A Novel Framework to Describe Technical Accessibility of Open Data

Jolon Faichney and Bela Stantic
School of Information and Communication Technology
Griffith University
Gold Coast, Australia
email: {j.faichney,b.stantic}@griffith.edu.au

***Abstract*—Open Data is a recent and important movement that has economic, social, and political benefits. Despite a lot of attention in literature there are still limitations with the existing Open Data frameworks in describing technical accessibility of Open Data. In this paper, at first, we review the emergence of Open Data and the current state of frameworks and standards. We also describe our progress and findings working with Open Data at the local, state, and federal level in Australia. We then present a new Open Data Accessibility Framework (ODAF), which more completely defines levels of Open Data accessibility, guiding data custodians to make data more accessible for Open Data consumers.**

Keywords-Open Data; Open Government; Case Study; Framework.

I. INTRODUCTION

Open Data is a relatively recent movement, with the Unites States launching its Open Data portal in 2009 and the United Kingdom in early 2010 [1]. Open Data is a broad term, which has been defined as "accessible at marginal cost and without discrimination, available in digital and machine-readable format, and provided free of restrictions on use or redistribution" [1]. Open Government Data is a subset of Open Data, however Kloiber [2] states that the majority of uses of the term "Open Data" is used synonymously for "Open Government Data".

The United Kingdom has led the way in implementing and utilising Open Data being ranked number 1 in the world in both the Open Data Barometer [3] and the Global Open Data Index [4].

Due to the relative recentness of Open Data there are only several attempts to define frameworks for Open Data including the Open Definition (2005) [5], Sunlight Principles (2010) [6], Tim Berners-Lee's 5-star Linked Open Data (2010) [7], and Open Data Certificates (2013) [9]. Open Data Certificates [9] currently represents the most comprehensive framework combining three previous frameworks into four levels of Open Data publishing quality.

In Section 2, we present the widely accepted existing Open Data frameworks. In Section 3, we provide an overview of the current level of Open Data support and collaboration at the Local, State, and Federal levels in the City of Gold Coast region. In Section 4, we describe our experiences working with the City of Gold Coast outlining issues and challenges with the current frameworks. In Section 5, we propose and present a new framework describing the technical accessibility of Open Data. In Section 6, we discuss the challenges to adopting the proposed Open Data Accessibility framework. In Section 7 we present our conclusions and proposed future work.

II. BACKGROUND

A. Open Knowledge Definition

Underpinning the majority of Open Data definitions is the Open Definition provided by the Open Knowledge Foundation, now at version 2.0, which states: "*Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness*" [10].

The Open Definition is broad, primarily focussing on the licensing of Open Data rather than the technical aspects.

B. Sunlight Foundation Open Data Principles

In 2010, the Sunlight Foundation defined 10 principles of Open Data (extending the previous 8 Sebastopol Principles): Completeness, Primacy, Timeliness, Ease of Physical and Electronic Access, Machine readability, Non-discrimination, Use of commonly Owned Standards, Licensing, Permanence, and Usage costs [6].

Many of the Sunlight Foundation principles are now covered in the Open Definition 2.0, specifically the last five principles listed above.

It's worth noting that the first two principles of Completeness and Primacy show the prioritising that the complete, raw, original data is made available. This is an important priority for Open Data as it means that the public has access to the original data. However, our experience discussed in the next section shows that non-raw, processed data can be of benefit for Open Data adoption. The Sunlight Foundation principles do not promote processed data other than making the data available in open, machine-readable formats.

C. 5-star Linked Open Data

Based on our experience, raw, unprocessed data can make Open Data less accessible. Tim Berners-Lee introduced the 5-star Linked Open Data framework with an emphasis on technical accessibility [7]. Each level makes the

 ISBN: 978-1-61208-445-9

data more accessible to applications. The five levels of the Linked Open Data framework are shown below:

1. Make the data available on the web in any format with an open license
2. Make it available as structured, computer-readable data (not in image or PDF formats)
3. Use non-proprietary formats such as CSV and XML
4. Use URIs within data so that other websites can point to resources
5. Link data to other data to provide context

The 5-star framework puts an important focus on technical accessibility. Open Data which does not have inherent links, only has to satisfy the first 3 levels. The 3rd level stipulates that the data must use non-proprietary (or open) formats. This is already covered in the Sunlight Foundation and Open Definition 2.0. However, it is important to note that the 5-star framework has emerged from Berners-Lee's work on linked data, the influence can be seen in the 4th and 5th levels which centre around linked data . Therefore the 5-star framework doesn't provide a greater level of detail in technical accessibility apart from adding levels for linked data.

D. Open Data Certificates

The Open Data Institute (ODI) has developed the Open Data Certificates [9] which combine the three previously discussed Open Data frameworks into four levels of Open Data access, which are:

Raw – A great start at the basics of publishing open data

Pilot – Data users receive extra support from, and can provide feedback to the publisher

Standard – Regularly published open data with robust support that people can rely on

Expert – An exceptional example of information infrastructure

The Expert level technical requirements can be summarised as follows:

- Provide database dumps at dated URLs,
- provide a list of the available database dumps in a machine readable feed,
- statistical data must be published in a statistical data format,
- geographical data must be published in a geographical data format,
- URLs as identifiers must be used within data,
- a machine-readable provenance trail must be provided that describes how the data was created and processed.

The expert level provides greater technical details than the preceding frameworks. However, in the course of working with Open Data, in our case we have found that the above frameworks do not adequately describe the requirements of software applications, which require technical access to the Open Data and we have developed an Open Data Accessibility Framework.

However, before we describe the Open Data Accessibility framework (ODAF) we will discuss our experiences working with Open Data.

III. Open Data in the City of Gold Coast

The City of Gold Coast, located in the State of Queensland, Australia, is unique in that there is strong support for Open Data at the local, state, and federal levels.

This section describes Open Data adoption at the federal, state, and local levels, and Griffith University's participation.

A. Federal Government

Australia is ranked at number 7 in the world in the Open Data Barometer [3] and is currently ranked number 5 in the world alongside New Zealand in the Open Data Index [4].

The Australian Federal Government launched its open data portal *data.gov.au* in 2012 and appointed the role of Director of Co-ordination and Gov 2.0. The open data portal can be used by any individual or organisation within Australia to host open data including federal, state, and local governments. The portal was migrated to the Open Knowledge Foundation's CKAN [11] platform in 2013 and currently hosts 5,200 data sets from 159 organisations. Since 2012 the federal government has run a national hackathon called *GovHack* where participants from around Australia compete for a pool of prizes. In 2014 GovHack was run in 11 cities with 1300 participants and observers competing for $256,000 in prizes. The federal Minister for Communications gave the keynote speech at the 2014 GovHack awards.

B. State Government

The City of Gold Coast resides in the state of Queensland, the second largest state in Australia, but the third-most populous. In 2013, the Premier of Queensland launched the state's open data initiatives, which included a competition titled the *Premier's Open Data Awards*. Unlike the GovHack hackathon, the Premier's Open Data Awards runs for several months providing participants time to work on larger projects. The Premier presented the awards to participants at both the 2013 and 2014 award ceremonies.

Despite the federal government providing the *data.gov.au* portal for all levels of government to use, the Queensland state government launched its own portal *data.qld.gov.au*, which currently hosts 1577 data sets.

C. Local Government

In 2013, the City of Gold Coast began to spearhead its open data initiative through a collaborative effort between the Economic Development and Information Services Departments. This involved establishing a data portal,

 ISBN: 978-1-61208-445-9

engaging with departments to identify and release data, running community forums to educate the public on open data, and supporting other open data initiatives.

The City of Gold Coast supported Griffith University in running a Premier's Open Data Awards information event in 2013 and also sponsored and helped organise the 2013 and 2014 local GovHack events.

The City of Gold Coast has decided to use the federal government's *data.gov.au* portal to host its data.

The City of Gold Coast has also been active in sponsoring development of apps which utilise Open Data including apps developed by Griffith University.

D. Federal, State, and Local Government Interaction

In 2013 the Director of Co-ordination and Gov 2.0 stated that the City of Gold Coast region was unique within Australia in having strong support from local, state, and federal government levels. In 2013 the Economic Development office of the City of Gold Coast along with state and federal departments arranged for Tim Berners-Lee to speak at Griffith University.

The City of Gold Coast has been very supportive of events and initiatives run by both state and federal governments. The federal government has also been very supportive of the Gold Coast region.

IV. Open Data Case Study

This section describes our experiences working with Open Data for three mobile apps and identifies issues during the process.

A. Cultural Challenges

Our first experience with Open Data in the City of Gold Coast began in 2013 with a smartphone app for disability car parks initiated by Regional Development Australia Gold Coast. Having no knowledge of the City of Gold Coast's Open Data support, the committee developing the app first asked the question, can we access the data? Fortunately, the City of Gold Coast had just started their open data initiative with the ultimate goal of "open by default". Despite the new open data initiative it took Council's enterprise architect three weeks to get the data due to traditional mindsets, policies, and procedures towards data protection.

The disability app is shown in Figure 1(a) and has more recently been expanded to show disability toilets and access ramps. The data required for the app is simply a list of latitude/longitude points for disability car parks on the Gold Coast in addition to polygon outlines of the carparks. The data is not sensitive, nor should it require a license, as the carparks can be seen simply by driving around the city. However, the traditional policies of the local council would've made it difficult to acquire and utilise the data. However, due to the council's Open Data initiative, which aims to not only release data publicly, if possible, but also under a license that allows the data to be freely used, re-used, and re-distributed, we were able to easily utilise the data once made available. Additionally the data was then made available for the general public on the federal Open Data portal *data.gov.au*.

We have since worked on two further Open Data-based apps including GC Heritage shown in Figure 1(b), for displaying heritage sites, and GC Dog Parks shown in Figure 1(c) and (d). Despite cultural and policy changes within the City of Gold Coast, acquiring data can still be a time consuming process as data is prepared in formats not previously required. However, the benefits of releasing this data are that the community now has easy access to disability, heritage, and dog park site information.

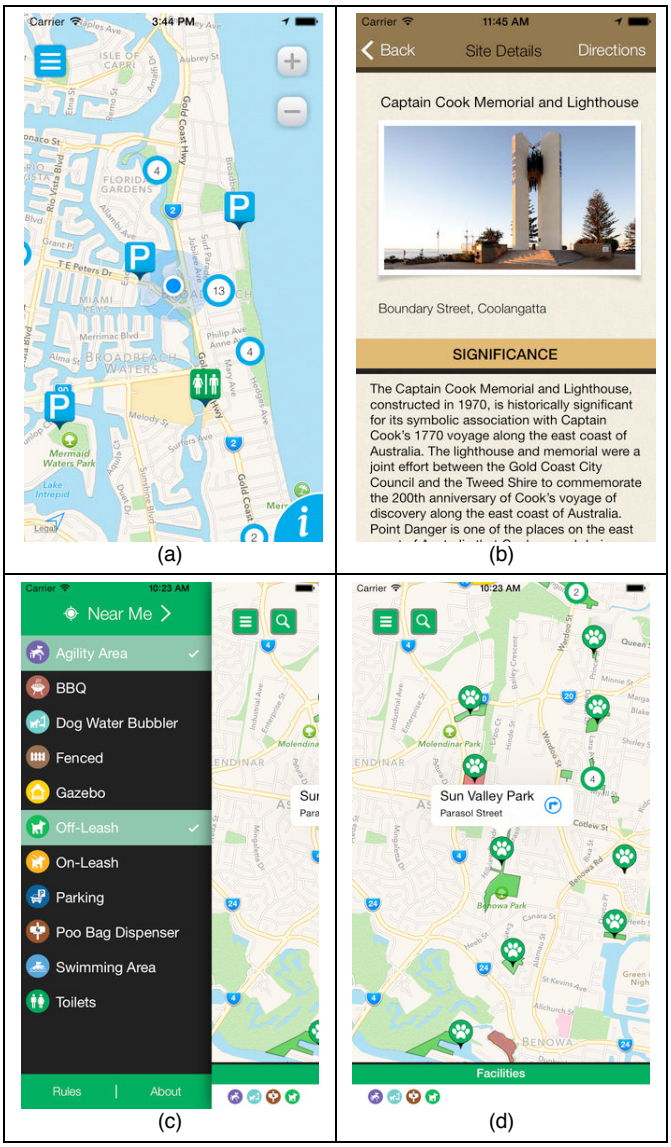

Figure 1. Apps developed by Griffith University using Open Data provided by City of Gold Coast: (a) Access GC, (b) GC Heritage, (c) and (d) GC Dog Parks

B. Data Cleaning Challenges

While working on Open Data for disability car parks there were a number of data cleaning steps required. The data cleaning challenges were as follows:

1. The data came in two files: on-street and off-street car parks, even though the app required no distinction between on-street and off-street car parks.
2. The files weren't named in a way to be able to identify which were on-street and off-street car parks.
3. The files contained *all* car parks in the City of Gold Coast not just disability car parks.
4. The two different files used different notations to identify a disability car park.
5. There were some minor formatting errors in the files.
6. The files were in a large XML file format with a lot of unnecessary data, the files were converted to CSV files suitable for mobile applications reducing the file size by about 100 times.

As it can be seen above, there were 6 data cleaning challenges. We faced similar issues with the additional data sets for other applications, which extended Access GC and for the GC Heritage and GC Dog Parks data. We proposed that the City of Gold Coast adopt our data cleaning process so that future updates to the data would be ready to use. However, the City of Gold Coast was not able at this stage to adopt a data cleaning process for the following reasons:

1. There is a lot of data to be made available and the highest priority is to release the data in the most accessible form,
2. The data custodians responsible for maintaining the data sets do not currently have the responsibility of cleaning it once it is exported,
3. The data custodians don't have the resources to facilitate regular data cleaning for Open Data purposes.

In addition, some of the cleaning processes require programming skills which the data custodians may not have.

Our experiences with using Open Data to date have formed the motivation to develop an Open Data Accessibility Framework. By having a technical accessibility framework, Open Data providers will be able to allocate sufficient resources to ensure Open Data is more accessible and more broadly adopted.

V. OPEN DATA ACCESSIBILITY FRAMEWORK

The 5 levels of the Linked Open Data Framework are aimed to address technical accessibility of Open Data. However, it is possible to achieve level 5 in this framework whilst still presenting many technical challenges to users of the data.

Our aim is not so much to replace the 5 levels but rather expand the 3rd level (use non-proprietary formats) to provide greater detail on technical accessibility.

We have identified six technical aspects that affect Open Data accessibility. These are not so much levels but rather checkboxes. Not all will be attainable by Open Data providers but provides a measure to evaluate the technical accessibility for both Open Data producers and consumers.

The Open Data Accessibility Framework is summarized in Table 1 and described in the following subsections using specific examples from our experiences working with Open Data.

TABLE I. OPEN DATA ACCESSIBILITY FRAMEWORK

Open Data Accessibility Framework (ODAF)
Resource Naming
Data Coalescing
Data Filtering
Data Consistency
Data Formats
API Accessibility

A. Resource Naming

When working with disability carparks we were provided with two files: *carparks.kmz* and *parking.kmz*. One represented on-street parking and the other represented off-street parking. It wasn't clear which file was which. The files or URLs should clearly indicate the contents of the file. In this case a name such as *onstreet_parking.kmz* and *offstreet_parking.kmz* should be used.

Resource names may also benefit from additional information such as the date of release of the data and the region they are from.

There currently is no standard for naming Open Data resources however the Expert level Open Data Certificates do stipulate that URLs should contain dates [9].

B. Data Coalescing

In the disability carpark example the data came in two files: on-street and off-street. There is little need for a distinction between the two types of carparks in most usage scenarios. In addition, the distinction would be more appropriately indicated as an attribute of a carpark record rather than being provided in separate files.

Open Data providers should aim to provide data as single files where there is no need for separate files.

Another example would be providing data separated into files by zip code. Most software applications will find it easier to deal with a single file and have the zip code as an attribute of the data rather than separated into individual files.

The Open Definition 2.0 states that "*the work shall be available as a whole*" [10] and the Sunlight Foundation principles state that "*Datasets released by the government should be as complete as possible, reflecting the entirety of what is recorded about a particular subject*" [6]. However, neither definitions stipulate whether this refers to a single file or multiple files, additionally the focus is on the primacy or

 ISBN: 978-1-61208-445-9

the original raw data, rather than data processed to be more accessible.

C. Data Filtering

In contrast with the previous requirement of Data Coalescing, there are often requirements for data to be filtered. For example the on-street and off-street parking data for the Gold Coast region consists of 22.7MB of uncompressed KML files. In contrast, the extracted disability carparks represented in CSV format were less than 200KB. Mobile apps are an important use of Open Data and a 22.7MB XML file would place a heavy resource burden on a mobile app.

It would be useful for datasets to be filtered for particular domains, in this case disability. Note that this requirement is not in conflict with the previous requirement of Data Coalescing. Data Coalescing should remove unnecessary separations of data whereas Data Filtering should provide useful application-oriented data separation.

A key point that we will address in the next section, is that Data Filtering and Data Coalescing must be driven by the Open Data consumer, as the Open Data producer may not be aware of the needs of the consumer.

D. Data Consistency

Open Data frameworks have identified the need for data cleanness. However, equally important is the need for notations to be consistent between files. As an example our work with carpark data used two different notations to represent disability carparks between the onstreet and offstreet files. One file used an identifier NUM_DISABLED_SPACES followed by a number, whereas the other file used simply the keyword "Disabled Parking".

The Expert level Open Data Certificate stipulates that URLs must be used consistently; however, there is no mention of consistency of other data types [9].

E. Data Formats

The carpark geospatial data provided was in an XML format (KML – Keyhole Markup Language). XML is a nested, structured data format, which is more challenging to process for mobile and web apps than the record-based CSV file format. XML however allows for complex data relationships to be represented and may in fact better represent the original data.

To minimize resource usage, CSV and other similar formats are more suitable for mobile and web applications due to reduced file size and simplicity in processing data.

The disability app data requirements were simply GPS co-ordinates of the disability resource. No additional information was required. Removing unnecessary data resulted in a file size reduction factor of 100.

Note that the recommendation here is not to replace the original data with a filtered CSV-like format, but to provide data in forms that are most suitable for mobile and web applications *in addition* to the original raw data.

The Expert level Open Data Certificate [9] recommends that geographical data be made available in geographical formats such as KML, however, our experience is that these formats are not the most suitable for mobile apps and preprocessing is often required.

F. API Accessibility

Most of the Open Data provided to us has been through files. However there are advantages to providing an API. One advantage is that the entire file does not need to be transferred. One dataset that we had access to was almost 1TB and had to be transferred on a hard drive.

APIs also provide additional benefits such as allowing the data to be filtered and destination formats to be determined at the time of the request.

Open Data portals, such as CKAN (Comprehensive Knowledge Archive Network) [11], utilized by the UK, EU, and Australia, allow for data uploaded in one format to be accessible through an API. However, the API doesn't allow searching and filtering of the data.

Emerging Open Data portals such as Open Data Architectures and Infrastructures (Open-DAI) [12] are beginning to provide support for data filtering. Alternatively technologies such as Elasticsearch [13] can be used to provide comprehensive RESTful API functionality however this would be beyond the skillset of most data custodians.

VI. DISCUSSION

Technical accessibility is an important factor in Open Data adoption. The Open Data Accessibility Framework (ODAF) we proposed identifies six factors that improve Open Data technical accessibility. We will now discuss some of the considerations and consequences of ODAF.

Technical accessibility aims to make it easier for Open Data consumers and software to process Open Data. ODAF identifies characteristics that improve technical accessibility that will require changes to the data and the processes that produce them. We will now discuss these implications.

Firstly, the most important aspect of Open Data is making the raw data available. Even though ODAF promotes changing the data and often removing data, it is important that the raw, original data is still made available. ODAF does not promote reducing the availability of data, but instead providing *additional modes* of the data.

Secondly, ODAF does not prescribe specifically what changes should be made. ODAF does not specify how data should be coalesced, filtered, made consistent, or which formats or APIs to provide it with. Ultimately it is up to the data custodians and consumers to determine these. ODAF is therefore a checklist that describes how successful the Open Data Consumer is in responding to the Open Data Producer.

It would be unrealistic for the Open Data Producer to provide data in every possible combination that could be

 ISBN: 978-1-61208-445-9

required. However, by keeping the ODAF factors in mind, it should result in better quality data sets at the outset.

By adopting an API-based approach the Open Data Producer can satisfy many of the ODAF requirements. An API can often coalesce many data sets into one API resource. Naming is likely improved, and API queries aid with specific queries. APIs generally allow multiple data format responses such as XML, JSON, and CSV and customized fields.

The onus however is still on the Open Data Producer to adopt processes that make the data more accessible. This may be beyond the resources that have been allocated to make Open Data available.

The most important step of Open Data is to make the original raw data available. However to allow Open Data to be useful and widely adopted it must also satisfy the ODAF requirements. This may require adopting an API-based Open Data Portal. However, existing Open Data Portals are limited in their ability to clean, filter, and coalesce structured data. Open Data Portals must be extended to provide querying abilities within structured data to satisfy the requirements of ODAF.

VII. CONCLUSION

In this paper we have explored existing Open Data frameworks and highlighted their weaknesses in describing requirements for technical accessibility. Based on our own experiences working on three Open Data projects and also being involved with Open Data initiatives at the local, state, and federal government levels in this work we propose the Open Data Accessibility Framework which presents factors which improve the technical accessibility of Open Data.

Adopting the ODAF factors will require a commitment from Open Data Producers to listen to their consumer's needs and make appropriate changes. It will require more resources to make the Open Data more technically accessible. Ultimately it should result in the data being available through an API. APIs can open up other opportunities such as crowdsourcing data, transitioning from e-Government to "we-Government" [14][15], progressing to what O'Reilly defines as "Government as a Platform" [16].

Open Data is an emerging initiative. Great progress has already been made in adoption at all levels of government throughout the world. Much of the progress has been at the policy and cultural level. There has been a focus on releasing data in a timely manner including the proposal of the timeliness measure *tau* [17]. However, much more work needs to be done at the technical level and this ODAF is a framework that defines attributes of technically accessible Open Data.

REFERENCES

[1] M. Heimstadt, F. Saunderson, and T. Heath, "From Toddler to Teen: Growth of an Open Data Ecosystem," JeDEM, vol. 6, no. 2, 2014, pp. 123-135.

[2] J. Kloiber, "Open Government Data – Between Political Transparency and Economic Development", Masters Thesis, Utrecht University, 2012.

[3] T. Davies, "Open Data Barometer 2013 Global Report", http://www.opendataresearch.org/content/2014/666/open-data-barometer-2013-global-report, 2013 [retrieved: March, 2015].

[4] Open Knowledge Foundation (OKFN) Global Open Data Index, http://index.okfn.org/place/, December, 2014 [retrieved: March, 2015].

[5] Open Knowledge Foundation (OKFN), Open Definition 1.0, http://opendefinition.org/history, August, 2005 [retrieved: March, 2015].

[6] Sunlight Foundation, "Ten principles for opening up government", http://sunlightfoundation.com/policy/documents/ten-open-data-principles/, August, 2010 [retrieved: March, 2015].

[7] T. Berners-Lee, "Is your Linked Open Data 5 star?", http://www.w3.org/DesignIssues/LinkedData.html, 2010 [retrieved March, 2015].

[8] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data – The story so far", Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 2009, pp. 1-22.

[9] Open Data Certificates, https://certificates.theodi.org, 2013 [retrieved: March, 2015].

[10] Open Knowledge Foundation (OKFN), Open Definition 2.0, http://opendefinition.org/od/, December, 2014 [retrieved: March, 2015].

[11] Comprehensive Knowledge Archive Network (CKAN), http://ckan.org [retrieved: March, 2105].

[12] R. Iemma, F. Morando, and M. Osella, "Breaking Public Administrations' Data Silos", JeDEM, vol. 6, no. 2, 2014, pp. 112-122.

[13] O. Kononenko, O. Baysal, R. Holmes, M. Godfrey, and D. Cheriton, "Mining Modern Repositories with Elasticsearch", MSR 2014, Hyderabad, India, May, 2014, pp. 328-331.

[14] D. Linders, "From e-government to we-government: Defining a typology for citizen coproduction in the age of social media", Government Information Quarterly, vol. 29, no. 4, 2012, pp. 446-454.

[15] T. Nam, "Suggesting frameworks of citizen-sourcing via Government 2.0", Government Information Quarterly, vol. 29, no. 1, 2012, pp. 12-20.

[16] T. O'Reilly, "Government as a Platform", Innovations, vol. 6, no. 2, 2011, pp. 13-40.

[17] U. Atz, "The Tau of Data: A new metric to assess the timeliness of data in catalogues", Proceedings of the International Conference for E-Democracy and Open Government, Krems, Austria, May, 2014, pp. 258-268.

 ISBN: 978-1-61208-445-9