Open Data Life Cycle and Infrastructure Bar Camp

[2014-12-04, 14:00-15:10, Lisbon, LNEC]

Facilitators: Jan, Harris, Peter W.

Scribe: Jan

Minutes

Attendees: 3 + 2 facilitators + 1 facilitator/scribe

Introduction

- Aims of this bar cam are to discuss the Open Data lifecycles and the infrastructure that support the publication of Open Data. When discussing the Open Data lifecycles attention should be paid to the involvement in the curation of datasets and on the user engagement throughout the life cycle.
- Existing life-cycle models
  - Linked Open Data Life Cycle
  - Data Curation model
  - Engage Life Cycle models
  - COMSODE model
- Domains in Open Data publication
  - Planning
  - Preparation
  - Publication and cataloguing
  - User relationship management
  - Archiving
- Three bar camp questions outlined
  - Should the users be involved in the data management (curation) of open data?
  - Should the user engagement (collaboration) be performed throughout the whole life cycle?
  - What is the optimal IT infrastructure for public sector to support the IT-driven economy?

Q: What do you mean by the "model"?

- It is the structure of the dataset, the fields and the units of measure etc.

Q: Does the feedback loop exists in reality?

- Note: feedback loop in which the feedback provided by the users is utilized to improve the quality of the published datasets.
- For example in the Czech Republic feedback led to improvement of datasets published by the Czech Telecommunication Office.

Q: Can we apply the presented models to the real time data

- Presented models are applicable to the real time data se well however an appropriate infrastructure to support the steps of the life cycle is needed.

Based on the bar camp questions three factors facilitation PSI/Open Data publication and reuse were identified:

- User involvement in data curation
- User engagement throughout the life cycle
- Proper IT infrastructure supporting the publication of PSI/Open Data

Description of the above mentioned factors as well as the answers to the SharePSI 2.0 Lisbon Workshop questions (“What X is the thing that should be done to publish or reuse PSI?”, “Why does X facilitate the publication or reuse of PSI?”, “How can one achieve X and how can you measure or test it?”) are provided in Conclusions section.

Three questions for the bar camp – highlights of the discussion

1) Should the users be involved in the data management (curation) of open data?
    - Why (not)?
        - Opinion: No. Users can combine datasets but it is not necessary to involve them in the curation process.
            - Data is already in a system, just publish raw data directly from this system
            - When you build a system it just needs an API, so not involving users simplifies the development/procurement
            - The most important thing is making data available, taking into account the different needs of people
                - Sometimes data is 'tagged' by one group with one term, but with another group with another term. For enhanced discoverability it is helpful if the users can enrich data with elements that are user provided, but this should be done in a way that doesn't change the original data nor disguise the provenance.
        - Opinion: Yes. For example In Scotland users are involved in curation of datasets in the ALISS service. Data from various sources of data related to the health care domain are integrated in this service. Users can provide perspectives on the data assets.
        - Q: Can users change the published data in ALISS service?
            - A: Users cannot change the original data but they can contribute to the metadata (metadata enrichment).
        - For discovery it is good if the data are enhanced by the users – data are tagged but the original data should not be polluted.
        - In general, involvement of users in the data curation would in fact depend on what is meant by the curation.
    - How (if yes)?
        - Metadata enrichment like in the ALISS service example.
        - Format change
            - E.g. transformation of data from PDF to excel
            - However raw and machine readable data should be published from the start rather than letting users to do the transformation.
2) Should the user engagement (collaboration) be performed throughout the whole life cycle?
    - Why (not)?
        - There should be collaboration between the providers of data and the users throughout the whole life cycle.

    - Using the Engage life cycle model as an example in case of the ALISS service users are involved in most of the life cycle phases, at least up to some extend (pre-processing phase is not present in ALISS service life cycle).
    - Users might be involved for example in curating, duplication checks, pre-processing, creating/gathering the data.
  - How (if yes)? How the users are involved?
    - Examples how users are involved in the ALISS service:
      - Create phase – people work with organization to collaboratively spot the assets (parks, walks, health facilities etc.).
      - Curate phase – adding meaning to the data, adding metadata (description, tags).
      - Publishing phase – users can search the assets using a set of provided tools and API.
    - Involvement of the users might be also dependent on the data. It might difficult to achieve it with for example sensors data.
      - However people can for example reuse the earth quake sensors data.
    - User can by engaged by being involved in the data area and having the opportunity to collaborate.

3) What is the optimal IT infrastructure for public sector to support the IT-driven economy?
   - Why optimize the infrastructure
     - Users expect for example quick response from the government web sites. It can be achieved by the right architecture of the page and the location of the component parts. So we should ask what the kind of consideration of design in order to support businesses requirements is.
     - We develop on technology designed for 20th century – expensive disks, expensive memory. Is this the architecture to support for example massive mobile usage of the web site? Can't we do better?
     - Certain "path designs" are expensive from the computation perspective in the infrastructure designed for 20th century.
     - We should try to specify the infrastructure for the specific delivery.
       - Aim for memory-only computing
       - Why do update and delate in the data stores and not just use persistent data stores
     - Most relational data will fit into 1TB database – we can get improvements with data stores optimized for handling this amount of data
     - What are the candidate checks what the best candidates are
   - Why not optimize the infrastructure (counter arguments)
     - Government should not be involved in these so low level problems.
       - However for example the Scottish government is different – it has a large IT department that is able to deal with this kind of problems.
     - Governments should look for the right people to the job and contract them
       - However the procurement lacks behind the technology. Procurement often aims for not what the best is but what is safe.
     - Cloud based service might be a solution (do not buy your hardware, use some service).
       - (+) Provides flexible cost structure.
       - (-) It might be insecure.

        - Infrastructure is a crucial thing to get the things working. If we provide documents as a service (web pages) we should be expected to be provide data as a service.
        - How
            - Two possible approaches:
                - Using SW and HW as a service – do not procure your own infrastructure.
                - Building strong IT department.
            - Abandon infrastructure with roots in the 20th century and use the infrastructure developed for the 21st century which can better satisfy needs of publication data on the web.

Conclusions

User involvement in data curation

What X is the thing that should be done to publish or reuse PSI?

- In order to facilitate publishing and reuse of PSI users should be involved in data curation.

Why does X facilitate the publication or reuse of PSI?

- Metadata enriched by the users can improve discoverability of the datasets and it can also help to add meaning to the data.

How can one achieve X and how can you measure or test it?

- Metadata enrichment – users might be allowed to enrich the metadat, e.g. they can tag the datasets or add/improve datasets description.
- Users might perform transformation between various formats of data and provide back the transformed datasets.
- However original datasets should stay untouched. I.e. it should possible to distinguish between the original data/metadata and the user generated content.

User engagement throughout the life cycle

What X is the thing that should be done to publish or reuse PSI?

- User engagement facilitates publication and reuse of PSI. User engagement should not just one phase of the Open Data life cycle but users should be engaged throughout the whole life cycle.

Why does X facilitate the publication or reuse of PSI?

- If the users are engaged from the start they can help to identify datasets that are in demand and thus it facilitates reuse of the published data. It also helps publishers to focus on the right datasets.

How can one achieve X and how can you measure or test it?

- Depending on the phase of the Open Data life cycle or the type of data users might be involved for example in collaborative selection of datasets, metadata enrichment, search or use of the provided tools and APIs.
- In general provides of the data should be able to provide the users opportunities to collaborate.

Proper IT infrastructure supporting the publication of PSI/Open Data

What X is the thing that should be done to publish or reuse PSI?

- In order to facilitate the PSI publication and reuse proper IT infrastructure supporting the publication should be in place.

Why does X facilitate the publication or reuse of PSI?

- In order to be able to publish data in a way that meets users' expectations an appropriate IT infrastructure is necessary. Users expect for example quick response from the government web sites. If the publishers are not able to satisfy expectations of the users it might hinder the reuse.IT infrastructure is a crucial thing to ensure that the level of the provided data services meets the users' requirements.

How can one achieve X and how can you measure or test it?

- Procure SW and HW as a service or build a strong IT department that is capable of providing and supporting the required infrastructure.
- Abandon infrastructure that is made up technologies with roots in the 20th century and use the infrastructure developed for the 21st century which can better satisfy needs of publication data on the web.
- Use architecture and the up-to-date technologies that are specifically designed/optimized for the intended delivery
    - E.g. in memory computing, persistent data stores, data stores optimized for the expected amount of data etc.