Ágnes Czibik [1], *Bence Tóth*[2], *Mihály Fazekas*[3]

How to Construct a Public Procurement Database from Administrative Records?

With examples from the Hungarian public procurement system of 2009-2012

Government Transparency Institute reports: GTI-R/2015:02

December 2015, Budapest, Hungary

[1] Government Transparency Institute: http://www.govtransparency.eu/

[2] University of Cambridge and Government Transparency Institute

[3] University of Cambridge and Government Transparency Institute, corresponding author: mfazekas@govtransparency.eu

Government Transparency Institute

The Government Transparency Institute's foundation in 2015 was motivated by growing need and opportunity to do independent research and advocacy on transparency, corruption and quality of government in Europe and beyond. The Institute is a non-partisan think tank independent of governments, political parties or special interest groups. The aim of the Institute is to systematically explore the causes, characteristics, and consequences of low quality of government in an inter-disciplinary approach drawing on political science, economics, law, and data science. We believe, it is only through understanding and precisely measuring government behaviour we can build good government.

Authors

Ágnes Czibik, Government Transparency Institute

Bence Tóth, University of Cambridge, Government Transparency Institute

Mihály Fazekas, University of Cambridge, Government Transparency Institute

Date of publication

21 December 2015

Keywords

public procurement, database, data cleaning, data quality

Acknowledgements

The authors thank István János Tóth for his role in forming the concept of this paper and for his useful comments.

The authors are also thankful to Johannes Wachs for his comments; and to the participants of DIGIWHIST project (EU Grant Agreement number: 645852) for their suggestions. An earlier version of this paper has been prepared for the kick-off meeting DIGIWHIST in Cambridge (UK) in April 2015.

Contact information

Government Transparency Institute

E-mail: info@govtransparency.eu

Website: http://www.govtransparency.eu/

Government Transparency Institute

Executive summary

This report aims to provide a comprehensive outline of the potential challenges of building a database from publicly available public procurement records and the possible solutions to the identified problems. We use Hungary as an illustrative example as most problems faced in other countries are present there, and so any finding will be widely applicable. The chapters introduce 1) the process of public procurement and the publicly available announcements in Hungary; 2) the administrative processes which generate errors in the source data as well as cause problems for database building; 3) the general process of correcting errors in the database; and 4) specific problems and solutions with examples. The result is a comprehensive checklist, which could help building a database from public procurement records in any country where these are publicly available.

TABLE ES.1: CHECKLIST

Data availability			
	Regulatory framework		
		Thresholds	*Above which value threshold is it compulsory to publish public procurement procedures?*
		Special sectors e.g. defense	*To which sectors do special regulations and exemptions apply?*
	Data format		
		Paper	*Are there enough resources to process them? E.g. by scanning and using optical character recognition.*
		PDF – non-searchable	*Are there enough resources to process them? E.g. using optical character recognition.*
		PDF - searchable	*Are there enough resources to process them? E.g. by retrieving information from them.*
		XLS	*Are there enough resources to retrieve information from the XLSs?*
		HTML	*Are there enough resources to retrieve information from the HTML?*
		XML	*Are there enough resources to retrieve information from the XML files?*
	Free or fee is charged?		*Are there enough resources to pay the fee?*
Announcements to include in database			
	Variety of announcement types and templates		*How many announcement types exist in the country? How many templates are used per announcement type?*
	Duplications		*Are there announcements which are published more than once?*
	Framework agreements		*Are both framework agreements and contracts based on framework agreements published? How can they be identified?*
	Unsuccessful tenders		*Are unsuccessful tenders published too? How can they be identified*

	Erroneous notices and their correction	*What is the official procedure if a mistake has been discovered in a notice? How this procedure affects our database?*
Covering the full tender cycle - linking announcements to each other		
	Types of links among announcements	*Do announcements contain direct links to each other and/or are they assigned a tender ID?*
	Reliability of links	*How many announcements have no links despite legal requirements?*
Quality of reported information		
	Formats (for all variables)	*What kind of characters are allowed to be typed in the entry fields of the template? Do some characters conflict with variable definitions (e.g. alphabetic characters in numerical variables)?*
	Missing data (for all variables)	*How many missing values are there for each variable? What is the reason for missing values?*
	Checking nomenclatures	*Are NUTS codes, CPV codes, official registry numbers, postcodes, etc. valid, existing values in the database?*
	Monetary amounts	*Are all amounts valid, reasonable values?*
	VAT	*Is VAT included?*
	Currency	*Are the amounts always given in national currency?*
	Special price constructions	*Do special price constructions appear? E.g. a reserve rate, a price given as an interval, long description instead of one exact value*
	Unit prices	*Are unit prices used? If yes, is the number of units procured given?*
	Outliers (for all variables)	*Are there outliers, unlikely to be true values?*
Identifying actors		
	Procuring bodies	*Are procuring body IDs included in the announcements? If not, is there an official register of public organizations in the country?*
	Bidders/winners	*Are bidder tax IDs included in the announcements? If not, is the official register of companies available?*

Government Transparency Institute

Introduction

Public procurement is a crucial area of public spending due to its high corruption risks, the large amounts involved and public visibility. In spite of the clear importance, very few databases exist which would allow governments and citizens to monitor corruption risks and inefficiencies in public procurement across Europe. Furthermore, even though all EU countries publishing a large amount of micro-level information on public procurement tenders, this information is usually very difficult to analyze systematically because it is fragmented and contains mistakes. We may have tools to acquire information on individual contracts, but we have very limited ways to systematically analyze larger sets of contracts in order to discover regularities and factors which could indicate corruption, collusion, or inefficient public administration.

The authors at the Government Transparency Institute have been working on Hungarian public procurement data since 2011[1]; they developed a cleaned and standardized public procurement database using the Hungarian Public Procurement Authority's website which contains all publicly available public procurement related announcements in Hungary. This report summarizes our experiences regarding the construction of the Hungarian public procurement database.

This report aims at giving a detailed outline of the potential problems and the steps of the correction using Hungarian public procurement data as an example. Two factors make Hungary an ideal example for other countries. On the one hand Hungarian procuring bodies are obliged to provide a wide range of information about their public procurement tenders using a publicly accessible central website; on the other hand, the frequent legal changes and the lack of appropriate control mechanisms in the data generation process result in a database with many diverse and changing mistakes. In sum, the Hungarian database has a wide range of problems frequently occurring in other countries, hence offers many lessons for a wider audience.

This report adds to the slowly emerging open data movement assessing the quality of publicly released government contracting data, increasingly on the micro level: e.g. the Canadian public procurement data quality initiative (https://sites.google.com/site/do101mtl/seao/iqd-1), public procurement database building and data quality projects in the Czech Republic (http://zindex.cz/, http://www.profily.info/), and Open Contracting Data Standard's comparison site on data scope across the globe (http://ocds.open-contracting.org/opendatacomparison/datasets/).

1. The process of public procurement and related announcement types

Figure 1.1 shows schematically the process of public procurement in Hungary specifically, but the basic elements are similar in most countries. When a public organisation decides to purchase a product or service and this purchase is subject to public procurement regulations, the institution publishes a call

[1] As part of the Corruption Research Center Budapest (CRCB) until Summer 2015. An earlier version of this paper has been prepared for the DIGIWHIST (EU Grant Agreement number: 645852) kick-off meeting in Cambridge April 2015.

for tender or it sends invitations to tender to selected companies. National regulation defines which procedure is to be applied and which notices have to be made publicly available in the official public procurement journal and on a central public procurement website.

After the bids are received, the procuring body evaluates them and selects the winner according to the award criteria. The result is published in a contract award announcement. If no valid bids were received or the prices were too high for the institution, an announcement is published about the reason of the failure.

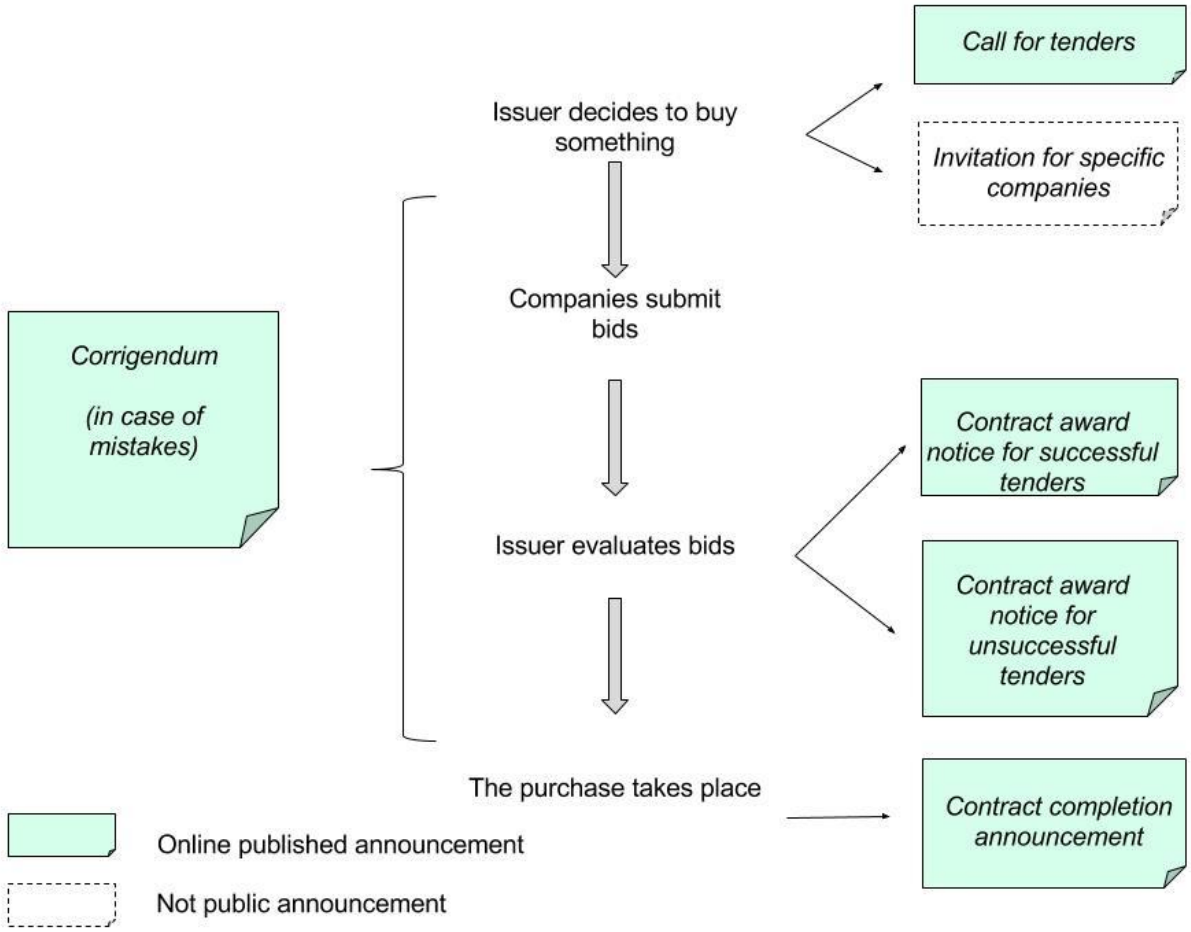
After the contract is fully executed (e.g. the construction is finished), a contract completion notice is published, which contains the final cost of the purchase and other details. In Hungary this type of announcement was only mandatory to publish before 2012.

If the procuring body or the Public Procurement Authority discovers a mistake in an already published announcement during the public procurement procedure, it may publish a corrigendum with the corrected information. However, in the majority of the cases the former erroneous version remains available in the online system.

Contract award notices and call for tender notices have the most important role in building a public procurement database because they contain the most important bits of information e.g. the procuring body name, the winner company name, the value and content of the contract. We use contract award notices as the anchor in the database because they have to be published in every single tender, whereas this is not the case for the other announcement types such as call for tenders. Contract completion notices are also important to include because they are the only source of information on the final result of the projects (e.g. payments made).

It is important to note that one announcement type can have several different templates depending on e.g. the sector or the procedure type. For example, in Hungary 38 different online forms can be used to publish a call for tenders.

FIGURE 1.1: PUBLIC PROCUREMENT AND RELATED DOCUMENT TYPES



2. Problems in general: Which administrative processes generate errors in the databases?

Public procurement related announcements are available on the website of the Hungarian Public Procurement Authority[2] in HTML format. Unfortunately, several difficulties occur when we begin to scrape and parse them in order to build a database. We present below the most common problems which are also present in other countries to varying degrees.

- Instability of the legal environment

The national regulation of public procurement is changing from time to time and it is not always clear which rules – old or new – apply in the transition period. Even if this was clear, mapping the legal

[2] Main page: http://kozbeszerzes.hu/
The announcement search page: http://kozbeszerzes.hu/adatbazis/keres/hirdetmeny/

environment constantly and dealing with the inconsistencies in the database requires a significant amount of effort.

- Lack of protocols/neglect of protocols

The rules of publication of notices are often not well specified or sometimes the submitter of the data does not comply with the rules. E.g. the deadline of publication is violated, some announcements cannot be found online although according to regulation they should be, there are important data missing in the announcements, etc. The quality of data in the online published announcements is not checked and there are no consequences for procuring bodies if the entry fields are blank without a good reason or records are inconsistent. Sometimes this is not an incidental mistake but an explicit method to conceal corrupt transactions and escape scrutiny.

- Lack of pre-defined organisation IDs

The unique official numeric codes of procuring bodies and bidders – such as tax number and State Treasury registry number - are not used. Instead of these the names of the organisations are displayed in manually entered text format. Both the long and short version of the company names can appear and misspelled names are also present. This makes difficult to aggregate the contracts of a company or institution. Another aspect of this problem is that announcements which belong to the same procurement procedure usually do not have a common ID, which would make it possible to link them together.

- Free text fields without restrictions

The raw data retrieved from the online HTML forms are often inconsistent because of the lack of control mechanisms in the entry fields. E.g. letters can be typed in fields which are numeric and values can be entered in various formats (‘1 000’, ‘1.000’, ‘1,000,-Ft’, ‘thousand’). Lots of problems could be traced back to this feature.

- Lack of cross-references in the data entry process

If the same information has to be entered repeatedly in different announcements (e.g. address of the procuring body in the call for tenders and in the contract award notice) it is not sure that exactly the same information is entered, generating additional errors. This problem could be avoided by using well-designed software which fills in these entry fields automatically.

3. Solutions in general

Some of the above mentioned problems are possible to solve automatically during the database building process by following these steps:

1) retrieving information form online public procurement announcements,

2) analysing the distribution of the values and different value formats,

3) correcting the most frequent error types with a combination of automatic codes and manual corrections.

However, missing information typically cannot be imputed after the fact, limiting even the most resourceful data cleaning efforts. In addition, the distribution of error types is such in Hungary that there

are many rare errors each of which affect only few cases, but jointly affect many announcements making correction efforts excessively costly.

We found that the following correction procedure creates the best results: we apply automatic correction to the most frequent mistakes at first, and then we correct manually the rest of the errors to a reasonable extent. This extent might vary greatly for different variables. Later on, manual corrections could be fed into a machine learning algorithm which minimises the need for manual corrections in the future.

In order to maximize auditability in case of a suspected miscalculation, we keep our transformed and corrected variables well-documented; 4-5 versions are stored of the most problematic variables in different phases, including the original, raw, uncleaned version. These versions are generally the following:

1) raw version: exactly as it was in the announcement
2) automatically corrected version
3) manually corrected version: only applied to those cases, where there are data in the raw cells but the automatic correction did not work.
4) final version: standardised format, which contains both the automatically and the manually corrected values

4. Typical problem types and their solutions

4.1. Defining the scope – the maximum of potentially available information

It is important to determine the maximum number of public procurement announcements which are publicly available. This is affected by a number of factors which are specified below.

4.1.1 Changes in regulation

Changes in regulation influence fundamentally the quantity and quality of publicly available information on public procurement procedures so elaborate mapping should precede the actual data collection. Besides desk research it is often necessary to inquire at the national authority or to ask for help from public procurement lawyers because the text of the law does not always indicate clearly the practical realisation of rules.

It is important to know what kind of public procurement procedures exist in a country and what kind of announcements have to be published; hence, what kind of documents we can expect to be available. It is also possible in practice that in spite of the legal obligation some publications are missing. (E.g.: the call for tender announcement has to be published in case of an open procedure but often only the contract award notices can be found.)

4.1.2 Different data formats

The availability of public procurement announcements does not mean necessarily that they are in a format which can be easily processed and used. They can appear a) on paper, b) electronically, but only scanned as a picture, c) as a searchable PDF, d) in online HTML forms, e) in XML format, f) in a coherent, organised downloadable database.

All these formats are possible to process - automatically or manually - but with different level of effort. It should be decided which formats are affordable to deal with. In case of Hungary, online HTML forms were used in the first place, and additional data was retrieved from PDF documents from before 2005.

4.2. Identifying the population of announcements to include in the database

The great majority of information needed to build a public procurement database is contained in a few major announcement types. The contract award notice and call for tenders are the most important types, but correction notices and amendments are also essential for a high quality database.

The quality of the database can be increased by filtering out non-relevant, unnecessary and erroneous announcements. Duplications, erroneous announcements, framework contracts and unsuccessful procedures were filtered out in the Hungarian public procurement database.

4.2.1 Identifying announcement types and related templates

Each announcement type appears under different names on the search page of the Hungarian Public Procurement Authority. This is partly due to the numerous different templates which are in use and partly because of other, unclear reasons. Careful consideration is needed when deciding which announcements should be included in the database.

The variety of templates has to be taken into consideration also when the information retrieving programs are developed because different templates contain the same information at different sections and even important keywords might differ in them.

FIGURE 4.2.1.1: VARIATIONS FOR 'CALL FOR TENDER' ON THE PUBLIC PROCUREMENT AUTHORITY'S WEBSITE

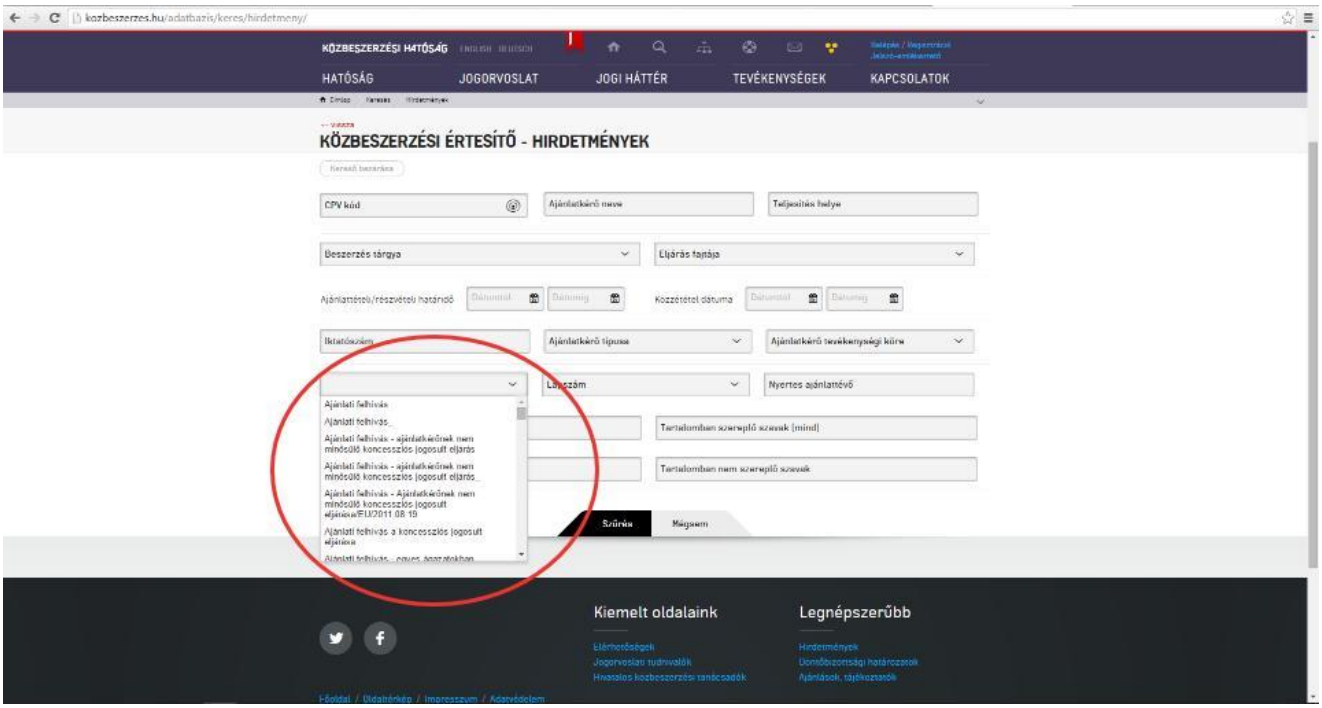
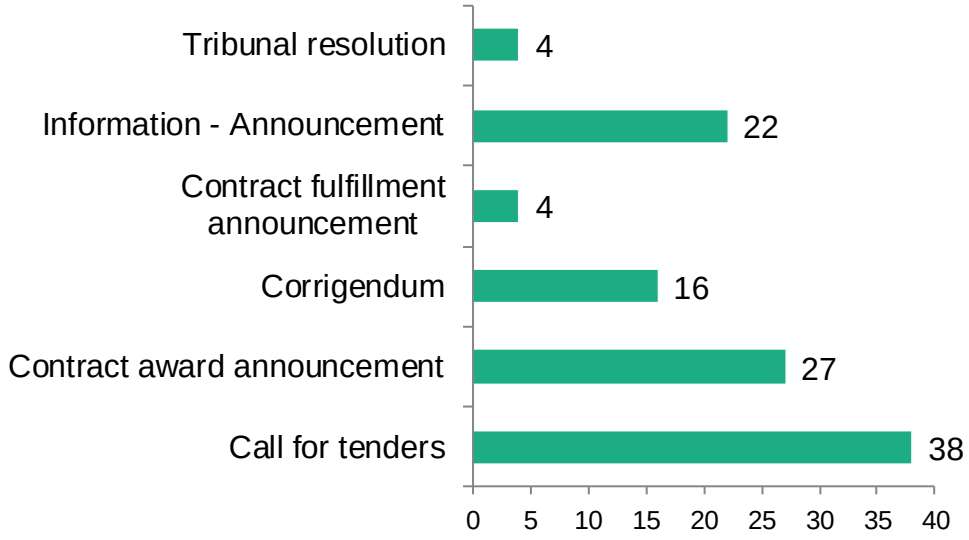


TABLE 4.2.1.1: EXAMPLES FOR DIFFERENT VARIANTS OF CALL FOR TENDERS

Hungarian name	English translation
Ajánlati felhívás nyílt eljárásra EUHL - hatályon kívüli	Call for tenders, open procedure, EUHL - repealed
Ajánlati felhívás nyílt eljárásra KÉ - hatályon kívüli	Call for tenders, open procedure, PP Bulletin - repealed
Ajánlati felhívás KÉ	Call for tenders PP Bulletin
Ajánlati felhívás_ KÉ	Call for tenders_PP Bulletin
Részvételi felhívás KÉ	Call for participation PP Bulletin
Részvételi felhívás/EU/2011.08.19. EUHL	Call for participation/EU/2011.08.19. EUHL
Részvételi felhívás - egyes ágazatokban_ KÉ	Call for participation in specific branches_PP Bulletin
Eljárást megindító felhívás - 121. § (1) bekezdés b) pontja/KÉ/2013.07.01 KÉ	Call for starting a procedure - 121. § (1) paragraph b) /PP Bulletin/2013.07.01

FIGURE 4.2.1.2: NUMBER OF VARIANTS OF ANNOUNCEMENT TYPES ON THE PUBLIC PROCUREMENT SEARCH PAGE, 2005-2012



Calculations: CRCB/GTI

4.2.2 Duplications

In some cases, the same information may be reported more than once on the public procurement website, using different announcement types. This means that the same phase of tendering is reported more than once. For obvious reasons we have to identify any such duplication and remove them. In Hungary, this double reporting happens mostly if an announcement has to be published not only in the national official journal but also in the Official Journal of the EU.

According to the Hungarian regulation, all announcements published in the EU's Official Journal have to be published also in the national journal (Public Procurement Bulletin) within five working days. We filtered out the EU Official Journal version of the announcements. This information – whether the announcement was published in the national journal or in the journal of the EU – is given in the HTML form of the announcement (See Figure 4.2.2.1).

TABLE 4.2.2.1: CONTRACT AWARD ANNOUNCEMENTS BY SOURCE, 2005-2012

	2005	2006	2007	2008	2009	2010	2011	2012	Total
EU Official Journal	0%	0%	0%	29%	33%	20%	6%	31%	19%
National Journal	100%	100%	100%	72%	67%	80%	94%	69%	81%
N	5591	9975	7425	14496	23541	31334	18761	19337	130460

Source: calculations by CRCB/GTI

FIGURE 4.2.2.1: THE PLACE OF THE 'EU'S OFFICIAL JOURNAL INFORMATION' IN A HUNGARIAN CONTRACT AWARD ANNOUNCEMENT

Note: EUHL stands for EU Hivatalos Lapja, which means EU's Official Journal in Hungarian. 'Hirdetmény típusa' is 'Type of the announcement'.
The URL of this announcement: http://kozbeszerzes.hu/adatbazis/mutat/hirdetmeny/portal_0150_2013/

4.2.3 Corrections

Some of the information published on the public procurement website turns out to be incorrect later. If the Public Procurement Board (PPB) is informed about the error, a separate correction notice is published about this. However, in some cases also the original announcement is re-published with corrected data. (But the original, erroneous announcement remains available too.) It is not clear which factors determine whether this re-publication happens or not.

Although the elimination of errors is necessary, the ambiguous practices used by the PPB raise a number of questions in terms of transparency and legal certainty. On one hand the separate correction notices contain the registry number of the original, erroneous announcement so it is possible to link together these two documents. On the other hand, if the original announcement is re-published, it is not indicated in either document. Thus, it could happen easily that our database contains both the

correct and incorrect version. We can never be sure if an announcement has a corrected version somewhere else.

Various methods can be applied to find the corrections and the corrected, republished announcements. The following three were applied in case of Hungary:

1) Firstly, all correction notices were downloaded and the ID numbers of the original, incorrect announcements were retrieved from them. We searched for the corrected, republished version of these announcements using the a) short description of the contract, b) the name of the procuring body and c) the date of publication. If a corrected version was found, we used that one and we filtered out the incorrect version. If we did not find a corrected version, we corrected the original incorrect version manually using the information in the correction notice.

2) Secondly, we searched specific keywords in the contract award notices such as 'corrig*', 'correct*' and 'amend*' because these words often appear in the corrected, new version of announcements. If we found a corrected, republished announcement, we tried to find the related incorrect version and the correction notice, if there was any.

3) Thirdly, there is a section in all announcements where the ID numbers of formerly published, related announcements are listed. If a contract award notice contains a reference to another contract award notice, it is highly possible that the former one is an incorrect version and the latter one is the corrected, republished version. However, this cannot be decided automatically, only by human labour.

We filtered out only 128 contracts[3] because of the fact that they were corrected later and we wanted to keep only the correct versions. Besides this, we manually corrected the incorrect versions if there were no corrected versions published.

The relatively small number of cases suggests that neglecting this problem would not have significantly harmed the quality of the database. However, some errors were substantial, (e.g. the incorrect contract value was 1 billion instead of 1 million). Unfortunately, even a single large error can lead to less robust analyses.

4.2.4 Unsuccessful tenders

Not every call for tender results in a successful contract award: sometimes tendering fails because there is no bidder or submitted prices are too high for the procuring body. In Hungary, an announcement is published even in this case, but the place of the winner's name and address is blank and the cause of the failure of contracting is indicated in the announcement after this text: "V.2.2) If the procedure is unsuccessful, or the cause of unsuccessful contracting"[4]

TABLE 4.2.4.1: RATE OF SUCCESSFUL AND UNSUCCESSFUL CONTRACT AWARD ANNOUNCEMENTS, 2005-2012, %

	2005	2006	2007	2008	2009	2010	2011	2012	Total
Unsuccessful	12	12	7	9	10	13	10	10	11
Successful	88	88	93	91	90	87	90	90	89
N	5413	9455	6888	12696	21130	28630	17443	16882	118537

Source: calculations by CRCB/GTI

[3] 4 contracts in 2009, 81 contracts in 2010 and 43 contracts in 2011

[4] The original Hungarian text: "V.2.2) Ha az eljárás eredménytelen, illetve szerződéskötésre nem kerül sor, ennek indoka"

4.2.5 Framework agreements

Public institutions and companies may sign a framework agreement which defines the commodity or service the procuring body wants to buy, the maximal value of the transaction and the end date of the contract, but it remains flexible regarding the exact quantity and total price. These agreements do not result in an immediate payment, but the actors should contract again and again based on the framework agreement - up to the maximum value indicated in the agreement. These new contracts are 'contracts based on framework agreement'. We should not take into account both the framework agreement and the contracts based on it because this way we would count the price twice. So we filtered out framework agreements from the Hungarian database.

Different approaches can be applied to find the framework agreements:

1) In some years, explicit information is available in the announcements regarding framework agreements.
2) We looked for extremely high contracting values in the database and checked them manually – especially the description of the content of the contract. These outliers were often framework agreements.
3) We searched specific keywords in the short description of the announcement (e.g. "framework"). The results cannot be taken automatically for framework agreements; they have to be checked manually.

TABLE 4.2.5.1: RATE OF FRAMEWORK AGREEMENTS BY YEAR, 2005-2012

	2005	2006	2007	2008	2009	2010	2011	2012	Total
Framework agreements	0%	3%	7%	6%	5%	2%	2%	6%	4%
Not framework agr.	100%	97%	93%	94%	95%	98%	98%	94%	96%
N	5413	9455	6888	12696	21130	28630	17443	16882	118537

Source: calculations by CRCB/GTI

4.3. Linking announcements to each other

The complete documentation of a public procurement procedure consists of a few different types of announcements that are separate documents and are not linked to each other in a clear, unambiguous way on the public procurement website, because of the lack of a unique tender ID. Establishing a link between announcements that belong to the same procurement procedure is a critical point of building the public procurement database because only several announcements together can provide all information which is necessary for corruption risk analyses.

Our goal was to solve this deficiency by elaborating a method to connect call for tenders, contract award notices, contract completion and amendment notices and corrigenda together.

We aimed at connecting the following types of announcements:

a) call for tenders (CfT) – contract award notices (CAN);
b) corrigendum (C) – call for tenders (CfT);
c) corrigendum (C) – contract award notices (CAN);

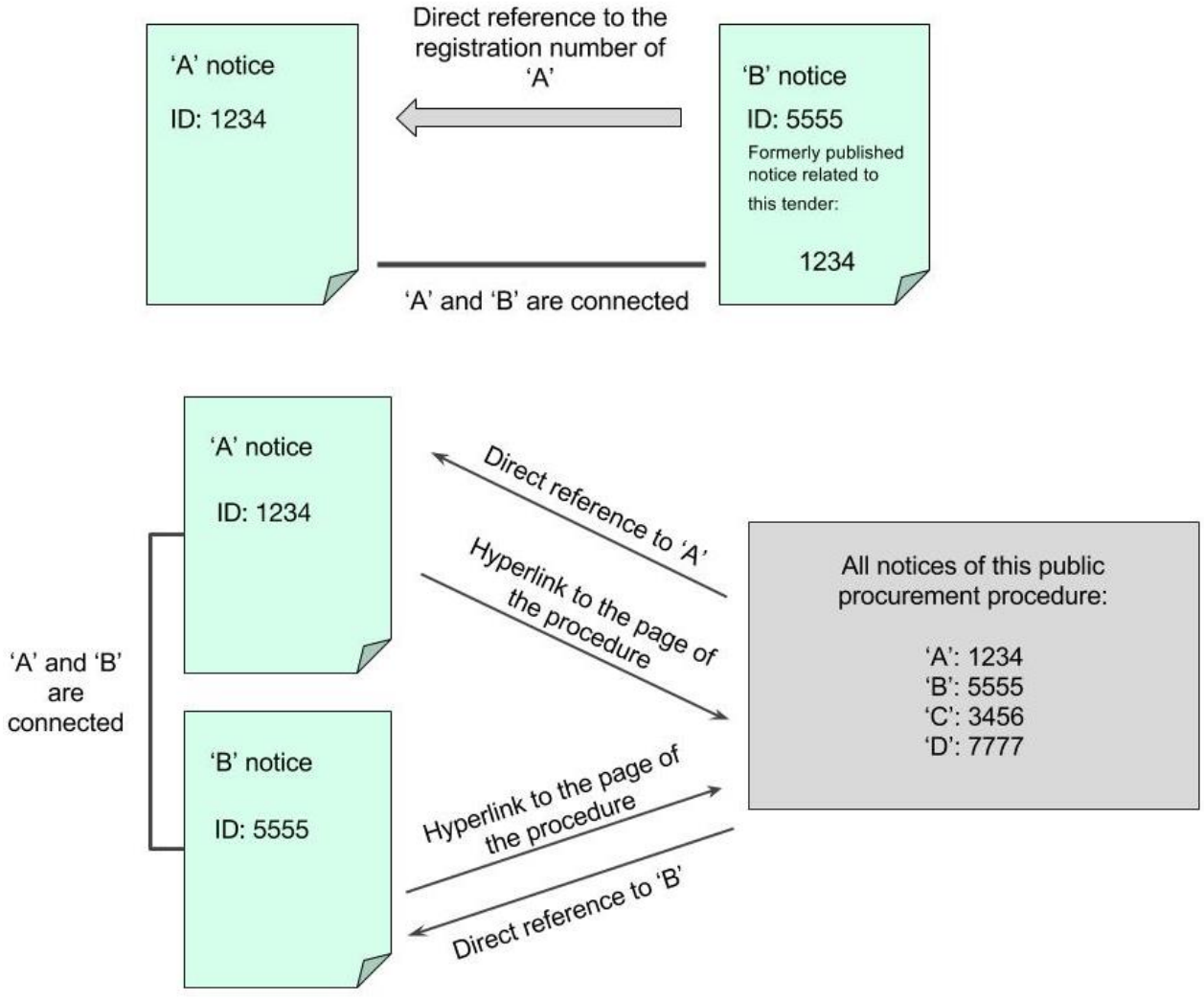
d) contract amendments (CA) – contract award notices (CAN);

e) contract completion notice (CC) – contract award notices (CAN).

Four different approaches were tried with varying rates of success. Eventually only the most reliable linking method was used. However, the other methods might be used more successfully in other countries, so we present all of them below.

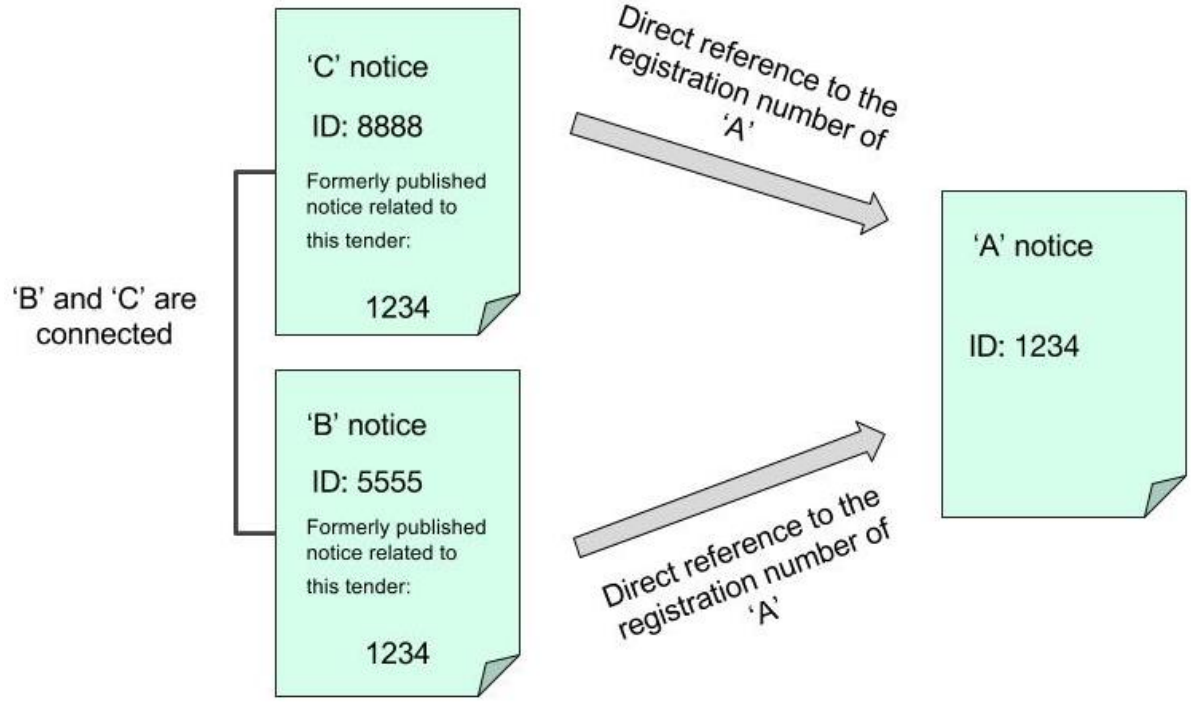
a) Direct connection - This method is based on unique announcement IDs. In this case the registration number of the formerly published document (e.g. the call for tenders) is directly indicated in the latter document (e.g. in the contract award notice) so they can be linked together. Since 2013 also a new form of direct connection is available in Hungary: each public procurement procedure gets a unique ID and a separate webpage is created for them where all related announcements are listed. However, it often turns out that the list of related announcements is incomplete whether the procedure has a separate summary page or not.

FIGURE 4.3.1: TWO TYPES OF DIRECT CONNECTION BETWEEN ANNOUNCEMENTS



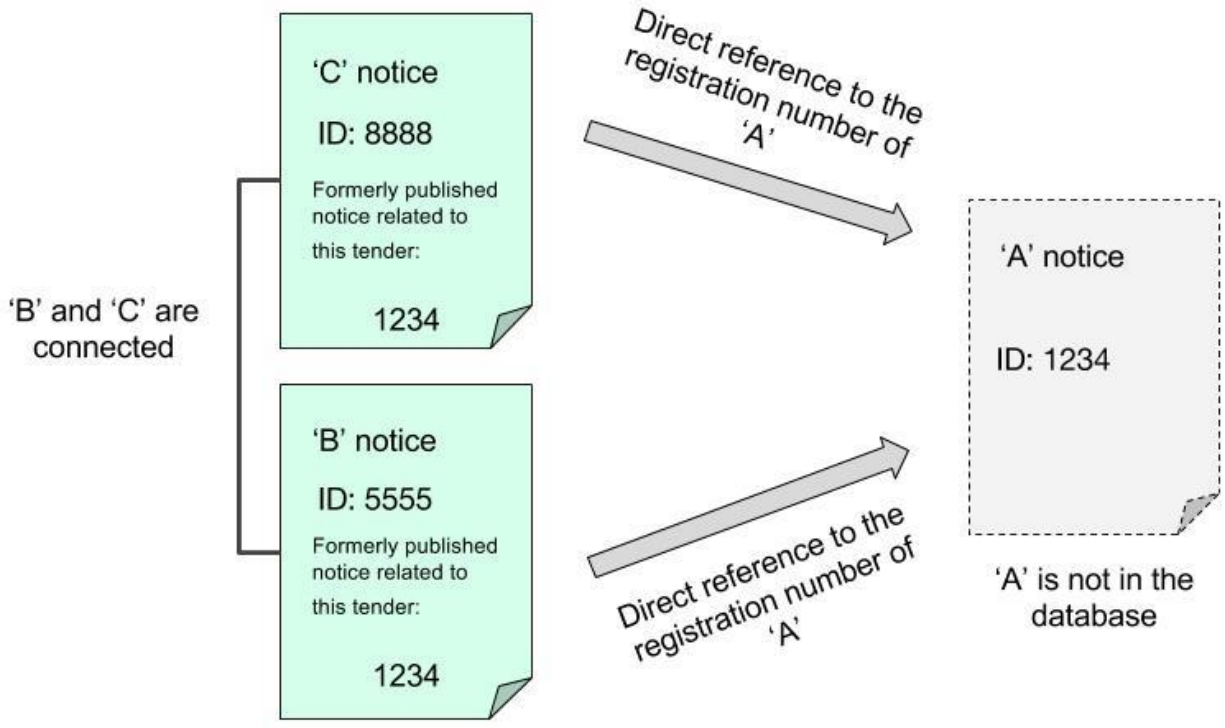
b) Indirect connection – This method is also based on registration numbers. In this case there is not any direct reference in the announcements to each other but if the registration number of ‘A’ announcement appears in both ‘B’ and ‘C’ announcements, we can link ‘B’ and ‘C’ together. All three announcements are in our database. (This is important because the third connection method is different in this regard.)

FIGURE 4.3.2: INDIRECT CONNECTION BETWEEN ANNOUNCEMENTS



c) Indirect phantom connection – This method is very similar to the indirect connection method but in this case we do not have ‘A’ in our database, we only know that an announcement should exist with the registration number which is mentioned in ‘B’ and ‘C’ announcements. This method leaves us with less chance to check whether the connection can be confirmed or not.

FIGURE 4.3.3: INDIRECT PHANTOM CONNECTION BETWEEN ANNOUNCEMENTS

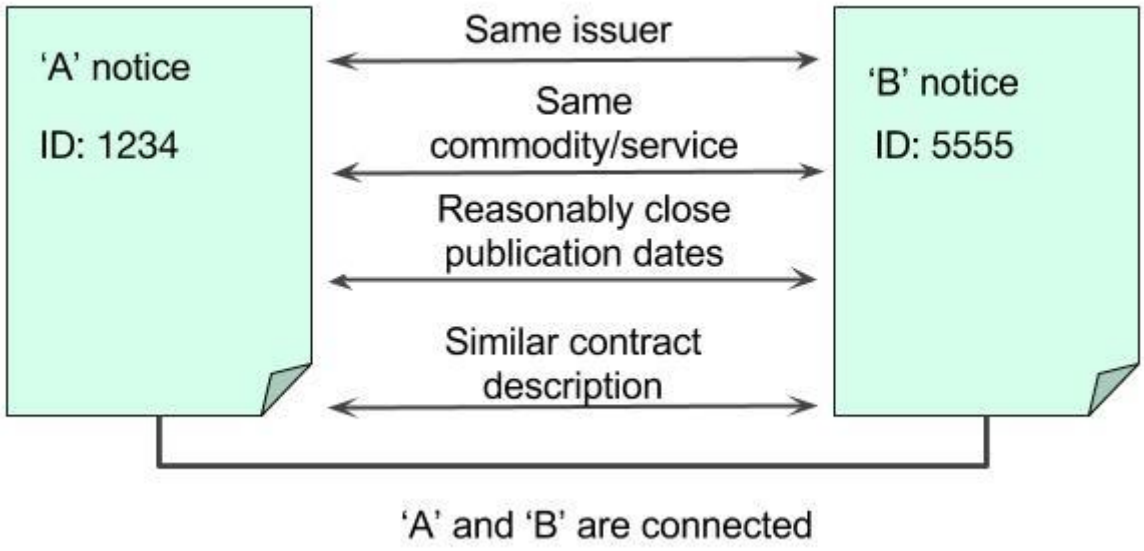


d) Proxy connection – In this case we cannot use registration numbers because the analysed announcement does not mention any. Instead of registration numbers we try to find announcements that contain the same procuring body, the same commodity registration numbers and their publication dates are reasonably close to each other. Another important criterion is that the call for tenders should have been published earlier than the contract award notice. We also used the text description of the subject of the contract: we analysed the similarity of these descriptions in two announcements. We defined the following variable:

$$rmatch = \frac{(l - r) * 100}{l}$$

where 'l' is the length of the longer description and 'r' is the Damerau-Levenshtein distance between the two descriptions. We used 95% as a threshold so if rmatch reached 95%, we accepted the connection.

FIGURE 4.3.4: PROXY CONNECTION BETWEEN ANNOUNCEMENTS

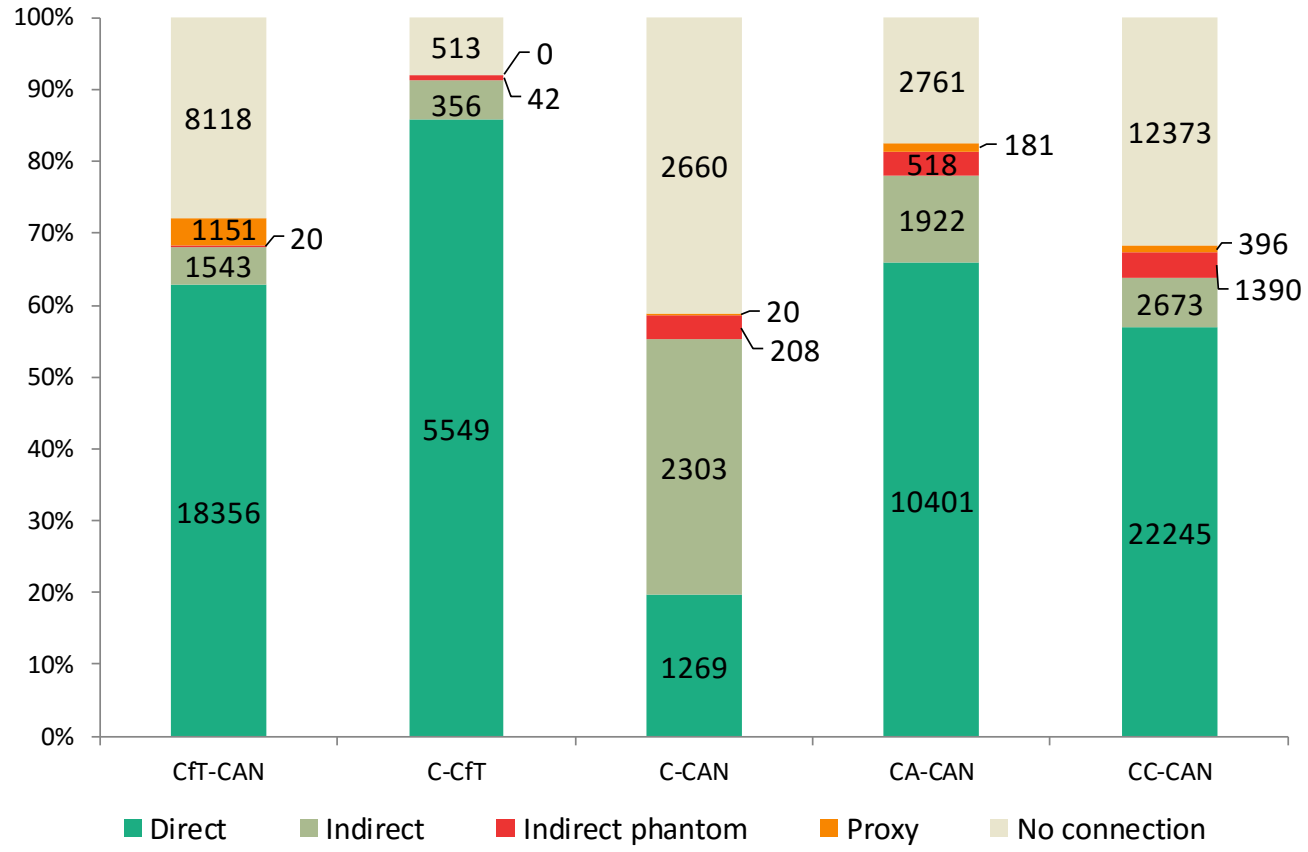


It is important to note that indirect, indirect phantom and proxy processes are exposed to type II errors: although their conditions are strict, false connections might be established.

Figure 4.3.5 shows the success rates of the different connection methods (direct, indirect, indirect phantom and proxy) in case of the five types of document pairs. The variance of the success rates is rather significant: in case of C-CAN pairs, the success rate is lower than 60%, while in case of C-CAN pairs it is above 90%.

In case of the most important relations (CfT-CAN and CC-CAN), 30% of the connections are missing, meaning that in 30% of the CfTs it is not known who won which procurement and what is the final contract value. Indirect and indirect phantom connection processes perform poorly: in most of the cases only 10% of the missing connections can be established with these methods. The proxy method works only in case of CfT-CAN, C-CAN and CA-CAN pairs.

FIGURE 4.3.5: SUCCESS RATES OF DIFFERENT CONNECTION METHODS BY ANNOUNCEMENT TYPES, 2005-2012



Source: calculations by CRCB

Table 4.3.1 shows a more detailed illustration of the success rates of the four connecting methods. We used call for tenders published between 2005 and 2012 in the national journal and we tried to identify their contract award notices with the above mentioned methods. We used the direct method at first and we succeeded in 63% of all calls for tenders. After this we applied the indirect method **to the rest of the calls for tenders**. This increased the rate of linked calls for tenders by 5 percentage point. Thirdly, we applied the indirect phantom method but this hardly increased the success rate.

The fourth method – the proxy connection – is independent from the former three methods. We applied it to all of the calls for tenders and we were able to connect 41% of the call for tenders to contract award notices.

TABLE 4.3.1: CALL FOR TENDERS WITH AND WITHOUT CONTRACT AWARD NOTICES BY CONNECTION METHODS, 2005-2012, %

	Direct		Indirect		Indirect phantom		Proxy	
	pcs	%	pcs	%	Pcs	%	pcs	%
Not connected	10832	37%	9289	32%	9269	32%	17230	59%
Connected	18356	63%	19899	68%	19919	68%	11958	41%
Total	29188	100%	29188	100%	29188	100%	29188	100%

Source: calculations by CRCB

As an indicator of the quality of matching we examined the description texts' similarity (the above defined 'rmatch' variable) in two groups of connected announcements: in the first group there are also a direct link between announcement pairs while in the other group there is no direct link, only indirect or proxy connection. It is clear that the similarity is significantly higher if there is a direct connection between two announcements, so it is more probable that the announcements are really connected to each other and this is not a false positive match.

TABLE 4.3.2: AVERAGE VALUE OF 'RMATCH' IN CFT-CAN PAIRS, 2005-2012

	Mean	St.Dev.	N
'rmatch' without direct connection	45,95	0,21	22260
'rmatch' with direct connection	80,5	0,21	17306

Source: calculations by CRCB/GTI

As neither the two indirect methods, nor the proxy connection process have led to significantly more document connections, finally we decided to use only the direct method to decrease the risk of creating invalid connections.

4.4. Retrieving and correcting the necessary information from announcements

4.4.1 Typos

The online public procurement templates are filled up with information by procuring bodies and the content is not checked by automatic mechanisms. The free text fields allow procuring bodies to type in various kinds of erroneous data. These problems come up mostly when we try to match a value to some kind of official codes (CPV codes, NUTS codes, company ID-s etc.) or we are looking for a connection between two or more announcements. As a general rule we may assume that every variable has to be corrected and standardised before we analyse them.

4.4.2 No standard data formats in monetary value variables

Several data format problems can be traced back to the frequent use of free text fields in the online templates of the Public Procurement Authority. The most complicated cases emerge in connection with monetary variables: the estimated and the actual price of the commodity or service. We illustrate the

problems and the correction methods regarding the final contract value. The potential problems could be classified into the following types.

a) Alphabetic versus numeric characters: Because it is possible to enter not only numeric but also alphabetic characters, the procuring bodies are not forced to give one exact value with numbers. This means on the one hand that they can type 'thousand' instead of '1000', on the other hand they are also allowed to write a long, detailed description of the contract value – which is actually not a rare case in Hungary.

b) Different formats: Even if the procuring bodies type in the price with numbers, they often apply a variety of punctuation and space-based formats.

c) Unit prices: The price is often given as a unit price for one year/month/piece of commodity or service. If the quantity or period of time is given too, it is possible to compute the total price. However, the concrete method should be tested carefully manually because an incorrect method can easily create several new mistakes.[5] We applied two approaches to identify unit prices: 1) we looked for specific keywords in the announcement (e.g. '/month', '/person', '/kW'); 2) we assumed that very small values are unit prices. The threshold should be defined empirically and it should be tested manually. In Hungary it was ~200EUR.

d) VAT: The values can be indicated with or without VAT in the announcements, so this should be standardised. If there is no unambiguous information whether the price contains VAT or not, we may assume that the more frequent case applies.

e) Currencies: The contract values may be indicated in different currencies; this should be standardised too. We used monthly average exchange rates from the webpage of the Hungarian National Bank.

f) Outliers: The extremely low and high values should be analysed manually as they might reveal systematic errors. Too small values might be unit prices, too high values might be typos.

g) Special constructions: it could be worth retrieving information about special price constructions such as 1) values given as an interval or 2) reserve rates. These cases might be rare but they might be used as an explicit method to raise prices after the company won the tender with an unreasonably cheap offer.

In the following we present a few real examples to illustrate the variety of the content of the "Value of the contract" part of the announcements.

[5] We did not use the unit price multiplying method at last because it generated more mistakes than the number of values it corrected.

TABLE 4.4.2.1: EXAMPLES FOR VARIOUS FORMATS OF MONETARY VALUES

Hungarian	English
8443200	8443200
15.000.000,-	15.000.000,-
44 700 063	44 700 063
12950981,2	12950981,2
2.000.000,- Ft	2.000.000,- Ft
36 hónapon keresztül, 2125 ±25% főre 4000 Ft/fő	36 months, 2125 ±25% persons, 4000 HUF/person
112 500 Ft/szerződés, összesen várhatóan 39 600 000 Ft	112 500 HUF/contract, total expected value: 39 600 000 HUF
9,55 Ft/Wh (84 GWh/év-re)	9,55 HUF/Wh (84 GWh/year)
35,7 Ft/db	35,7 HUF/pcs
2006. évre 39 000 000 Ft + áfa + 30% szükség esetén	For 2006: 39 000 000 HUF +VAT +30% if necessary
bruttó 47.201.712,-	gross 47.201.712,-

As various kinds of correcting methods are applied to the values, it is highly recommended to document every phase of the process and save all transformed variables, so that we could go back if a mistake is discovered. We created variable names the following way in general:

[announcement type]_[content of the variable]_[version number],

E.g: ‘can_c_v_0’, where ‘can’ stands for contract award notice, ‘c’ stands for contract, ‘v’ stands for value and ‘0’ indicates that this is the first, raw version of the variable. We used the following version numbers in general:

- 0: raw, string version
- 1: automatically processed numeric values
- 2: manually corrected values - we create this values only in those cases where there is a string version but the automatic correction did not succeed. In other cases this variable has a missing value. We also indicate in this variable with a special value (e.g. -1), if the raw variable indicates unit price
- 3: this variable contains unit prices. E.g. if the raw version is 10000HUF/month, then can_c_v_3=10000.
- fv: this is the final version of the variable.

If it is necessary to transform further the variables, additional suffixes are used, e.g. _nhuf for ‘net value, in HUF’, or _nhuf_nunit for ‘net value, in HUF, not unit price’.

TABLE 4.4.2.2: EXAMPLES FOR THE PHASES OF THE CLEANING PROCESS, 2005-2012, %

currency [6]	VAT[7]	raw	automatically processed	manually corrected	well-defined unit prices	final version	net value in HUF	net, HUF, no unit price
HUF	no	8 400 000	8400000			8400000	8400000	8400000
HUF	no	13.920.00,-		1392000		1392000	1392000	1392000
HUF	no	49.800.000,-/year		-1	49800000	747000000[8]	747000000	747000000
HUF	no	6 000	6000			6000	6000	

Source: calculations by CRCB/GTI
Note: -1 value in the manually corrected column means: unit price
-2 value in the manually corrected column means: missing or we cannot decide

As Table 4.4.2.3 demonstrates, monetary variables are possible to correct in most cases: in Hungary, 87 percent of contract award notices contain some kind of information about the value of the contract and we were able to retrieve correct, standardised values from 81 percent of announcements.

TABLE 4.4.2.3: RATE OF MISSING AND VALID VALUES OF THE CONTRACT VALUE VARIABLE IN DIFFERENT PHASES OF THE CLEANING PROCESS, 2005-2012

	raw	automatically processed	manually corrected	well-defined unit prices	final version	net value in HUF	net, HUF, no unit price
missing	13%	18%	98%	99%	16%	17%	19%
valid	87%	82%	2%	1%	84%	83%	81%
N	118537	118537	118537	118537	118537	118537	118537

Source: calculations by CRCB/GTI

4.5. Identifying actors

The procuring bodies and the bidder companies are indicated only with text information in Hungarian procurement notices. This practice causes many problems because a specific company may appear under several differently spelled names: short name, long name, misspelled name etc. That is why identifying the actors is an essential part of the database building process. We solved this problem by matching standard IDs to the actors from official registers. We acquired these databases from more different sources:

a) The company register can be purchased from the Hungarian Central Statistical Office for the years 2009-2012.

b) The IDs of public institutions can be acquired from the Hungarian State Treasury free of charge

c) We retrieved company IDs from company information webpages automatically

We applied the following steps in the matching process:

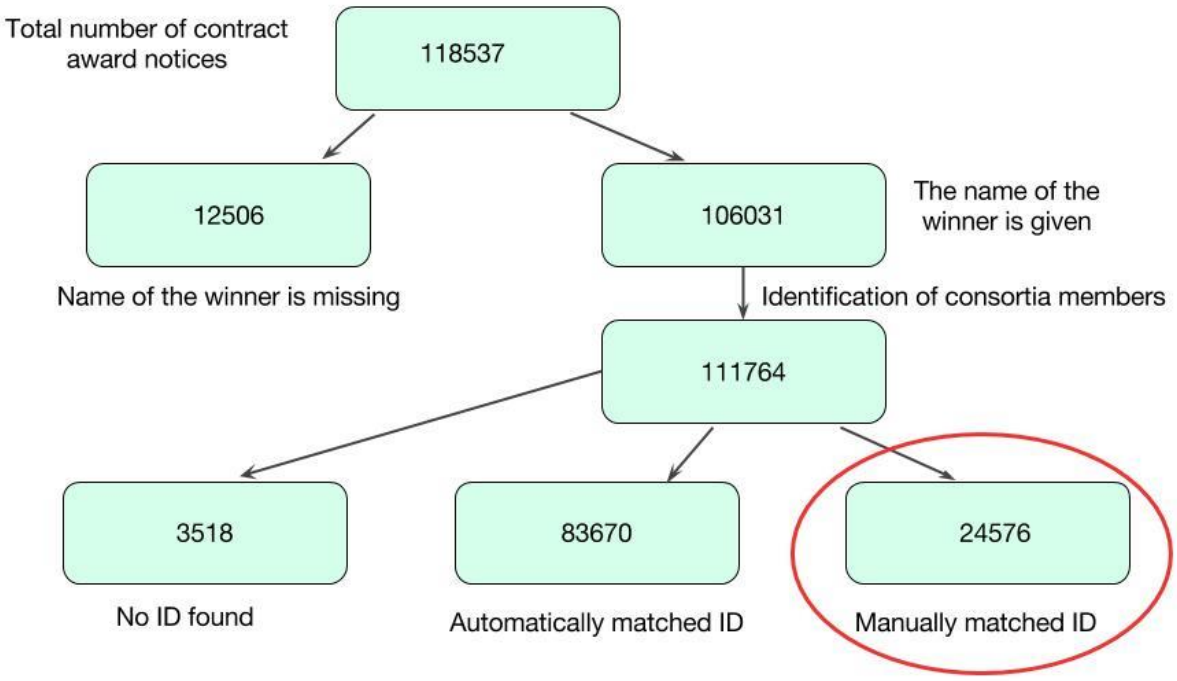
[6] The currency was indicated in the announcement and it is stored in the database.
[7] The VAT information was indicated in the announcement and it is stored in the database.
[8] It was indicated in the announcement that the contract will be effective for 15 years. This information is also stored in a variable in the database.

1) We identified cases where a consortium won and we retrieved the names of the members of the consortium. Consortia can be detected by searching specific keywords such as ‘consortium’ and ‘joint bidders’.
2) We standardized company names and addresses as much as possible by for example translating abbreviations such as ltd into limited company.
3) We matched automatically an ID to the winner company, if the company name was completely identical in the announcement and in the company register.
4) We matched an ID to the winner company, if the similarity between the names in the announcement and in the company register was above a previously defined threshold. We defined this threshold empirically: different values were tried and the rate of correctly and incorrectly matched IDs and the rate of unmatched company names were also tested.
5) We listed the remaining unmatched company names and searched them manually on company information webpages or on the web in general.

There are 118,537 contract award notices in the Hungarian database between 2005 and 2012 but in 12,506 cases there are no company names in them at all. If we separate consortia members, we get 111,764 company names to which we should find the correct official ID.

IDs could be matched to 97 percent of these company names, but it is important to stress that 22 percent of the matches were made manually because the automatic methods did not find the correct ID. Another problem occurred during the automatic ID-matching: sometimes it connected incorrect IDs to the company names, so a systematic check should be part of the method too.

FIGURE 4.5.1: THE PHASES OF IDENTIFYING WINNING COMPANIES, 2005-2012, %



Source: calculations by CRCB/GTI

Government Transparency Institute

Summary

As a summary of the report we would like to condense the information we presented in detail. The following checklist may help assessing and planning the roadmap of building a database from public procurement data in any country where these are publicly available.

TABLE S.1: CHECKLIST

Data availability			
	Regulatory framework		
		Thresholds	*Above which value threshold is it compulsory to publish public procurement procedures?*
		Special sectors e.g. defense	*To which sectors do special regulations and exemptions apply?*
	Data format		
		Paper	*Are there enough resources to process them? E.g. by scanning and using optical character recognition.*
		PDF – non-searchable	*Are there enough resources to process them? E.g. using optical character recognition.*
		PDF - searchable	*Are there enough resources to process them? E.g. by retrieving information from them.*
		XLS	*Are there enough resources to retrieve information from the XLSs?*
		HTML	*Are there enough resources to retrieve information from the HTML?*
		XML	*Are there enough resources to retrieve information from the XML files?*
	Free or fee is charged?		*Are there enough resources to pay the fee?*
Announcements to include in database			
	Variety of announcement types and templates		*How many announcement types exist in the country? How many templates are used per announcement type?*
	Duplications		*Are there announcements which are published more than once?*
	Framework agreements		*Are both framework agreements and contracts based on framework agreements published? How can they be identified?*
	Unsuccessful tenders		*Are unsuccessful tenders published too? How can they be identified*
	Erroneous notices and their correction		*What is the official procedure if a mistake has been discovered in a notice? How this procedure affects our database?*
Covering the full tender cycle - linking announcements to each other			

	Types of links among announcements	*Do announcements contain direct links to each other and/or are they assigned a tender ID?*
	Reliability of links	*How many announcements have no links despite legal requirements?*
Quality of reported information		
	Formats (for all variables)	*What kind of characters are allowed to be typed in the entry fields of the template? Do some characters conflict with variable definitions (e.g. alphabetic characters in numerical variables)?*
	Missing data (for all variables)	*How many missing values are there for each variable? What is the reason for missing values?*
	Checking nomenclatures	*Are NUTS codes, CPV codes, official registry numbers, postcodes, etc. valid, existing values in the database?*
	Monetary amounts	*Are all amounts valid, reasonable values?*
	VAT	*Is VAT included?*
	Currency	*Are the amounts always given in national currency?*
	Special price constructions	*Do special price constructions appear? E.g. a reserve rate, a price given as an interval, long description instead of one exact value*
	Unit prices	*Are unit prices used? If yes, is the number of units procured given?*
	Outliers (for all variables)	*Are there outliers, unlikely to be true values?*
Identifying actors		
	Procuring bodies	*Are procuring body IDs included in the announcements? If not, is there an official register of public organizations in the country?*
	Bidders/winners	*Are bidder tax IDs included in the announcements? If not, is the official register of companies available?*