HIRMEOS
High Integration of Research Monographs in the European Open Science infrastructure

HIGH INTEGRATION OF RESEARCH MONOGRAPHS IN THE EUROPEAN OPEN SCIENCE INFRASTRUCTURE

WP5

Technical specifications for the Open Annotation Service

01/03/2018

The project has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement 731102

HIGH INTEGRATION OF RESEARCH MONOGRAPHS IN THE EUROPEAN OPEN SCIENCE INFRASTRUCTURE

Deliverable D5.1

Technical specifications

Grant Agreement number	:	731102
Project acronym	:	HIRMEOS
Project title	:	High Integration of Research Monographs in the European Open Science infrastructure
Funding Scheme	:	EINFRA-22-2016
Project's coordinator Organization	:	CLEO-CNRS
E-mail address	:	pierre.mounier@openedition.org
Website	:	http://www.hirmeos.eu
WP and tasks contributing	:	WP5
WP leader	:	Ubiquity Press
Dissemination level	:	PU
Due date	:	28/02/2018
Delivery date	:	9/04/2018 (Postponed 1 month)

Contents

I. Context

Ubiquity Press (UP) is developing an annotations service (T5.1). The software that collects the annotations will run in their servers, and a public API will be provided (T5.2). Publishers will be able to register DOIs in the service developed in WP6, and then perform requests to the WP6 API to obtain the annotation metrics per DOI.

The API will provide metrics for annotations along with the existing metrics which will be implemented as part of WP6.

UP will also provide a JavaScript library that generates a widget that will be used to annotate the HTML, ePub and PDF version of a chapter, and help the partners in getting it to work on their platforms (T5.4-7).

II. API Standards

Since the main client of this API will be the JavaScript metrics widget, it makes sense to use JSON as the API format.

Timestamps

Ideally we would like to have a consistent date/time format for responses, we could use the standard "Complete date plus hours, minutes and seconds" as specified by W3C, e.g. 1997-07-16T19:20:30+01:00

Response standards

Along with the actual result of the query, the response should also incorporate some sort of **status** in order to let the client know whether the query was successful or not. These statuses could either be codes (e.g., 200), strings (e.g., success), or both (e.g., `{"code": 200, "status": "success"}`).

III. Frontend implementation

The partners will be provided with a JavaScript libraries they can embed in their websites, to create annotations around the HTML, ePub and PDF version of the document.

When choosing frontend technologies, the aim is to re-use existing open source projects / libraries, so that any effort in improving the code whenever needed, will be made available to the wider community, minimising the fragmentation.

The development of the chosen technologies / libraries is supported by well established institutions, with a strong reputation for developing future-proof open source projects: mainly Hypothes.is and CrossRef.

Ubiquity Press will set up a CDN to serve the required frontend libraries and guarantee a fast response, in order to allow the partners avoid performance degradation when loading the content.

The aim of the CDN is:

- to pin the JavaScript libraries to a specific version, which has been proven to work in the context of the project
- Ubiquity Press will test new releases of the libraries, and plan the upgrades to the new versions, letting the partners know when an upgrade is due, and if any modification on their pages is required in order to get the upgrade to work

All the widgets described in the following sections are reporting annotations to the live Hypothes.is backend, which is only exposing public annotations in the API which will be accessed by WP6. This means that metrics will only reflect public annotations.

Hypothes.is does not associate an annotation to a DOI, but to the URL where the annotation has been created. The CrossRef Event Data service will take care of associating an annotation with the DOI (see next section).

HTML

The JavaScript library provided will be a specific version of the one provided by Hypothes.is (https://github.com/hypothesis/client).

Detailed instructions on how to add the client to the page are available on the official documentation (https://h.readthedocs.io/projects/client/en/latest/publishers/embedding).
The Ubiquity Press Tech Team will help the partners to overcome any technical problem they might incur in while implementing the widget.

ePub

The version will be selected looking at the most recent fully functional integration between EPUB.js and Hypothes.is available (https://github.com/futurepress/hypothesis-reader). This integration is based on the upcoming version of the EPUB.js library (v3).

PDF

Hypothes.is has an official integration with pdf.js (https://github.com/hypothesis/pdf.js-hypothes.is). The repository contains instructions on how to embed the PDF viewer in the page, and enable Hypothes.is annotation.

IV. Backend implementation

a. CrossRef Event Data API

The CrossRef Event Data service scrapes DOIs to associate them with information coming from different data providers, including Wikipedia, Twitter and Hypothes.is. It is already used in WP6 to gather part of the information needed in HIRMEOS, and the code of the service is completely open source.

Regarding Hypothes.is specifically, the association between the URL of the annotation and the DOI is made looking at the Dublin Core metadata available in the page when it is scraped (see https://www.eventdata.crossref.org/guide/data/transparency-authority-trust/#landing-page-to-doi-mappings). It is thus important that the partners will have properly formatted DC information embedded in the HTML, in order to get a successful association of the URL to the DOI (and of the annotation metrics on the DOI).

At the most basic level, only a DublinCore meta tag should be strictly required, e.g.:

```
<meta name="dc.identifier" content="10.11647/OBP.0111"/>
```

but Ubiquity Press recommends (and will assist with the setup of) proper DublinCore integration.

CrossRef maintains a public list of all the publishers being scraped (see https://www.eventdata.crossref.org/guide/data/artifacts and https://www.eventdata.crossref.org/guide/service/artifact-registry); Ubiquity Press will make sure that all the partners' content URLs are correctly registered as data sources.

b. Chapter-level DOI

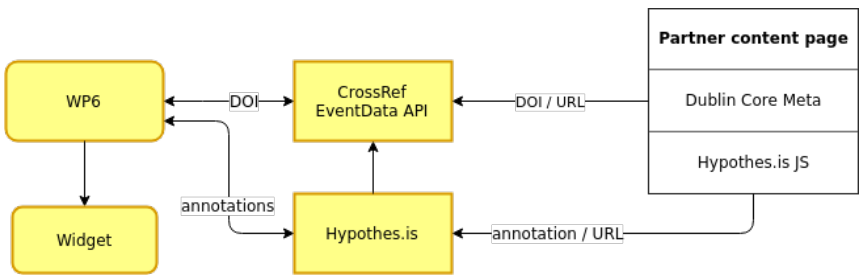
In the context of monographs, it happens very often that a single chapter could have its own DOI. This will lead the CrossRef Event Data service to collect metrics around the chapter, not associating them with the DOI of the book.
In the scope of future improvements, WP6 is set to build an internal hierarchy of DOIs (see WP6 tech specifications, section III.D), which will enable the aggregation of chapter-level metrics into book-level metrics. See WP6 technical specifications for further information.

V. API queries

a. Workflow

The WP5 is implemented as part of the CrossRef Event Data plugin which will be added to the WP6 source code (available at https://github.com/hirmeos/metrics/tree/master/src/providers/crossref).

- **WP6**: The publisher registers an account on the proposed Altmetrics service.
- **WP6**: After logging in, the publisher loads a CSV containing the following columns
  - URL of the published document (chapter, book)
  - DOI of the document; used as PK
- **WP6**: After a new CSV file has been successfully uploaded, the service launches a scraper, which will process all the active plugins in the system; the CrossRef Event Data plugin is active and will perform a query to the data source in order to retrieve (among other statistics) the annotations on all the DOIs in the database.
- **WP5**: the response from the CrossRef Event Data API contains the unique identifier of the annotation in Hypothes.is; the annotation is fetched from the Hypothes.is API, saved in the database and labelled as such, thus enabling aggregation, in order to expose the annotations-related metrics in the WP6 API.



An example of the data returned by the CrossRef EventData API related to Hypothes.is (describes an annotation on a specific DOI):


```
{
  "license": "https://creativecommons.org/publicdomain/zero/1.0/",
  "obj_id": "https://doi.org/10.1007/bfb0105342",
  "source_token": "8075957f-e0da-405f-9eee-7f35519d7c4c",
  "occurred_at": "2015-11-04T06:30:10Z",
  "subj_id": "https://hypothes.is/a/NrIw4KlKTwa7MzbTrMAyjw",
  "id": "00044ac9-d729-4d3f-a2c8-618bcdf1d252",
"evidence_record": "https://evidence.eventdata.crossref.org/evidence/20170412-hypothesis-de560308-e500-4c55-ba28-799d7b272039",
  "terms": "https://doi.org/10.13003/CED-terms-of-use",
  "action": "add",
  "subj": {
    "pid": "https://hypothes.is/a/NrIw4KlKTwa7MzbTrMAyjw",
    "json-url": "https://hypothes.is/api/annotations/NrIw4KlKTwa7MzbTrMAyjw",
    "url": "https://hyp.is/NrIw4KlKTwa7MzbTrMAyjw/arxiv.org/abs/quant-ph/9803052",
    "type": "annotation",
 "title": "[This article](http://arxiv.org/abs/quant-ph/9803052) was referenced by ["Decoherence"](http://web.mit.edu/redingtn/www/netadv/Xdecoherenc.html) on Sunday, September 25 2005.",
    "issued": "2015-11-04T06:30:10Z"
  },
  "source_id": "hypothesis",
  "obj": {
    "pid": "https://doi.org/10.1007/bfb0105342",
    "url": "http://arxiv.org/abs/quant-ph/9803052"
  },
  "timestamp": "2017-04-12T07:16:20Z",
  "relation_type_id": "annotates"
}
```


In this example, the `pid` field contains the unique ID of the annotation in Hypothes.is and `title` holds the content of the annotation.

While the content of the annotation is directly available through the CrossRef EventData API, a dedicated Hypothes.is client will be used to query the Hypothes.is API in order to get the content of the annotation. This is done solely for the purpose of creating an Hypothes.is API Python client which will be available externally to the WP5 or WP6 projects, and to keep the WP5 implementation future-proof in case CrossRef will change the API in the future (the EventData API is still in beta phase).
The behaviour of the annotation API queries is described in section V at the end of this document.

b. Exposing the results

In a fashion similar to the tech specifications for WP6, the annotations results will be exposed by the WP6 API in different ways (note that the filters applied are source-agnostic and the source of the metric is not the main view of this endpoint):

1. **by source**: the response contains an array of metrics sources (e.g. Hypothes.is) and the related number of annotations at the datetime of the request; useful for a text-based widget.

   ```
   GET: /altmetrics/{doi}?view=source&source=hypothesis
   ```

   ```
   {
          "count": 2
          "next": "https://example.org/altmetrics/{doi}?[filters]&page=3",
          "previous": "https://example.org/altmetrics/{doi}?[filters]&page=1",
          "total_quantity": 3,
          "results": [
                 {
                        "id": "41e4544f-2adf-418f-8871-8f7e2853ec15",
                        "source": "hypothesis",
                        "quantity": 1,
                        "scraped_on": "2017-08-08T13:42:55+00:00"
                 },
                 {
                        "id": "aed09375-4fe1-4e75-9a03-d24387b1e653",
                        "source": "hypothesis",
                        "quantity": 2,
                        "scraped_on": "2017-07-08T13:42:55+00:00"
                 },
          ]
   }
   ```

2. **by time**: the response contains an array of datetimes representing the different scrapes, and the relevant metrics related to each source; useful for a graph-based widget.

   ```
   GET: /altmetrics/{doi}?view=time&start_date={dt}&end_time={dt}
   ```

   ```
   {
          "count": 2
          "next": "https://api.example.org/accounts/?page=3",
          "previous": "https://api.example.org/accounts/?page=1",
   ```

```
        "results": [
                {
                        "id": "41e4544f-2adf-418f-8871-8f7e2853ec15",
                        "timestamp": "2013-11-18T13:42:55+00:00",
                        "hypothesis": 1,
                        "facebook" : 0
                },
                {
                        "id": "aed09375-4fe1-4e75-9a03-d24387b1e653",
                        "timestamp": "2013-11-19T13:42:55+00:00",
                        "hypothesis": 0,
                        "facebook" : 1
                },
        ]
}
```


Additionally, the following endpoints could be provided:

3. **by source, expanding context**: the response will contain the whole content of a citation on a social media (e.g. the tweet object or Facebook post)
   `GET: /altmetrics/{doi}?view=source&expand=context&source=hypothesis`


```
{
        "count": 2
        "next": "https://example.org/altmetrics/{doi}?[filters]?page=3",
        "previous": "example.org/altmetrics/{doi}?[filters]?page=1",
        "total_shares": 2,
        "results": [
                {
                        "id": "41e4544f-2adf-418f-8871-8f7e2853ec15",
                        "source": "hypothesis",
                        "quantity": 1,
                        "scraped_on": "2017-07-08T13:42:55+00:00",
                        "objects": [
                                {
                                        "ref": "894904696690941952",
                                        "content": "annotation text",
                                        "created_at":"2016-07-08T13:42:55+00:00"
                                }
                        ]
                },
                {
                        "id": "aed09375-4fe1-4e75-9a03-d24387b1e653",
                        "source": "facebook",
                        "shares": 1,
                        "scraped_on": "2017-11-19T08:18:31.809000",
                        "objects": [
                                {
```

```
                                        "ref": "383407285386411",
                                        "content": "post text http://ddd.com",
                                        "created_at":"2017-07-01T13:42:55+00:00"
                                }
                        ]
                },
        ]
}
```


4. **by scrape ID**: using the unique identifier of a scrape process (which will be executed at regular intervals); items having the same scrape ID have been scraped together.
   `GET: /altmetrics/{scrape_id}`


```
{
        "count": 2
        "next": "https://example.org/altmetrics/{doi}?[filters]?page=3",
        "previous": "https://example.org/altmetrics/{doi}?[filters]?page=1",
        "total_shares": 3,
        "results": [
                {
                        "source": "hypothesis",
                        "quantity": 1,
                        "scraped_on": "2017-07-08T13:42:55+00:00"
                },
                {
                        "source": "facebook",
                        "shares": 2,
                        "scraped_on": "2017-07-08T13:42:55+00:00"
                },
        ]
}
```


VI. Annotation queries in WP6

The purpose of this endpoint is to provide a response of all the annotations for a specific DOI from CrossRef (related docs). Differently from the endpoints and filters defined in the previous section, the main objective of this view is to show annotations specifically. The syntax of the filtering allows to add annotations from different sources in the future, if needed (e.g. view=annotations&source=hypothesis).

The default ordering of the response data is by time. The main content of the response data will be an array of annotations, along with some information about the annotation.

`GET: /altmetrics/{doi}?view=annotation&start_date={dt}&end_time={dt}`

```
{
        "count": 2
        "next": "https://example.org/altmetrics/{doi}?[filters]?page=3",
        "previous": "https://example.org/altmetrics/{doi}?[filters]?page=1",
        "results": [
                {
                        "published": "2017-07-08T13:42:55+00:00",
                        "id": "41e4544f-2adf-418f-8871-8f7e2853ec15",
                        "body": {
                                "source_id": "qNv_Ei5kngsfgf54GXdF63",
                                "scraper_id": "00054d54-7f35-4557-b083-7fa1f028856d",
                                "text": "The various scenarios presented should be specified as
                        being global averages of expected sea level rise. The sea level rise
                        observed locally will vary significantly, due to a lot of different
                        geophysical factors.",
                                "license": "https://creativecommons.org/zero/1.0/",
                        },
                {
                        "license": "https://creativecommons.org/publicdomain/zero/1.0/",
                        "published": "2015-07-09T13:42:55+00:00",
                        "source_id": "aBv_Ei5ZSnWOWO54GXdFPA",
                },
        ]
}
```