Data Mosaics: standards and prototype for a data mashup platform

D1 – Technical background, usage scenarios and storyline of a demo

Author: Philippe Duchesne

Joint
Research
Centre

1 Introduction

Now that large amounts of online - and often open - data become available, along with efficient visualization tools for their respective media types, one of the next challenges is to make sense of these data in the scope of particular domains and use cases. A fair part of the sense making lies in the ability to connect parts of data elements and tie them together so that their combination carries more information than the sum of the individual elements.

When creating a unified and structured view of heterogeneous data sets that fits their particular needs, users face a series of challenges, such as:

- they do not own the various data sources they are referring to, nor can they edit them;
- they need to collect and structure only parts of resources that are identified by URIs;
- they need to store the result of their curation, exchange it and possibly collaboratively edit and annotate it; and
- they need to view that result in a transparent interface that overlays resources with their respective annotations and related fragments.

Technologies that answer these problems exist or are emerging, but a comprehensive end-to-end solution to produce data mashups (mosaics) is yet to be produced. Use cases, on the other hand, are numerous and cover a wide spectrum of domains, ranging from open science to data journalism and fact checking.

This project aims at exploring the possible features and technical components of such a solution, and producing a prototype tailored for a specific scenario. It will build upon the preliminary work done in the Data Mosaics project [1].

2 Definitions

These terms need to be clarified for the proper understanding of this project:

- Resource: An entity that can be identified and retrieved by a URL with no fragment identifier, and described by a mime type.
- Fragment: a fragment of a resource, identified by a fragment identifier.
- Annotation: an entity that links a (fragment of a) resource to another (fragment of a) resource, using a directed link, and potentially described by extra metadata (link type, author, date, etc.).
- (Data) Mosaic: a document that aggregates resources and fragments referenced by URL, linked by annotations to form a directed graph of elements that pictures a certain context.

3 Existing technologies & standards

3.1 Fragments

The notion that underpins the whole process of "mashing-up" data is the ability to extract and convey fragments of data entities. In a web environment where data entities are identified by URLs, fragments of these entities should naturally be represented using the notion of URL fragments, as defined in the URI specification [2]. This specification however leaves the fragment inner syntax open and lets each media type define its own ad hoc syntax, leading to potentially vendor-specific, heterogeneous solutions, e.g. the temporal fragment syntax of YouTube, or the page fragments for PDF files.

A harmonization effort, focusing on audio and video media types, is done in the MediaFragments proposed recommendation [3], while the Web Annotation Data Model [4] lists the media type specifications that define specific fragment syntax.

The Data Mosaics project gathers these proposals and specifications, and attempts a formalization of possible media dimensions (temporal, spatial, text, ...), and of their respective fragment syntaxes [1], as shown in the table below.

Fragment syntax	Representative media types	Description
`#t=`*`timerange`*	`video/*, audio/*`	used for **temporal** fragments, as defined in MediaFragments
`#xywh=...`	`image/*, video/*`	used for fragments in **pixel space**, as defined in MediaFragments
`#xpath=...`	`text/html, application/xml, application/json`	used for any **tree-structured** media type
`#bbox=`*`x1,y1,x2,y2`*	`application/vnd.google-earth.kml+xml, application/gml+xml`	used for fragments in the **geospatial** space
`#col=`*`range`*`&row=`*`range`*	`text/csv`	used for **tabular** data fragments
`#line=`*`range`*	`text/plain`	as defined in RFC5147
`#page=...`	`application/pdf`	used for **paged media**, such as pdf or slides
`#id=`*`elementId`*	`application/rdf+xml`	identifies an element by ID in the namespace of the RDF resource

This list of fragment syntaxes will be the basis for the expression of fragments in the following sections.

3.2 Web Annotation Data Model

The Web Annotation Data Model [4] is a W3C Proposed Recommendation for a specification that describes a structured model and format to represent and share annotations. The specification covers a wide range of use cases and addresses several key points in the perspective of this document:

- The need to formalize fragment references: the Web Annotation Data Model develops the wider notion of *selectors*, of which fragments are a subclass. The specification also lists existing fragment specifications (in a similar effort to what is done in [1]).
- A complete vocabulary to describe the annotation model and store annotations as triples.
- The notion of 'annotation collection' that is similar to Data Mosaics.

The Web Annotation Data Model, and its vocabulary, should therefore be considered to express the Data Mosaic model; if possible a Data Mosaic should be an instance of an Annotation Collection as defined in the Web Annotation Model.

3.3 AnnotatorJS

The AnnotatorJS project (http://annotatorjs.org) offers a mature solution to annotate documents, with an extensible Javascript library at its core, and a wide range of plugins to support various media types. It differs however from this initiative in several aspects:
AnnotatorJS focuses primarily on annotations per se, whereas this document addresses the notion of mosaics.

There is no explicit support for semantic aspects (tagging annotations with existing vocabularies) in AnnotatorJS.

As a consequence of the points above, the storage solution of AnnotatorJS does not offer the query capabilities of a true linked data solution for storing and exploring data mosaics

So AnnotatorJS is a mature implementation of the client side (visualisation and authoring) component, and may be considered for integration in the technological stack described in this document.

4 Business case and scenario

This project will prototype a data mosaic platform applied to the use case of annotating and cross-referencing resources related to a European Union directive, more specifically the INSPIRE Directive (2007/2/EC).

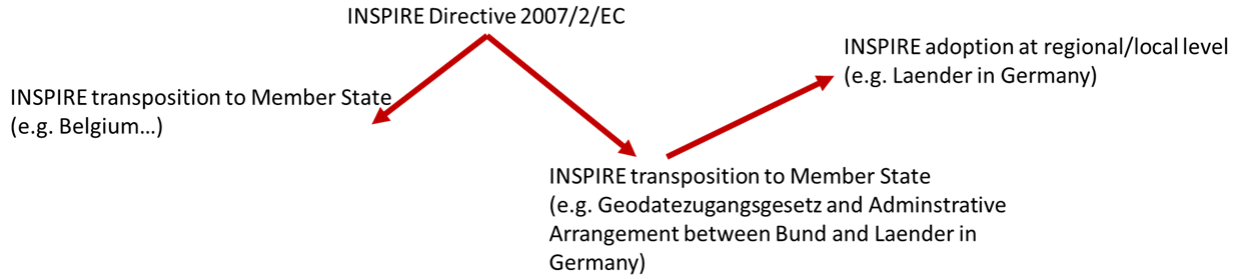
The following scenarios constitute several variants of that use case, with different user profiles and document types.

4.1 Scenarios

4.1.1 National Transpositions

- **Who?** Policy analysis of the INSPIRE Directive (and possibly national transpositions).
- **What?** To interconnect the separate obligations of the legal text with the according implementing rules, but also other legal texts that may influence the statement in the INSPIRE Directive itself (possibly including those particular to a given Member State).
- **Why?** In order to ensure that the overall policy context is met and that all obligations are followed up (probably also in legal texts in a given Member State).

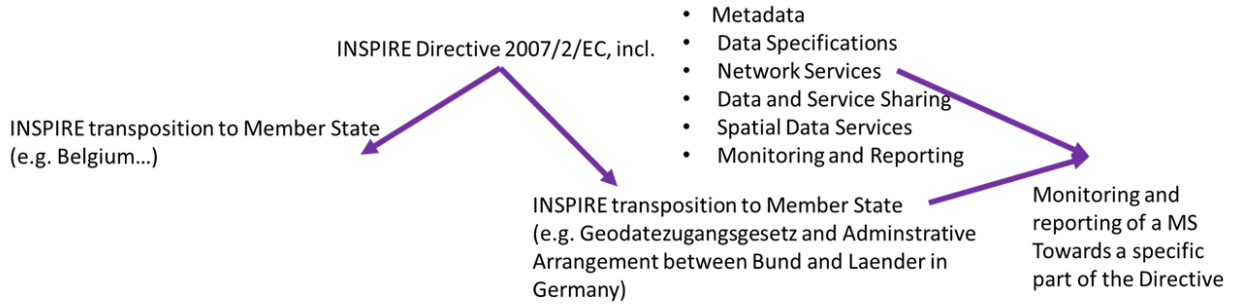
Example:



4.1.2 Member states reporting

- **Who?** Policy makers dealing with the INSPIRE Directive.
- **What?** To annotate the EU legal documents and passages thereof with the respective documentation of transpositions in the Member States (depending on the particular organisational structures).
- **Why?** To reveal the diverse approaches to implement INSPIRE on national and subnational levels and to keep an overview of what could already be reached in different Member States.

Example:

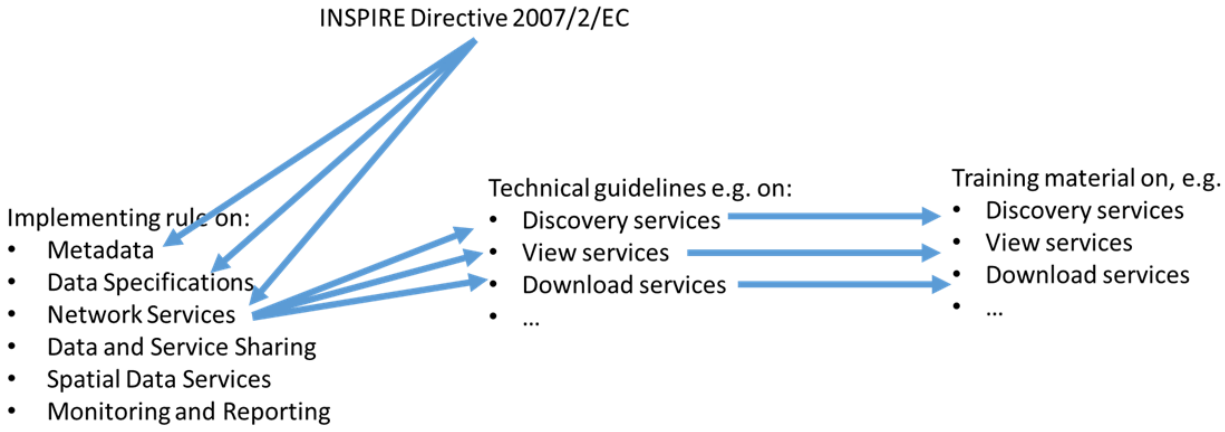


4.1.3 Technical guidelines and training material

- **Who?** Trainers about INSPIRE Directive.

- **What?** To annotate passages of the legal document with subsequently published implementing rules, latest versions of technical guidance documents, and examples of implementations.
- **Why?** To explicitly interconnect the separate documents and provide a single visual access point to the main components, starting from the INSPIRE Directive as a root (the legal basis).

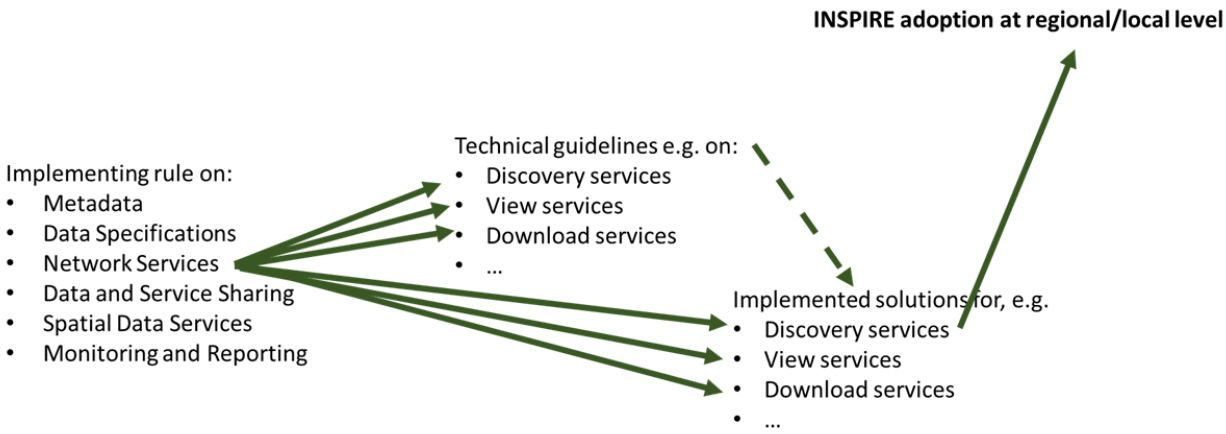
Example:



4.1.4 Sharing technical solutions between implementers

- **Who?** Implementers of the INSPIRE Directive.
- **What?** To annotate passages of the legal document with their respective solution…
- **Why?** To illustrate that/how they meet the requirements that are laid down in the directive, and to communicate and share their interpretations with colleagues (e.g. in other Member States, other local administrations or intermediate administrative units.).

Example:



4.2 Relevant data sources and media types

This use case involves mostly PDF documents, but also video streams (for the training material) and geospatial services (for the implementation reports).

More specifically, the following categories of datasets are involved:

- The Inspire directive itself: PDF document hosted at http://eur-lex.europa.eu.
- The implementation rules: PDF documents hosted at http://eur-lex.europa.eu.
- The national and regional transpositions of the directive into legal texts: PDF documents hosted at member states governmental websites.
- The member states implementation reports: PDF or XML documents hosted at member states governmental websites.
- The geospatial services that constitute the implementation: OGC services hosted at member states governmental websites.
- Other: press releases, scientific papers, conference proceedings.

5 References

[1] Duchesne P. (2013) *Aggregating media fragments into collaborative mashups*, W3C, Open Data on the Web Workshop '13, http://www.w3.org/2013/04/odw/odw13_submission_28.pdf

[2] Berners -Lee T., Fielding R., and L. Masinter (2005) *Uniform Resource Identifier (URI): Generic Syntax*, IETF, RFC 3986, STD 66. http://tools.ietf.org/html/rfc3986

[3] Mannens E, Troncy R, Pfeiffer S, Van Deursen D (eds) (2012) *Media fragments URI 1.0*, W3C Proposed Recommandation. http://www.w3.org/TR/media-frags

[4] Ciccarese P., Sanderson R. and Young B. (2015), *Web Annotation Data Model*, W3C, Working Draft 15. https://www.w3.org/TR/annotation-model

[5] Duchesne P. (2014) *Stitching data mashups from geodata fragments*, W3C, LinkedGeoData Workshop '14, http://www.w3.org/2014/03/lgd/papers/lgd14_submission_56

[6] Archer P., Goedertier S., Loukas N. (2012) *Study on Persistent URIs,* http://philarcher.org/diary/2013/uripersistence