Publishing information as data.

Guide to best practices and minimum requirements for structured data publication

***also* for small, underfunded and non-technical organizations.**

Contents

Publishing information as data. 1
Guide to best practices and minimum requirements for structured data publication 1
Introduction to this document....1
Thanks 2
Team....2
Glossary 2
Process flow....3
Modeling your information :....4
Minimal requirements and recommendations : 5
FORM....5
CONTENT....5
Examples 6

Introduction to this document

This document intends to explain in abstract and by example how to publish information and data acquired, even by small, probably underfunded and non-technical organizations in a re-usable, meaning machine readable fashion. The reason this work was done was to address an apparent reticence in many organizations to publish information in a structured format . **Note :**

This document was written over the course of two days and is by necessity limited in scope. Most notably practical recommendations related to anonymization of information before publication are not given. This is a glaring omission purposely introduced specifically because of the importance of the subject, it therefore preferable to blatantly not address this rather than to have a stunted description suggest completeness.

Should you, as reader of this document, have questions. Or maybe you would like some help with setting up a process to publish your organization´s information please do not hesitate to contact us at info@chokepointproject.net . If you do not need our help, but are publishing information in a structured re-usable fashion, please let us know. If you want to receive updates to this document, same thing. Send us a mail and let us know.

Thanks

This guide and the accompanying examples were developed during the "Developer Summit" attached to the Stockholm Internet Forum 2013 organized by the Swedish Ministry of Foreign Affairs, hosted by .SE and funded by Sida, we would like to thank all of them for giving us the opportunity to have developed this work.

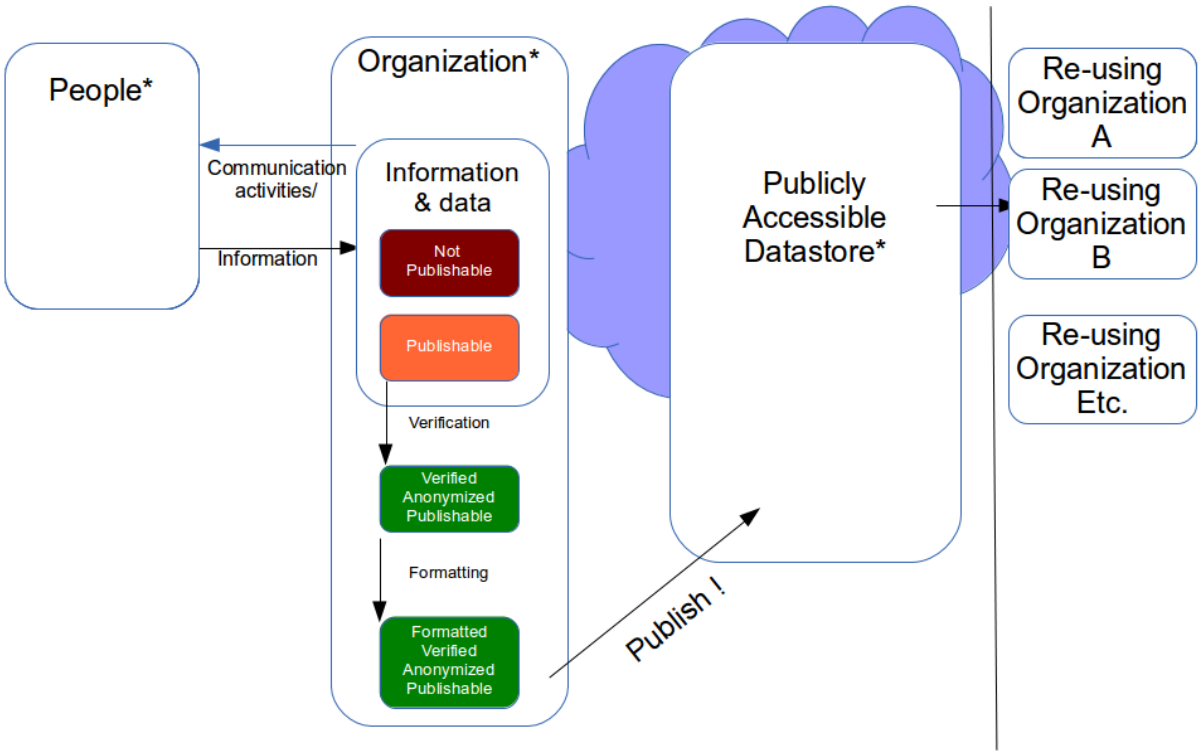
Team

Ruben Bloemgarten	ruben@abubble.nl	www.chokepointproject.net
Mohammed Ali	mohamed.ali@arabdigitalexpression.org	www.arabdigitalexpression.org
Radhouane Fazai	radhouanef@gmail.com	www.democracyinternational.com
Raphaël Vinot	raphael.vinot@gmail.com	www.circl.lu
Mireille Raad	mir.mir@gmail.com	Freelancer
Gustaf Bjorksten	gustaf@accessnow.org	accesssnow.org

Glossary

Machine readable:	http://en.wikipedia.org/wiki/Machine-readable_data
Human readable :	http://en.wikipedia.org/wiki/Human-readable_medium
Information :	
Data :	
Encoding:	http://en.wikipedia.org/wiki/Character_encoding
UTF-8:	http://en.wikipedia.org/wiki/UTF-8
CSV :	Comma Separated Values http://en.wikipedia.org/wiki/Comma-separated_values
Data modeling :	The process of structuring information with fixed descriptions. http://en.wikipedia.org/wiki/Data_modeling

Process flow



* **People** : These would be the people that your organization has interaction with, they and the people in your own organization tell you things about the activity that is pertinent to you. NOTE: the people you communicate with should not be the information you record, pertinent is the information (each piece being a “data point”.)

* **Organization** : You

* **Datastore** : A place where data is stored, this could be a website hosted by you, an online storage facility such as Google docs and similar.

* **Not Publishable** : This is information that you do hold inside your organization but is not suitable to be exposed to the public. You should know best what these are. If you feel uncomfortable making this determination look for outside help to aid in this determination.

* **Publishable** : This is information that you **know** to be suitable for publication, i.e. those pieces of

information that you are already publishing. Information that is already in the public sphere.

* **Verification** : Despite you already having made the determination of what is publishable or not, before publishing your report to the world, take a last look to see if it is really does not contain information that should not be published.

* **Formatting** : This is where your choice of formatting is executed. In the case of using an spread sheet form, such as in the attached example (Google Docs spreadsheet) this simply means exporting the generated csv file.

Modeling your information

This sounds more complicated than it is, this is simply the moment that you are determining the fields in your spread sheet (or whichever format you have chosen). These fields will each contain a “field descriptor”, which simply means a human understandable word describing the nature of the information that will be contained in the fields. A few basic elements that should always be present are : a time stamp and a location related to the row of information made up by the fields you have described in your model.

A good way to go through this process is to have two or more people sit down together where at least one person is intimately familiar with the goals and work done in your organization and at least one person who is familiar with data modeling (this does not necessarily need to be an expertise greater than being familiar with creating spreadsheets).

Minimal requirements *and recommendations*

FORM

- **Publish in a structured plain text format (csv, xml, json, sql)**
  The first line of generated document should state the encoding used in the file (if possible use UTF-8)
- **Publish new documents at regular intervals. Keep them alive by publishing continuously.**
  Keep your publications alive, try to publish daily, weekly or monthly updates.
  By continuously updating you are strengthening the information you have previously published.
- **Publish to a constant location.**
  This means that the url where a file can be retrieved by others is constant, such as http://www.myorg.com/export/YYYYMMDD.orgname.report.lang.csv .
  If you publish very often, once a day or faster, it is strongly recommended to also introduce a folder structure to your publication : add folders organized by year (YYYY), month(MM) and day (DD).
- **Use a naming convention when publishing.**
  YYYYMMDD.ORG.REPORT.LANG.EXT
  → 20130521.SocialScanForMokkatam.ar
  Where :

YYYY	***=***	***the year of publication***
MM	***=***	***the month of publication***
DD	***=***	***the day of publication***
		- Using this format, starting with the date, starting with the year, makes it much easier to sort the reports by date, in turn facilitating automated reprocessing of the information you are trying to get into the world. Its also easier for people using eye-to-brain processing power to quickly find and understand what they are looking at.
ORG	***=***	***the organization publishing the report***
VALUE	***=***	***additional descriptor field (you might be publishing many different reports!)***
LANG	***=***	***language of the report. Even if you only publish in one language, by simply adding this descriptor it might entice others to translate the report content.***

CONTENT

- **Never rename a field, create a new one.**
  After you have created the structure (the data model) for your report file, and have started to publish your data, you might feel that a descriptor is not a clear as you initially thought. However, renaming the field descriptor name, could very well break the processing done by others, and will make analyses over time much more difficult to perform. Simply create a new field and stop using the old field.

- **Never remove a field, leave it blank**
  After you have created the structure (the data model) for your report file, and have started to publish your data, you might feel that a particular field is no longer useful. Just stop using it, removing the field will change the "row order", breaking any automated processing and as before, will also make analyses over time very hard.
- **Do not use abbreviations !**
  Abbr. R diff to underst.
  Using fewer letters may be optimal use of characters, it makes it extremely difficult for others to understand what you might mean.
- **Be descriptive.**
  Imagine that you have a very bad memory and never document your work but still want to understand the report in 50 years time.
- **NEVER EDIT YOUR REPORTS AFTER PUBLICATION.**
  This is equal to lying. Also it means information already processed by others is now no longer correct. When wanting to change a piece of information, just publish a new report.

Examples

During these two days the following examples were created based on the principles outlined in this document :

- **Input form :**

https://docs.google.com/forms/d/1mboXDFFcwAn9zHilQ2MbVd-KELslG2Eo4BNgf-L1a4c/viewform

- **output file example** **:** 20130521.SocialScanForMokkatam.ar.csv
- **odg format process flow** **:** 20130521.DataPublishingProcessFlow.en.odg
- **png format process flow** **:** 20130521.DataPublishingProcessFlow.en.png
- **odt version of this file** **:** 20130523.A_Guide_To_Publishing_Information_As_Data.en.odt
- **pdf version of this file** **:**