DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 6: Case Studies

6-Case Studies

OpenEdition: Towards a European infrastructure for open access publication in humanities and social sciences

Pierre Mounier, EHESS & OpenEdition, France

OpenEdition

OpenEdition is a public infrastructure based in France since 1999. It is dedicated to SSH open access publication and scholarly communication. We are supported by 4 higher education and research institutions: CNRS, Aix-Marseille University, EHESS and Avignon University. It is exclusively dedicated to the dissemination of Humanities and Social Sciences research in *open access*. We think that dissemination on the web, not only on the Internet, is really important because it is a way to encourage and foster uptake for this kind of content in the different communities, scientific or not.
We are based in France but we are working to make our platforms and infrastructure more international, so we are working with an international network of partners in different European countries and even outside Europe. For example we have a partnership with Torino University in Italy to develop a specific program: OpenEdition Italia. We also have a partnership in Lisbon for the development of Portuguese content; another one in Germany with the Max Weber Stiftung; in Spain with the Uned University; and we also have partnerships outside Europe: In Canada with the Public Knowledge Project (PKP) which develops the Open Journal System (OJS) and in the US.

OpenEdition is an infrastructure and also a portal: openedition.org. This portal gives access to four platforms. We created and designed a platform for each type of document or information we want to disseminate:

1. Revues.org: it was the first platform created in 1999 and it is dedicated to journals. On this platform, we disseminate many open access journals from different countries and from all disciplines of Social Sciences and Humanities: history, philosophy, sociology, anthropology, geography, etc. Even though we are based in France, we are working with scientific committees in the different countries who want to disseminate their journals in any language on the platform. So the platform is multilingual: French, Italian, German, English, Spanish and Portuguese.
2. OpenEdition Books: It is a platform we have set up four years ago to help academic publishers disseminate open access their catalogue of academic books. It is multilingual too and at the moment we are working with approximately 70 publishers.
3. Hypotheses.org: This platform is dedicated to academic blogging. It is very different from the previous platforms where the disseminated materials are peer-reviewed. Hypotheses.org is for direct communication -there is no peer-review- but it is still interesting because researchers, librarians, research teams, institutions, projects, laboratories can disseminate information about what they are doing. It is completely different from the publication of a book for example, it is in fact the result of several years of work before publication. The content of the book is certified to be good quality, but it publicly appears many years after the research has been done. With a blog, it is the opposite, as soon as a researcher or a research team carry on a research, then they can disseminate their information by themselves about what

they are doing at the very moment they are doing it. So access to information is much quicker and more direct. Plus, as it is blogging, you have a commentary function, so the readers can directly comment on the blog and then you can have interactivity, some sort of a scientific discussion between authors of the blog and the readers. It is actually really complementary to the two firsts platforms.

4. Calenda.org: it is simple but really useful, it is a scientific agenda for SSH. It means that every organiser of a scientific event (workshop, seminar, conference, call for paper, etc.) can disseminate on this platform the information. DARIAH is one of the main partners of the platform, for example all events organised by and though DARIAH are disseminated as well on this platform.

Some figures

Revues.org hosts at the moment 437 journals, 100 600 documents on open access. There is a wide diversity in terms of languages, represented countries and disciplines.
For the books, as it is a younger platform, we disseminate at the moment around 3 000 books and 2 500 are open access. It is growing rapidly as we are going to work with more and more publishers.
For hypotheses.org, when we opened the platform in 2008, it was a bet because we didn't know if there would be any uptake in the community around this form of new communication. We had some professors saying that it would never work because "*blogs are for kids or teenagers*", "*not for academic content because it is not peer reviewed*". But we open the platform to see how it goes and it was a very good surprise because we now have more that 1 600 living blogs producing or having published 15 200 posts in open access. So, the uptake was enormous in fact and the most interesting thing about this platform is that the scientific community invented its own usage of the platform - we didn't anticipate it. For example, we didn't think people organising a seminar could use it for a seminar to announce new sessions, to record and spread sessions, to publish the bibliography of the seminar, to continue the discussion on the blog, etc. And it is the same with research projects, some ERC and ANR projects opened a blog to disseminate information about their project. Some of them opened a blog even before the project started. You can easily open a blog and start publicising information about an idea, maybe gathering new partners and then applying for funding. It is not compulsory to have funding to open a blog, you can open it before and it can be a good element on your file when you apply for fund.

Overview of visibility gained through the platforms for the disseminated content

This year, we are going to have a little bit more than 60 millions visits on the platforms, that is approximately 30 millions unique visitors. Those platforms are also meant not to be read and used only by scientific community and specialised scholars, but also by citizens, society at large. There are a lot of example of how those contents, like a chapter on a very narrow focused research monograph, are shared on Facebook or cited on twitter by academics and non-academics. This is the point of open access and of the web, because it can help improve and foster up taking by the community of this kind of content. The readership is worldwide in fact and disseminated in many countries.

Digital edition

What is *digital edition* and what is the difference between a *digital* and a *digitised* edition?

For us at OpenEdition:
Digital publishing means to publish on the web, not putting some pdf files on a server, but to *convert those files and content and put it on a nice layout to be read and accessed in a web browser*. The goal is give access to a very complex information stored inside SSH publication. With an article, you have a lot metadata, a very complex text, highly structured, you have many information to display, so it is a real work to have it and to present it in a nice and usable way on a web page. As a platform, a digital edition is to give access to the same text in many different formats because we know that the readers are disseminated around these formats. Some people prefer to have a pdf file, other prefer web pages or epub format, so we do it and people can have their content on reading devices, softwares, smartphones, etc. To produce a digital edition, you also need a digital publishing way of working: you also have to work particularly with giving access to the data stored inside a text. For example, the images which are accessible in a web page are also accessible in full resolution with metadata attached. But you can also embed a flash file, for example an interactive map, so the reader can show or hide different levels of information on this specific figure, or even to embed videos.

Collaboration inside the DH community

To carry on this publishing functionality, we have created and developed **Lodel**. It stands for LOgiciel D'édition ÉLectronique and it is published on Github as an open source publishing software, like OJS, but different. It is a CMS based on XML TEI. The content published on OpenEdition Books and on Revues.org is converted into TEI XML and stored into our databases. We can then automatically generate, from the XML, the HTML, ePub or pdf file. We use a TEI than what was previously presented because we do not encode primary material, we encode publications. We use a subset a different tags, not all tags from TEI, only a limited number. LODEL works by converting content. As we work in the humanities community, we know that most of the authors work with text processing softwares like Word or Libreoffice. So we help editors to structure the word file they receive using styles (titles, normal, citation, etc.). They structure the information into the word file according to our own schema and they can upload the word file onto the publishing server. Then LODEL transforms the word file into XML. After that, you have dissemination through different formats.
LODEL => Structuring the information of front, converting it into XML and publishing in different formats and into different channels.
LODEL takes into account all the specificities of the humanities content, so it is not like Wordpress or OJS, which are not specifically meant for SSH. LODEL is meant for this specific content and that is why we are inside the humanities community. For example, it generates identification for paragraphs, it manages footnotes and complex structuration of a text, it can manage bibliography, it can manage multiple indexes of keywords, geographical data, chronological data, hierarchical, ethnic populations, it is multilingual and attribute DOIs.
We are finally also part of the DH community as we publish the Journal of the Text Encoding Initiative.

R&D: OpenEdition Lab

- **Bilbo** is an automatic annotation software that can parse a text, detect and automatically recognise bibliographical references inside a text, analyse those references to divide between the name of the author, of the publication and then query the CrossRef database to see if there is DOI attached to this reference. At the end, Bilbo transforms a plain text bibliography, which is the norm in SSH, and then add a link to the bibliographical reference, so the bibliographic is no more a plain text but hyperlinked and it is a little bit more useful for the reader. We want to add this functionality into the footnotes and the next step will to try to detect fuzzy references inside the text to see if there is an online reference available to be linked to. This project received a Google Grant.

- **Agoraweb** is an automatic detection of book reviews. The idea is that inside publication and blogs, you have book reviews, authors writing articles or blog posts, which are in fact reviews of published books. In journals, it is easy to detect because there is a specific section "Book reviews". But in blogs, it is much more difficult to detect because a blogger never says "*this is a book review*". So you have to parse through whole blogs inside hypotheses.org to automatically detect if blog posts are book reviews or not. After that, with Bilbo we can see which book is reviewed and then make a link between the book and the reviews. The final aim of this project is to gather on OpenEdition Books the book by itself and at the bottom of the page all the available book reviews published in different venues, in journals, in blogs, or anywhere on the web. This is a really useful information for the reader. Demo website of the OpenEdition Reviews of Books: http://reviewofbooks.openeditionlab.org/
- **Open Peer Review**: an experiment proposed to journals which are traditionally [anonymously] reviewed: Julien Bordier, 2016. « Open peer review: from an experiment to a model: A narrative of an open peer review experimentation ». <hal-01302597>.

You can find further readings here: https://lab.hypotheses.org/bibliographie

OPERAS

Our very new initiative is Open access Publication in the European Research Area for the SSH (OPERAS). The idea with OPERA is to be able to work at European level. We try, with our partners in Europe, to extend this network and to make everybody work together to set up an infrastructure for open access publication in SSH. We defined inside the network some common goals, the idea is that the players in this field of open access publication in SSH are very small and fragmented, so we need to gather all those players to adopt common standards for example or to share research and development costs (because it is expensive). It is better to do it all together and to share the results, to identify and adopt best practices, to assess sustainable economic models and to advocate for open access in SSH. For now, we are 19 partners from 10 countries (Germany, UK, Netherlands, Spain, Portugal, Italy, Croatia, Luxembourg, Greece and France) and we are open to extend this network to other countries. Our first partner is the Association of European University Presses (AEUP): an association that gathers main University Presses at European level. The idea is to set up a cyber infrastructure for open access publication and more specifically for books, because in SSH most of the books are not in open access and not even digital. So there is a lot of work to make books *accessible, disseminated, visible, indexed and used on the open way*.

HIRMEOS

This project's name stands for High Integration of Research Monograph in the European research area for SSH (HIRMEOS). It has been submitted last year to the Horizon 2020 framework and it is going to start in January 2017. The idea is to gather 5 open access book publishing platforms (OpenEdition Books, Ubiquity Press, OAPEN, The Gottingen University Press and EKT platform) and to implement 5 sets of service at the same time:

- Identification at all level on 5 book platforms: DOIs, ORCID identifiers for the authors and FUNDREF to ease indexing to OpenAIRE.
- Certification of quality provided by the Directory of Open Access Books (DOAB).
- Implement automatic recognition service on named entities inside the full text of a book for places, names, periods, dates and maybe topics. Then it will be given back to the platforms to enrich their index or to link to DBpedia for example.
- Implementation of an open annotation service: the reader will be able to annotate line by line the full text of the books in open access. The reader will also be able to answer to the annotation with a *forum feature* inside the annotation service: when someone annotates a line or a sentence, then it is published on the web and someone else can answer to this annotation and the author of the annotation can answer to answers and so on. The idea here is to rise the uptake of this content by developing conversational features around the books, giving possibility to the reader to discuss about a book on the same website where the book is disseminated.
- Develop book metrics: implementation of Altmetrics for the books. Altmetrics are a new impact metric, invented by PLOS, to measure the impact of an article by counting not only the number of downloads or views on a platform, but also the numbers of shares on the social media such as Facebook and Twitter, in order to aggregate that and to make a new metric to measure impact. Annotations will be counted by the Altmetrics for the book metrics, because the number of annotations attached to a book is also an impact measure.

That is why we work on this project with different partners: for example we work with ORCID to attribute ORCID IDs, with CrossRef for the DOIs, with Hypothesis for the annotation plugin service, but also with Huma-Num, with DARIAH, with OpenAire for indexing our content, etc.

Contact

Pierre Mounier is deputy director of OpenEdition, a comprehensive infrastructure based in France for open access publication and communication in the humanities and social sciences. OpenEdition offers several platforms for journals, scientific announcements, academic blogs, and, finally, books, in different languages and from different countries. Pierre teaches digital humanities at the EHESS in Paris. He has published several books about the social and political impact of ICT, digital publishing and digital humanities.

Associate Director for international development OpenEdition
Coordinator of OPERAS: http://operas.hypotheses.org
ORCID: http://orcid.org/0000-0003-0691-6063
Twitter: @piotrr70
Email: pierre.mounier@openedition.org

Czech Literary Bibliography

Vojtěch Malínek, Institute of Czech Literature, Czech Academy of Sciences, Czech Republic

The Czech Literary Bibliography is a basic infrastructure for interdisciplinary research into literary culture of Czech lands with tradition since 1947. It is one of the largest and more opened research infrastructures for individual national literatures in Central Europe. All of our data are fully available online. We are supported by the Ministry of Education, Youth and Sports of Czech Republic and we are now included in the Czech Republic Roadmap of Large Infrastructures for Research, Experimental Development and Innovation. The hosting institution is the Institute of Czech Literature of the Czech Academy of Sciences.

The information sources we are processing are:

- set of bibliographical and other specialized databases (biographical base Czech Literary Figures, literary prices, book editions, etc.)
- documents collections and card index catalogues

Basic numbers

- over 2 000 000 bibliographical records (articles, conference proceedings, books, etc.)
- nearly 40 000 biographical entries about authors and literary scientists
- over 1 500 newspaper and journal titles processed
- data instantly available without any limitation online

Chronological range

Our bases cover the whole modern Czech literature, which means from 1770 to present, and data are continuously added and updated. Processed documents are mainly in Czech of course, but also in German, Slavonic languages, etc.

Range of disciplines

- Czech literature and literary studies
- other national literatures and associated humanities and social science disciplines (theatre studies, history, philosophy, linguistics, journalism, etc.)

Geographical scope

Concentrated on Czech lands, but we are working with bohemical literature published abroad, in Europe and USA for example.

Average usage rate

150–200 accesses per day

Main activities

- Processing of databases
- Digitisation and software development

Retrospective Bibliography of Czech Literature□

It is a large card index compiled from the 1950s to the 1990s that covers a chronological range from 1770 to 1945. This card catalogue consist of nearly 1.7 million records, with 530 titles of newspapers and journals mainly in Czech and German. Thematically, articles of following types are processed: Czech and world fiction, translations, journalism and specialist literature. The catalogue is organised by authors, reference, subject and identification section.

RETROBI software: http://retrobi.ucl.cas.cz/

During the “Retrospective Bibliography of Czech Literature 1770–1945 card index catalogue digitization project“ (2010-2012), we have developed a software for digitization and online presentation of card indexes. We have scanned and prepared OCR transcriptions of all of the cards. This allows full-text searches in text representation. We have also developed a tool for editing of text data available: it means that data are available for any user for corrections and editing (example). Cards could be corrected or rewritten as a whole or semi-automatically structured. For registered and skilled users, tools for large-scale editing of a chosen database field are available. RETROBI system enables complex queries, variable export options and offers administrator interface with several advanced functions, etc. Since 2012 it was used for digitising of 3 others large card catalogues in different institutes of CAS. RETROBI software was created with the kind support of the Ministry of Education, Youth and Sports as the result of the “Retrospective Bibliography of Czech Literature 1770–1945 card index catalogue digitization project“ (VZ09004), completed between 2009–2011 under the INFOZ programme.

Contemporary Bibliography (since 1945)

The Contemporary Bibliography is completely processed in database form. We are now finishing the conversion of older datasets into nowadays standards. So it now meets common standards for contemporary librarianship and bibliography (MARC21 exchange format, RDA cataloguing rules, Aleph software etc.). Everything is fully integrated to the national research information exchange networks. We are using various persistent identifiers.

Conclusion

This quite strongly structured data can be used for statistical and quantitative analysis of the literary field defined by the needs of researchers with regards for analysis of bibliographic data.

For example, you can search for the total number of records per year, it can put light on the way how political conditions can influence the number of records pro author, magazine etc.: after the of the Second War, the total number of records rapidly increased but after communists take power in 1948, the number significantly falls down, with its lower level in 1952.

You can make comparison of selected authors and analyse its connections to social contexts and measure the impact on their ranking with bibliographical data used in a quantitative and statistical way.

On the background of the data from RETROBI, the Top 10 Authors with highest number of records for the period 1770-1945 can be easily shown. The most “productive” of them was Arne Novák Arne with 7 370 occurrences.

Contact

Vojtěch Malínek, Institute of Czech Literature of the Czech Academy of Sciences

Institutional website: http://clb.ucl.cas.cz/

Facebook: http://www.facebook.com/ceskabibliografie
Email: malinek@ucl.cas.cz & clb@ucl.cas.cz

Turning the Polish Literary Bibliography into a Research Tool: Challenges, Standards, Interoperability

Maciej Maryl & **Piotr Wciślik** (Institute of Literary Research of the Polish Academy of Sciences)

Polish Literary Bibliography

The project we are currently working on is quite similar to the one described by Vojtech Malinek, Czech Academy of Sciences. Let me start with the use of the data, i.e. with some examples of research we would like to conduct on our data once the project is completed: The Polish Literary Bibliography – a knowledge lab on contemporary Polish culture.

Data-driven research into literary culture

The goal of our project is to make our bibliography not only easily accessible and searchable but also capable of serving as a discovery tool for researchers. We would like to enable them to perform similar tasks as in those studies which used rich bibliographical data for exploration of literary culture:

- Analysis of literary locations, e.g. comparison of the novels set in rural areas with the urban ones over time (see: Jockers 2013:45).
- Gender distribution of authors as compared with their coverage in the media. Katherine Bode has shown that since the 1990s more novels are being published by female authors, yet both critics and academic scholarship focuses more on male novelists. (Bode 2014: 132-134).
- Tracing co-publishing patterns as the source of the knowledge about the shape of literary networks. You can compare and interpret various affiliation networks from other countries (Long and So 2013 a: 150; 2013 b: 274).)
- Franco Moretti's analysis of the length of the titles in Victorian novels shows how the genre became standardised (Moretti 2013: 183). Moretti also traced the patterns in a genre lifecycle on the example of Victorian novels (Moretti 2005 15-19)
- Some existing bibliographical infrastructures already make some research tools available like exporting queries into csv format for further analysis (Australian Literary Bibliography) or tracing relationship between authors in the form of a network as in the case of Women Writers (NEWW).

In order to perform such research, we need reliable data, but our problem is that bibliographical data was collected not for research purposes but as reference material. So the data is prepared in a way that allows people better access to knowledge rather than statistical inferences. For instance, lots of information is preserved in unstructured annotations, which are difficult to be transformed into a database. Our challenge is to work with data collected as a reference resource, not research material□with the goal of forging a database ready to answer unpredicted questions, i.e. questions which were not taken into account when collecting the material.

PBL: Polish Literary Bibliography

The PBL is an annotated bibliography of Polish literature that has been published since 1954 as a project that is a comprehensive register of all articles, notes and other materials concerning Polish literary life. This can include different types of records: authors, works, reviews, articles, contests and awards, TV, radio, theatre and cinema adaptations, events, exhibitions and semantic annotations (e.g. ’literary theory’; ’exile literature’; ’psychology’; ’education’; ’censorship’), etc.
The Polish Literary Bibliography as a research tool on contemporary Polish culture is a□3-year project funded by the National Programme for the Development of Humanities (2015-2018) and is developed by Institute of Literary Research, PAS in cooperation with the IT partner: the Poznań Supercomputing and Networking Center (PSNC).

Goals

- Creating a database of 4 million bibliographic records on Polish literature (1939-2004)
- Integration of heterogeneous datasets (retrodigitisation of available data in printed format)
- Integration with Linked Open Data Cloud
- Data mining and visualisation tools for modelling processes of literary life□□

Challenges of ontology and retrodigitisation

Ontology
We apply schema.org ontology (with its BIB extension) in order to connect our data with Linked Open Data cloud. We would like to offer rich searching features that enrich an answer with additional information next to the results (cf. Google rich snippets). So we have to make connections between existing data and to retrieve automatically such information from the LOD cloud.

- Choosing the right data standard: Schema.org instead of bib-dedicated ontologies and data models (FRBR, RDA, BIBFRAME) because
  - it is not well equipped to handle theatre, cinematographic, radio and television instances of literary works (except FRBRoo) (=different events in the database, different ontologies)□
  - It is too complex to handle by metadata producers (including FRBRoo)□
  - Validity still TBC by community of practice

- Whereas Schema.org offers
  - a widely-used Internet standard
  - a robust model to handle PBL data model
  - not unprecedented in the library domain, there is a bibliographical extension with a careful list of ontologies elements and relevant vocabulary.

Process:

- Remediating the input tool (and habits of teams working with previous database)
- Converting the existing online database (1988-2002)
- Retrodigitising printed volumes (1944-1987; WW II; Bibliography of Samizdat).
- Integration with other bibliographies: we are figuring out how to map our entities, our bibliographical system onto schema.org ontology in order to ensure further cooperation with other national or regional bibliographies

Retrodigitisation of a huge collection in regard with the structure of the database
Challenges:

- Technology: Parsing & Lemmatization
- Tracing methodological inconsistencies
- Time factor
    - Subject-classification (new approaches to literary studies, e.g. gender studies)
    - new forms of literary life, e.g. online literary life
    - changing political geography: Yugoslavia and Soviet Union are no longer existing entities

Research challenges: local specificity

- Data-collection methodology has changed over the years
    - Selection of writers
    - Unrecorded debuts
    - Data censorship
- Geography
    - Changing borders
    - Literature in Exile
    - Polish literature = literature in Poland or in Polish?

- Shape of literary life: Official versus Underground publishing

Our overall goal is a conscious research into literary culture based on bibliographical data, which could be formulated as following: in order to come up with right research questions, one has to be sure to know how the data were selected, collected, censored, abridged, standardised, annotated, digitised, corrected and linked together.

References

Bode, Katherine. 2014. Reading by numbers. Recalibrating the literary field. London: Anthem Press.

Long, Hoyt and Richard So (2013a) ”Network Analysis and the Sociology of Modernism” boundary 2 40(2):147-182

Long, Hoyt and Richard So (2013b) ”Network Science and Literary History” Leonardo 3(46), 274-274.

Jockers, Matthew L. 2013 Macroanalysis. Digital Methods & Literary History, Chicago: Chicago UP.

McCarty, Willard. 2008. "Modeling in Literary Studies" A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, http://www.digitalhumanities.org/companionDLS/ (28.02.2016)
Moretti, Franco 2013 "Style, Inc.: Reflections on 7,000 Titles" Distant Reading, London: Verso.
Moretti, Franco 2005 Graphs, maps, trees. Abstract Models for Literary History, London: Verso.
Pacek, Jarosław, 2010, Bibliografia w zmieniającym się środowisku informacyjnym, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.
Shea, Christopher. 2008. "The geography of Irish-American lit" Brainiac [Blog], http://www.boston.com/bostonglobe/ideas/brainiac/2008/07/matthew_j_jocke.html (28.02.2016). □
Woźniak-Kasperek i Ochmański, red. 2009. Bibliografia : teoria, praktyka, dydaktyka : praca zbiorowa, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.□□

Contact

Maciej Maryl & **Piotr Wciślik** (Institute of Literary Research of the Polish Academy of Sciences)

Institutional website: http://chc.ibl.waw.pl/en/projects/pbl-lab/
Personal website: maryl.org
Twitter: @maciejmaryl
Email: maciej.maryl@ibl.waw.pl

Creation of Open Data Resources: Benefits of Cooperation

Kira Kovalenko, Institute for Linguistic Studies (Russia) & Austrian Centre for Digital Humanities (Austria) & **Eveline Wandl-Vogt,** Austrian Centre for Digital Humanities, (Austria)

I am a research fellow at the Institute for Linguistic Studies in Saint Petersburg and an invited researcher at the Austrian Academy of Science in Vienna, so I represent two organisation and I will be talking about cooperation between these institutions.

- Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences (Vienna), established in 2015: 1 department, 4 working groups, 50 researchers, director Dr. Karlheinz Mörth.
- Institute for Linguistic Studies (ILS), Russian Academy of Sciences, established in 1921: 6 departments, 120 researchers, director Prof. N. Kazansky

One of the biggest department of the Institute for Linguistic Studies is dedicated to lexicography. We create a lot of dictionaries, such as the dictionary of modern Russian language, the dictionary of the 18th and 19th century language and the Dictionary of Russian Dialects that started in 1965 and has so far 48 volumes published. It has more than 300 000 entries. The chief editor is prof. Sergey Myznikov, 8 compilers are working on it - and I am one of them. As a result of our discussion with Eveline Wandl-Vogt who is a member of the group compiling the Dictionary of Bavarian Dialects, we decided to combine our efforts on digitalisation of the dictionaries. Since then our plans for cooperation have enlarged, and now we have three common projects. The main aims of our cooperation are to:

- increase accessibility
- increase interoperability
- increase reusability
- enrich dictionary data
- interlink dictionary data
- create new workspaces
- open up dictionaries for research process and public curiosity

Projects

All started with the Dictionary of Russian Dialects and we decided to create an infrastructure for compilers and researchers. You can now see the dictionary *online* on the website of the institution and you will find almost all published volumes, but it is just a pdf format and of course we cannot correct the text or add new material to the volumes, etc. This is why we decided to create such infrastructure and we use TEI P5 to markup the dictionary. We also use Lemon model for searching information and for other technical issues we envisage Ontolex core model with extensions: *synsem*, *decomposition*, *variation* *and translation* and *linguistic metadata*. We are planning to have an infrastructure that we could add and from which we could extract what we need, automatically. It will be online, available for everybody and all users could have better access to the material.

Another project we are working on is the Russian Manuscript Lexicons infrastructure for researchers. Russian manuscript lexicons appeared as a new lexicographical genre in the middle of the 16th century and has been developed since that time. The infrastructure will include:

- alphabet arrangement according to their first apparition
- close connection to the [original] text
- more than 9 types
- from 700 up to 16 000 word entries
- foreign words from Greek, Latin, Hebrew, Church Slavonic, Ruthenian, Tatar, Arab, and German origin
- about 150 lexicons
- important source for historical lexicography

Approximately 15 of them can be found online on the National Library website, but in order to find something you have to look through a lot of pages and the process is very difficult if you need some particular words. During my PhD, I have manually copied some manuscripts representative for different types as a plain text, and now have them in text version; that is why it would be nice to have them in parallel text version and in facsimile version. We are planning to create such infrastructure where you could see this and then search necessary information. We are planning to use as well:

- TEI P5 to markup them
- cr_xq: a standards-based, fully configurable publication framework for XML data
- full-text search and field-specific searches
- synoptic view of facsimile and text
- computer-assisted collation and stemma creation
- facilitates creation of various indexes of tagged information in the ingested resources (lemma list, index of translations by language etc.)
- standards: METS, FCS-SRU, currently mostly used for TEI content (e.g. Austrian Baroque Corpus ABaC:us)

The last project is not started yet but it should begin next year, it is the Russian Plant Names in the Diachronic Aspect (from 11th to 17th centuries). It will be a database with a search engine. We applied for a grant and if we succeed, we would like to have such a database. We are a team of researchers from different background, we share interest for languages, folklore, literature, etc. We will use primary sources (manuscripts, printed books of the 11-17 cc.) and secondary sources (historical dictionaries, modern researches, etc.). So we are going to collect all this information and to create a database. We hope to have interoperability with the Austrian plant common names database (exploreAT!). It will also contribute to the project Biodiversity and Linguistic Diversity. In the end, it will be a collaborative Knowledge Discovery Environment. This database will include:

- Old Russian name
- modern Russian name

- foreign etymon (if loanword)
- type of representing a foreign phytonym (translation, transcription, transliteration, loan translation, hyperon)
- Latin name
- example of the use from text/dictionary
- bibliographical information (if printed: author, title, place of publication, date, page, genre; if manuscript: genre, author, name, location, page, etc.)
- use of the plant
- symbolical meaning

Then, it would be interesting to connect this database with international resources such as Europeana. You already can find some Russian names there, but not historical Russian plant names, so if we have such a database with material we have in manuscripts and old dictionaries included and enriched, it would be really interesting. It could help to have more modern and historical names represented in the international online resources.

Benefits of cooperation

- collaborative approach allows to establish sustainable workflows
- shared use of unified or de facto standards and infrastructures instead of starting creating new; develop new standards in a collaborative approach
- experimenting with new methods and emerging technologies
- gives a chance to open up new data
    - dictionaries
    - manuscripts (ease access to them)
    - lexical data
    - cultural data
- connection of data to the global resources (Europeana)
- more and better results for both cooperation partners
- sharing failures, risks and furthering learning and innovation
- improving competitiveness and visibility
- SHARING DIGITAL TRANSFORMATION AND INNOVATION

Contact

Kira Kovalenko, Institute for Linguistic Studies (Russia) & Austrian Centre for Digital Humanities (Austria) & **Eveline Wandl-Vogt,** Austrian Centre for Digital Humanities, (Austria)
Email: kira.kovalenko@gmail.com

Network of Dutch War Collections: pursuits and goals

Tessa Free, Netwerk Oorlogsbronnen, Netherlands

The Network of Dutch War Collections is a program of NIOD Institute for War, Holocaust and Genocide Studies, a research institute and a WWII-collection keeper in Amsterdam. The Network of Dutch War Collections is an independent program. It is facilitated by NIOD and a steering committee provides substantive direction. The goal on my organisation is to make scattered resources from and about the Netherlands in the Second World War digitally better findable and usable. In the Netherlands, there are approximately 400 institutes that keep documents about this period and there is a great diversity - from two papers to a couple of kilometers. Also, some documents are digitized and standardized metadated, and others are for example described on the personal computer of an almost retiring employee. We want to make 9 million sources findable and usable not directly for the public but firstly for the intermediaries: teachers, researchers, app or game builders, etcetera. So they can use the sources and reach the big public through their products (books, articles, classes, etcetera).

Scattered sources findable

An example: Westerbork memorial site was transit camp in the east of Holland. It is an important place of memory but a lot of documentation about it is scattered. Maps of the camp, extracts of diaries or movies are kept by different institutions and it is difficult to find all these documents in one place. We have different sources telling stories about one place but all kept by several institutes with different rules, different kind of publication, different possibilities to research it. We want to help all those different institutes by making the documents digitally better, findable and usable. For example, we can trace the story of a woman told by the different sources in different places.

Useful connected resources

By adding context, we connect resources. This is the most important thing for researchers (or other war source seekers) as it is the way you can tell the story of one person through all these different sources. Researchers are mostly looking for "*who*" (persons), "*what*" (themes), "*when*" (date) and "*where*" (places). So we are focussing on these questions to make the sources available through these four aspects. Besides making the resources better findable, we aim to make them better usable. By clearing property rights if possible [because some documents are copyrighted]. Or inform about these rights, so a researcher knows where to ask for publishing-permission.

Projects

- What: Build of a WW2 thesaurus implemented in different softwares
- Who: Personsportal, crowdsourcing
- Where: geocoded WW2 resources
- When: describing moments

R&D

- Open War Sources (Wikipedia)
- Pilotproject full-automatic access of the Central Archive of Special Legal Procedures (using OCR and NER (Name Entity Recognition) techniques)

Cultural change

Besides the projects, we work on explaining the importance of sharing, connecting and open publishing (if possible) to collection keepers. The Network keeps a War Collections-Portal where we harvest all the available resources. Nothing is hosted on the War Collections-website, it reflects information we harvest from the other institutes. Our work is sometimes uneasy to demonstrate because it is firstly a backside work and not visible at the first stages.

Contact

Tessa Free, Netwerk Oorlogsbronnen
Institutional website: http://oorlogsbronnen.nl/
Twitter: @tessafree or @oorlogsbronnen
Email: tessa.free@oorlogsbronnen.nl

Open Access Meets Productivity, "Scholarship, see effect of being an efficient source"

Adele Valeria Messina, University of Calabria, Italy

Case study: The Method of Online Academic Reviews and the Alleged Delay of post-Holocaust Sociology

At the beginning of this study there was a lot of confusion both about the use of EBSCO database and the delay of post-Holocaust Sociology. It was a research halfway and between Hemerographia and meta-Sociology.
This method has been chosen to verify this alleged delay of post-Holocaust Sociology in sociological literature through important indexes: the speed of publication and the scientific impact of research on academic public.

Beyond the digital research

How concretely open access answered to my own research question dealing with this sociological delay?
EBSCO hosts databases and open access to full text allowed to measure three important indexes:

- Productivity of sociologists
- (Their) Visibility
- (Their) Degree of Appreciation

It was then possible to address questions with regards with:

- how many written works a scholar has produced
- in which periods
- how many times the name of the authors appears in articles and reviews on EBSCO.

I also calculated the degree of appreciation in sociological works, thanks to the number of citations that academic environment has reserved for them. This method was important and useful because it allowed to verify this delay. "*Unknown papers emerged, unpublished reports and this fact cleared up doubts related to the question of the alleged delay of sociology*". By perusing the online academic sociological reviews (year by year) it was and it is possible to glance at and examine who promoted which research project and in which scientific reviews.

Three authors are special examples:

- Everett C. Hughes (1897-1983). It appears that Hughes wrote about the Holocaust in 1948, in a period very close to the World War II. In particular, he speaks about "banality of evil", an important category or concept, that we know instead thanks to Hannah Arendt in the 1960s. And this fact emerged from articles, letters, personal writings found thanks to open access in full text.
- Talcott Parsons (1902-1979). He is a famous author in Sociology, but for his works about the destruction of the Jews is completely unnoticed.

- Paul Neurath (1911-2001). After the conflict he put into writing his personal camp experience, but at the end of the war no publishing house was able to publish his work.

They are good examples that highlight the importance of open data for SSH and how perusal of sociological reviews permitted to determine “who” promoted “which” research projects and in “which” scientific reviews. It was also possible to comprehend the rhythms and delays within the academic environment for political reasons and research funding. The academic reviews are important for the sociological field because it is based on and built by the scientific reviews and dissemination of works. The online academic reviews do not replace or supplant paper or printed journals but support them permitting to have a wider range of analysis:

- Without this scientific and scholarly literature available online, thanks to open access, I could not have verified this delay.
- And without the perusal of well-qualified online reviews, several sociological studies would have been forgotten.

The idea behind this work is to contribute to the digitization of *unknown documents and manuscripts* related to modern and contemporary history and critical thought and to address the necessity of their usage and employment into current research. One of the central goal is to host unnoticed texts in a semantic open access platform in order to support the circulation and connection of data.

Links between open access and Humanities

It might be in a text, defined as tool of democracy in the sense that it is an expression of ideas with a freedom of speech. It is possible to speak of the new so-called “digital democracy” with digital tools.

That is why it is important to share data beyond digital tools, to link knowledge, to highlight relationship, to cross the bridge, made of words and concepts, that exists between any written and its readers. Right, on the relevance of these bridges, made of paper, Alessandra Cambatzu has written a great essay.

Thus, any text speaks to reader. Any reader talks back to the text. And this meeting, written or digitized, is powerful.

Contact

Adele Valeria Messina, University of Calabria
Email: adelevaleria@gmail.com & avmessina.freeebrei@gmail.com

Case studies on digital content reuse in the context of Europeana cloud

Eliza Papaki, Digital Curation Unit, ATHENA R.C., Greece

The work to be presented here was developed in the context of the Europeana Cloud project between 2013 and 2016. The Digital Curation Unit, ATHENA R.C. was leading the work on “Assessing Researcher Needs in the Cloud and Ensuring Community Engagement”.

Europeana cloud

Europeana Cloud was a 3-year project that started in February 2013 and ended in April 2016. Overview:

- Total Project Cost – 4.75m Euros
- EU Funding Contributing 3.8m Euros (80%)
- Matched Funding 950k (20k)
- co-funded by the CIP-ICT Policy Support Programme
- CIP-ICT-PSP-2012-6 - Project number 325091

Amongst its aims was to:

- build a cloud based infrastructure which would add new data to Europeana
- give solutions to content providers and aggregators to store, share and provide access to digital material in the area of cultural heritage more efficiently
- give researchers new services, tools with which they could access this work and share the content stored in the cloud - aiming for further strengthening its public impact.

Our work in Assessing Researcher Needs in the Cloud and Ensuring Community Engagement focused on:

- Developing an effective research content strategy for Europeana vis a vis Humanities and Social Sciences research communities and improving the understanding of digital tools, research processes and content throughout the research lifecycle.
- Engaging the Humanities and Social Sciences research communities in the use of Europeana as a valuable resource for Digital Humanities research.
- Managing Europeana Research, which aims at opening up cultural heritage content for use in research, by fostering collaborations between Europeana and the cultural heritage and research sector.

Methodology

In order to reach these aims we employed several methodological steps:

- Identification of research communities under the umbrella of SSH

- Web Survey to document the practices of researchers in Europe (tools, content and methods)
- Expert Forums which focused on the target audience of the project (documenting problems, gathering suggestions)
- Use cases on digital innovative tools (interviews were held with people employing tools, noting gaps and suggestions for further improvement)
- Focus Groups (discipline specific)
- Case Studies on particular research topics

Case studies:

Population Movement as a Result of Conflict in the 20th Century

This case study, which was conducted by Vicky Garnett of Trinity College Dublin, focused on conflicts of the 20th century:

- Greek/Turkish Conflict of 1920s
- Hungarian Revolution of 1956
- Yugoslav Wars of the 1990s

It imposed the following research questions:

- What was the international response to displaced population resulting from each conflict?
- How does this compare with the response to displaced populations in the 21st century?

Useful resources were found in newspapers, transcriptions and photographs but among the problems faced was that nothing could be found externally for the Greek and Turkish conflict and nothing online about the Yugoslav Wars.

Problematic resources within Europeana:

- Newspapers (at the time) were not properly searchable
- Metadata was often incomplete or inaccurate

Problems in accessing content beyond Europeana:

- Specifically for the Yugoslav Wars, the content is still sensitive and subject to
    - government embargoes
    - cultural sensitivity
    - lack of resources to maintain records

Considerations for Europeana

- This was a worthwhile study and there is much content beyond Europeana that can be potentially absorbed as part of a collection.
- Europeana needs to create stronger links to existing content and maintain that.
- Cultural sensitivities may be a problem in obtaining data, particularly for the more recent conflict of the 1990s

- Financial issues may also be a problem for maintaining records and content within local GLAMs.

Case study: Children's literature

This is a case study conducted by Eliza Papaki of Digital Curation Unit, ATHENA R.C. Definition of the topic: "*Children's Literature is (among many other things) a body of texts (in the widest senses of that word), an academic discipline, an educational and social tool, an international business and a cultural phenomenon*". Hunt, P. (2004). International Companion Encyclopedia of Children's Literature. Routledge.

Methodological steps:

- Matching Children's Literature to academic disciplines: History, Cultural History, Literature and Languages, Education, Library Studies, Philology, Textual Studies, Linguistics and Media Studies.
- Searching Europeana for related content resulted in: Text (1596) / Image (194) / Video (25) records ranging chronologically from 1450 to 2014 and geographically from all over Europe, mainly the United Kingdom.

Diverse content retrieved in Europeana can be considered as brainstorm of results.

Suggestions for promoting digital content reuse within Europeana

1. Enrichment of content in the topic of Children's Literature, even through potential collaborations with other digital libraries:
   - Gap in geographic coverage of available digital material
   - Collections as means of organising content
   - Easy access to references, databases, journals and books
   - Map of publications
2. Mobilising already existing associations of this area in respect to new digital resources, tools and services will increase the number of potential users and the community.
3. Further development of digital tools and services on such content gathers great attention among researchers who still employ more traditional, non-digital approaches in their research.

Findings from Europeana Cloud

- Different research disciplines use different types of data in different ways
- Data aggregation horizontal rather than vertical: Not useful for high-level, advanced research, but very useful for teaching
- Need for collection-level descriptions
- Need for user-friendly tools and services which will enable re-use of Europeana data

What is still needed

- Better understanding of how content and metadata is actually used, and its relationship with digital methods and tools

- Targeted outreach and engagement methods
- Empirical understanding of technical tools that will increase use of content and metadata.

Europeana Research

“*Europeana Research will help open up cultural heritage content for use in cutting-edge research. It will run a series of activities to enhance and increase the use of Europeana data for research, and develop the content, capacity and impact of Europeana, by fostering collaborations between Europeana and the cultural heritage and research sector. It will provide an important focus for the emerging communities of practice who rely on Europeana for their research, and support the European investment in digital cultural heritage*”.

Europeana Research is currently an ongoing project continuing and further enriching work conducted within Europeana Cloud in improving Europeana for research use and by enhancing the network of the research community.

Contact

Eliza Papaki, Digital Curation Unit, ATHENA R.C.
Institutional website: http://research.europeana.eu
Twitter: @eurresearch & @DigCurationUnit
Email: e.papaki@dcu.gr