October 10th 2018

CO.SHS 2018 Workshop

Summary

CO.SHS

CO.SHS Workshop

Wednesday, October 10th 2018, Université de Montréal

The mid-term edition of the CO.SHS workshop was held on Wednesday, October 10th 2018 at Université de Montréal. It provided the various teams participating in the technological developments around the **open cyberinfrastructure for the humanities and social sciences** with the opportunity to showcase the progress they have made over the course of this past year, to address the challenges and problems they met along the way, as well as to highlight their accomplishments. From a broader standpoint, this workshop was also a call for further reflection on the potential integration of some of the projects to the Érudit platform.

→ To find out more about the developed research projects and tools, the teams and partners involved in the project, the corpora made available to researchers and all of the latest developments, visit **co-shs.ca**

PRODUCTION

Nicolas Sauret and Marcello Vitali-Rosati, Canada Research Chair in Digital Textualities – **Stylo: A Semantic Editor for the Humanities and Social Sciences**

→ See the presentation

→ Test the Stylo prototype

→ View Stylo's documentation

→ See Stylo's source code

Over the past year, a functional prototype of the Stylo tool has been made available online. The tool is a combination of a text editor with internal markup, a metadata editor, and a bibliography management tool. It thus has a module for conversion into various formats and a versioning module, therefore allowing to generate multiple publications from a single document.

Stylo is now used in its entirety by the *Sens public* journal, taking over the production chain already being used by the editors. The authors are progressively being invited to use the tool in its annotation phase (proof preview). The members of the Canada Research Chair in Digital Textualities will pursue their experimentations with six pilot journals as part of the Revue 2.0 project (revue20.org) and wish to convince more editorial teams to try out the tool so that they can adapt and improve their prototype.

Robin Varenas and Sylvain Aubé, NT2 Lab – **SARQC: A Support Structure for Canadian and Quebec Journals**

→ See the presentation

→ See the source code for the Views OAI-PMH module: drupal.org/project/views_oai_pmh

→ Read last August's interview

The NT2 Lab's efforts this year have mainly been focused on interoperability with the Érudit platform based on the case of the *Captures* journal – one of the four scholarly and cultural journals part of the pilot project, along with *Voix et images*, *Lettres québécoises* and *Estuaire*. *Captures* already owns a website (revuecaptures.org) and the work undertaken up until now was mainly aimed at data exportation towards Érudit.

In order to achieve this, the team members worked on structuring Drupal content and on developing a module named Views OAI-PMH, which allows to convert a document from HTML to XML format (specifically towards the EruditArticle schema, unique to Érudit) and to generate an OAI-PMH output for the journal's data dissemination. The team is currently in the validation phase of the results of the OAI-PMH input with Érudit and is developing a Drupal distribution that will be interoperable with *Lettres québécoises*.

Davin Baragiotta, Érudit – **OJS on Érudit: Efforts for CO.SHS**

→ See the presentation

→ See Érudit's source code repositories github.com/erudit and gitlab.erudit.org

→ See the EruditArticle schema documentation

Several technical initiatives related to the production of XML journals were undertaken by Érudit over the past year, namely in collaboration with the Canada Research Chair in Digital Textualities, the NT2 Lab and the Public Knowledge Project (PKP). The goal is to improve Érudit's production chains by focusing on upstream processing and automating part of the processing while ensuring tight quality control.

At this time, 7 OJS2 journals have been added to Érudit in minimal processing, 15 OJS3 journals are soon to be minimally processed, and the complete processing of OJS3 journals is under development. The main challenges remain the extraction of bibliographical references, the semi-automatic structuring of main text and the migration of the XML EruditArticle schema towards JATS (Erudit Publishing Schema).

Juan Pablo Alperin, Public Knowledge Project – **PKP and XML**

→ See the presentation

→ Find out more about the Substance Consortium

→ Download Texture

PKP's ultimate goal is to get all of the 10,000 journals using OJS to produce full XML markup—all around the world, in all disciplines. To accomplish this, they need to achieve two things: the automation of XML creation, which is very complex; and the integration of XML into journals' workflows, which is simpler. Their current efforts are aimed at the latter.

The Substance Consortium, including PKP, Érudit, SciELO and eLife, is a community-driven initiative supporting the continuous development and maintenance of open-source tools for XML integration. The primary activity is the development of Texture, an online text editor producing JATS-XML documents. The first version was released during the SciELO 20 Years Conference last September. The next steps will be to introduce the XML files further upstream in the workflow on OJS3: from the end of production to the beginning.

However, the problem of the automation of XML creation remains. Open Typesetting Stack is still best markup solution for the moment, but it is very hard to maintain and to scale. Other solutions—like Grobid—will be explored by the Semantic Extraction Working Group.

DISCOVERY

Fabrizio Gotti, Philippe Langlais and Vincent Letard, RALI Lab – **Open Information Extraction**

→ See the presentation

→ Find out more about the extraction of information triples: arXiv:1809.08962

→ Read last September's interview

The RALI's work this past year was mostly focused on improving the extraction process for concepts and verbal relations. One of the greatest challenges of OIE (open information extraction) is the extraction of concepts expressed using paraphrases or non-verbal clues, which are very complex to grasp for a machine. In order to assess the extent of the problem, the members of the team created a benchmark by manually annotating information triples found in 57 sentences (Léchelle, Gotti, and Langlais, 2018). They are currently evaluating the potential of paraphrases to assist the extraction process, with the intent of injecting them into their Distylium extractor, which will work for both English and French.

The team is also developing internal links, that will allow to make article suggestions to readers, and external links. Marginal annotations will also be studied.

Luis Meneses and Ray Siemens, ETCL (Electronic Textual Cultures Lab) – **Social Media Engine**

→ See the presentation

→ Test the Social Media Engine prototype: sme.dhinstitute.ca/solr/sme/browse/?q=cinema

→ Read the interview from last March

The idea behind the Social Media Engine is that papers and knowledge should be connected, and it aims to change the research experience provided by digital libraries or open access repositories accordingly. It allows for a reorganization of research results on a specific topic according to the number of times the documents have been mentioned or shared on social media. It is based on a combination of topic modeling and TF-IDF and uses an Altmetric.com (Digital Science) data dump. In the past year, the ETCL team members have completed and improved the prototype and user interface. They also started the process of a user evaluation cycle—both qualitatively and quantitatively. The team has been very active in the Digital Humanities research community.

The ETCL will continue their work in this direction, especially to enhance the user interface and to expand the conversation.

Juan Pablo Alperin, Public Knowledge Project – **Altmetrics libres : Paperbuzz**

→ See the presentation

→ Test the Paperbuzz web service

→ See the source code for the API and PaperbuzzViz

→ Find out more about the Facebook project: arXiv:1809.01194

Since 2017—with the launch of the new Crossref Event Data service—PKP has been working on an open source altmetrics project called Paperbuzz. It is a community-driven solution, as opposed to the existing for-profit companies such as Altmetric.com (Digital Science). Crossref gathers Twitter, Wikipedia, Facebook, Reddit and WordPress events and gives access to this data at a very granular level, but doesn't calculate metrics. This is why PKP partnered with the nonprofit Impactstory to build a service that aggregates these events into metrics and makes them available through an API. PKP has also developed a Javascript visualization library that uses the Paperbuzz API to display the metrics by day, month, year and total. PaperbuzzViz is now ready to use and will be available for all OJS journals through a plugin and on the articles of journals hosted by the Open Library of Humanities.

However, the altmetrics data collection area is still full of challenges, especially since the big social media platforms are black boxes that do not disclose the way they gather and map the data. To solve this problem and to collect better quality Facebook data, PKP started another independently run project, in order to eventually push it back into Crossref Event Data.

EXPLORATION

Christopher Collins, Vialab – **Vialab Project Overviews: Interactive Tools for Humanities and Social Science Researchers**

→ See the presentation

→ Test the Textension tool

→ See the source code for various Vialab projects

The Vialab team is working on a variety of research and visualization tools designed for the humanities and social sciences. In the past year, the team members have completed the development of the Textension tool, which will soon be fully deployed. A second project, which has been progressing at a good pace, is called Slow Analytics (presented last year as Document Analyzer). It offers new ways to dive into a large corpus such as Érudit's; it allows for example to upload a document of interest and to use its terms in the search engine.

Three entirely new projects were started over the course of the past year. The Érudit Knowledge Map aims at mapping the knowledge transfer that occurs in the Érudit corpus, such as co-authorship networks in specific journals, or by institution, date, author or paper title. Document Matching is intended to help individual users to find which journal is most interesting for them by dragging and dropping a relevant document. Citation Galaxies is a visual analytics tool for bibliometrics that helps with citation context analysis. It also allows to manually build sentiment datasets.

Maxime Sainte-Marie, Canada Research Chair on the Transformations of Scholarly Communication – **Cleaning up the BAnQ Collection: The Current State of Affairs**

→ See the presentation

The approach presented last year for the cleaning of BAnQ's digitized journal and newspaper corpora turned out to be a dead end, as the digitization quality strongly varies from one document to the other and is often very poor. The documents, containing a strong proportion of redundancies, are mainly images in JPG format, as well as PDF and TXT files. Among the latter, some files are simply empty PDF conversions without textual markup, while others are completed and ready to use.

It is therefore necessary to reverse-engineer the corpus in order to reduce its size, and to follow this up with large-scale image processing, while keeping in mind that this technique is a significant challenge in itself. Only after completing these crucial tasks will the OCR and linguistic post-processing be possible.

Vincent Larivière, Observatoire des sciences et des technologies – **Érudit: From Publishing Platform to Bibliometric Database**

→ See the presentation

→ Find out more about the OST's services and databases

Over the course of the academic year 2017-2018, the Observatoire des sciences et des technologies (OST) team has officially launched the citation index project based on Érudit's scholarly journal corpus. As a reminder, a citation index was created in collaboration with Érudit, the OST and the Agence Universitaire de la Francophonie between 2015 and 2017, but the progress achieved as part of CO.SHS will allow for the constitution of a comprehensive and useable relational database. The structure of the tables will be similar to the one used by the OST for its other databases.

The database is currently being populated with Érudit data. One of the critical elements of the project will be the disambiguation of author's names, affiliations, and journals. This process will be partly automated with the algorithms used by OST every year to build the Web of Science database, but significant manual cleaning will also be necessary. Several types of research will be made possible by this citation index, namely on the position of women in the humanities and social sciences in Quebec.