In
- *Proceedings of DE2011*.
- Prier, E. and McCue, C.P. 2009. The Implications of a Muddled Definition of Public Procurement. *Journal of Public Procurement*, 9, 3&4: 326-370.
- Prier. E., Prismakova, P., and McCue, C.P. 2018. Analysing the European Union's Tenders Electronic Daily: Possibilities and Pitfalls. *International Journal of Procurement Management*, 11(6): 722-747.
- Rula, A. and Zaveri, A. 2014. Methodology for assessment of linked data quality. In *Proceedings of the 1st Workshop on Linked Data Quality at the 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany*, volume 1215 of CEUR Workshop Proceedings.
- Scannapieco, M. and Catarci, T. 2002. Data quality under a computer science perspective. *Archivi& Computer*, 2: 1–15.
- Shannon, Claude E. 1948 [2001]. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 379-423, 623-656, July, October, 1948. [Corrected version reprinted in ACM SIGMOBILE Mobile Computing and Communications Review 5(1): 3-55.]
- Strong, D. M., Lee, Y. W. and Wang, R. Y. 1997. Data quality in context. *Communications of the ACM*, 40, 5: 103-110.
- Tayi, G. K., and Ballou, D. P. 1998. Examining data quality. *Communications of the ACM*, 41, 2: 54-57.
- TED. 2016. *TED Processed Database*: *Notes & Codebook, Version 2.2*. Retrieved January 20, 2017 from http://data.europa.eu/euodp/repository/ec/dg-grow/mapps/TED(CSV)_data_information.doc.
- Ubaldi, B. 2013. *Open Government Data*: *Towards Empirical Analysis of Open Government Data Initiatives*, OECD Working Papers on Public Governance, No. 22. Retrieved January 20, 2017 from http://dx.doi.org/10.1787/5k46bj4f03s7-en.
- van Zeist, R. and Hendriks, P. 1996. Specifying software quality with the extended ISO model.
- *Software Quality Journal*, 5, 4: 273-284.
- Verhulst, S. and Young, A. 2016. Open Data Impact: When Demand and Supply Meet - Key Findings of the Open Data Impact Case Studies. GovLab, Retreived 12/18/2017 at: http://odimpact.org/files/open-data-impact-key-findings.pdf.
- Wang, R.Y. and Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12, 4: 5-33.
- World Bank. 2003. *Definition of E-Government*. World Bank: Washington, DC.
- Wormell, I. 1990. Information quality: definitions and dimensions. In *Proceedings of a NORDINFO Seminar, Royal School of Librarianship, Copenhagen*, T. Graham. London.
- Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J. and Auer, S. 2012. Quality assessment methodologies for linked open data. *Semantic Web Journa*l, 1, 5: 1-31.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R. and Alibaks, R. S. 2012. Socio-technical Impediments of Open Data. *Electronic Journal of e-Government*, 10, 2: 156–172.