European Commission

Data Mosaics: standards and prototype for a data mashup platform

Final Report

Author: Philippe Duchesne

Joint Research Centre

1 Summary of work carried out

The goal of this project was to build a stack of tools and possibly standards to annotate, and more generically cross-link fragments of data available on the web, and group these graphs of fragments into contextualized bundles called 'Data Mosaics'. The envisioned architecture consists of two main technical components and two formalized models that can become standards, namely:

- A web service to store and query Data Mosaics,
- A web client to author and browse these mosaics,
- A formalization of URI fragments for possible data types, and
- An abstract model for Data Mosaics.

This objective was achieved and illustrated by a specific use case: annotating documents and resources related to the INSPIRE Directive. This use case is representative in that it involves resources from multiple sources (official EC documents, technical guidelines, legal texts, Member State reports, etc.) and multiple types of data (mostly PDF and HTML, informational videos, geospatial services).

The project resulted in an operational demonstrator deployed online, as documented in deliverable D2, and showcasing (but not limited to) several scenarios related to the INSPIRE use case, namely:

- Linking the INSPIRE Directive to its transposition into member states legal texts.
- Annotating member states reports with relevant fragments of the directives and with actual online implementations.
- Annotating technical guidelines with training material.

2 Lessons learned

This demonstrator shows how Data Mosaics, along with URI fragments, constitute aggregated views can be used to convey domain specific knowledge and perspective, thus enabling users to easily stitch together and publish added value data mash-ups. Because such Data Mosaics exist and can evolve independently of the linked resources, they are documents on their own, like multi-faceted patchworks of remote data that can be collaboratively authored, annotated and exchanged.

Furthermore, usage of URI fragments brings semantics to any subparts of an online resource. While a resource identified solely by a URI is a semantic monolith that can only be tagged as a whole, the use of fragments allows the tagging of parts of video streams or of paragraphs of a text. Breaking that atomicity extends the realm of Linked Data beneath the surface of these resources. It makes such as solution not only an annotation tool, but a repository of linked data fragments, holding much finer-grained semantic information.

2.1 User interface

As for the user interface, an important aspect of the user experience is to have a seamless integration of these tools with the environment of the user, i.e. allow the user to consume annotations while keeping his browsing habits. The integration of the developed tool with common Internet browsers is a major feature to engage users in the data curation process. A seamless overlay of the mosaic functionalities on top of the natural browser interface avoids breaking the train of thought of the user and helps capturing one's ideas.

2.2 Modular architecture

The abstract model and the representation format for the Data Mosaics play a central role in the solution, by guaranteeing a decoupling of the web service and the web client from an implementation point of view. That way, it is possible (and desired in some situations) to have ad-hoc implementations of either the service or the client, while relying on the Mosaic model and representation as a pivot.

In that respect, the use of the W3C Recommendation (Web Annotation Model) as the basis for the model is crucial in that it makes the solution easily interoperable with potential future solutions that may emerge from this specification: overlaying annotations from such services would be immediate.

2.3 Access to authoritative sources

For a use case as the one developed here (annotating EU directives, their legal transpositions and member states reports), it is important to be able to annotate original documents hosted by authoritative portals. This assumes one of the golden rules of Open Linked Data to be true: resources should be exposed using permanent, resolvable URIs. While this is now obvious to many people, it still is not the case for several official portals. As an example, the French registry of legal documents (www.legifrance.gouv.fr) publishes PDF documents within its portal only, in a way that cannot be directly accessed using a simple HTTP GET operation. Such a direct access by URL should not be overlooked.

3 Recommendations

This project leads to several recommendations and directions for further exploration.

3.1 Supported resource types

While the use case of the prototype (annotating INSPIRE-related documents) is a very relevant one and demonstrates the core concepts of Data Mosaics, it focuses only on a small subset of the types of resources that are relevant to such an approach. Other media types such as videos, audio streams and tabular data are potential subject to annotations.

The application currently includes viewers for HTML, PDF, videos (including YouTube and Vimeo streams), geospatial services and data (GeoJSON, OGC standards, ESRI services).
Other datatypes may be supported. In particular, tabular data formats such as CSV should be implemented.

3.2 Controlled vocabularies

Deliverable D2 shows how controlled vocabularies can be used to tag fragments or annotations. It is good practice to use as much as possible such vocabularies to do so (and reuse existing vocabularies when available), to ensure semantic interoperability with other applications handling data in the same domain. It has been stated earlier how the solution described in this project can be seen as a repository of semantically linked fragments; relying on shared controlled vocabularies ensures that this linked data reaches out to external data and yields mutual enrichment.

3.3 User experience

Regarding user experience, the current prototype does not entirely achieve seamless integration, as it still is a separate web application. Complete seamlessness will be achieved if annotations can be fetched and overlaid while viewing a resource using the native functionalities of the browser. Having a web browser that natively supports such annotations would be the ultimate integration. Short of that, a native plugin would offer a much better user experience.

3.4 Graph View

Enhancing the graph view to provide true browsing through the graph would be a major improvement, especially for large sets of resources interconnected with each others. Especially, in a mass-market context where large number of resources would be annotated by a large audience, it is interesting to be able to browse the graph of linked fragments visually, expanding nodes and edges at will.

3.5 Social network integration

The current web application integrates social networks such as Twitter, LinkedIn, Facebook and Google as authentication authorities. As a further integration of these networks, the Mosaic Web Client can use these networks as catalogs of resources to pick from when building a mosaic: re-tweets, liked pages in Facebook, documents in Google Drive, etc.

3.6 Offline download

While the key feature of Data Mosaics is to be able to reference and display online datasets, it is sometimes necessary to be able to work offline. Implementing offline mosaics requires a way to extract the relevant fragments for every resource type, and pack them in a self-standing archive.

3.7 Other application domains

In terms of application domains, many other use cases involving heterogeneous online data are likely to benefit from such a platform, such as:

- information aggregation in disaster management,
- fact checking in data journalism,
- embedding experimentation data in open science, or
- referencing media and legal source in political debates.