Best Practice: Identify what you already publish

25 July 2016

This version
  http://www.w3.org/2013/share-psi/bp/iwyap-20160725/
Latest version
  http://www.w3.org/2013/share-psi/bp/iwyap/
Previous Version
  http://www.w3.org/2013/share-psi/bp/iwyap-20160627/

This is one of a set of Best Practices for implementing the (Revised) PSI Directive developed by the Share-PSI 2.0 Thematic Network.

Outline

Organisations might find deciding what information resources should be made available for re-use in machine-readable formats challenging. Information already published by organisations represent a good candidate for datasets to be published as open data. Therefore organisations should create and maintain inventory of already published information. However the amount of such information is often too large to be catalogued manually. Therefore automated scraping techniques should be applied to create inventories of already published information.

Links to the Revised PSI Directive

Selection

Challenge

Where to start when deciding what information resources should be made available for re-use in machine-readable formats?

Identifying what information should be made available in machine-readable formats for re-use might be challenging due to the lack of knowledge what information is already published and the amount of information might be too large to be catalogued manually.

Solution

An inventory or catalogue of already published data and information assets should be developed and maintained. This may be achieved manually or by using automated scraping techniques to gather details of information assets that are already published on the Web site.

Why is this a Best Practice?

Information is frequently published under a distributed process using a content management system. Inventory of already published information assets might be missing and it might be difficult to create it manually. Therefore organisations might find it challenging to know where to start when deciding what information resources should be made available for re-use in machine-readable formats.

An inventory of already published information helps organisations to understand what information they provide and what assets they can make more re-usable. Understanding of what datasets an organisation can possibly publish as open data is essential for selecting datasets for publication. Techniques such as site scraping allow organisations to periodically audit their Web site in order to assess what information assets they publish and in what form (open, closed, etc.).

Why is there a need for this Best Practice?

Information is frequently published under a distributed process using a content management system. creating an inventory of already published information assets might be missing and it might be difficult to create it manually. Therefore organisations might find it challenging to know where to start when deciding what information resources should be made available for re-use in machine-readable formats.

How do I implement this Best Practice?

A simple spreadsheet might serve as an inventory of the data/information assets, but depending on the volume of information and the requirements of the organisation, cataloguing solutions such as CKAN might be deployed.

Scraping software/libraries are needed, such as Scrapy. Metadata gathered using the scraping software can be used as facets for sorting and grouping the links. Faceted browsing features are provided by application such as Exhibit. If security is the concern, the scraper should be run on an isolated machine and only the headers should be processed.

Whether created manually or by automated means, the inventory should contain at least basic metadata about the data/information assets like the title, location, current format and terms of use. Additional metadata like the responsible person/unit, target data format or update frequency might help to manage the future publication process and it helps to make more precise estimates of the effort and costs needed to publish and maintain the open datasets.

Where has this best practice been implemented?

Country	Implementation	Contact Point
Scotland	The Scottish Government	Dr Peter Winstanley, The Scottish Government.
Helsinki Region Inforshare	Open Data Pipeline	Comment box included in the page

References

- Timisoara Workshop Session: Identifying what you already publish
- Krems Workshop Session: Extracting Structured Data from Unstructured Open Data

Local Guidance

This Best Practice is cited by, or is consistent with, the advice given within the following guides:

- (Austria) Open-Government-Vorgehensmodell Open Government Process Model
- (Belgium) Open Data Handleiding Open Data Handbook
- (Croatia) Preporuke o prilagodbi skupova podataka za javnu objavu i ponovno korištenje Open Data Guide, Croatia
- (CzechRepublic) Standardy publikace a katalogizace otevřených dat veřejné správy ČR Open Data Standards
- (Finland) Helsinki Region Infoshare
- (Finland) Avoimen Datan Opas Open Data Guide
- (Germany) Open Government Data Deutschland
- (International) Open Data Handbook, Solutions Bank
- (Lithuania) Viešojo Sektoriaus Informacijos platinimo gerosios praktikos Best Practices for Sharing Public Sector Information
- (Luxembourg) Recommandations pour l'ouverture des données publiques Recommendations for opening data
- (Malta) PSI Directive Implementation & Internal Data Sharing Platform (draft)
- (Netherlands) Handreiking bij openen van data Guidance on Open Data
- (Serbia) Open Data Handbook
- (Spain) Guía metodológica para planes open data sectoriales Methodological Guide for Sectorial Open Data Plans
- (Spain) Guía de aplicación de la Norma Técnica de Interoperabilidad de reutilización de recursos de informaciónApplication Guide for Technical Interoperability Standard on PSI re-use
- (Sweden) Vidareutnyttjande av information Om PSI och öppna data Reuse of PSI and open data
- (UK) Open Data Resource Pack
- (UK) Birmingham and West Midlands Localised Guide for Open Data

Contact Info

Dr Peter Winstanley, The Scottish Government.

Related Best Practices

- Publish overview of managed data
- Categorise openness of data
- Dataset Criteria
- Develop an Open Data Publication Plan

Issue Tracker

Any matters arising from this BP, including implementation experience, lessons learnt, places where it has been implemented or guides that cite this BP can be recorded and discussed on the project's GitHub repository