Monday 24th October

9:00 a.m. - 12:00 a.m. | Introduction

Welcoming remarks: Mirjam Friedová (Dean, Faculty of Arts, Charles University), Marek Skovajsa (vice-dean for research, Faculty of Humanities, Charles University), Pierre Mounier (EHESS, OpenEdition), Emiliano Degl’Innocenti (DARIAH-IT) & Lucie Doležalová (Charles University)

The status of data in publication, Joachim Schöpfel (Lille University)

The presentation will investigate the relationship between data and text in different document types, in social sciences and humanities. It will introduce different categories and types of research data, make the link with research fields and methods, and add some comments about differences between SS&H and STM. It will also question the future of the distinction between documents and data, in the environment of content mining. Other issues will be addressed for further discussion: sharing and reusage of data; impact and evaluation of data; the link between document, data life cycle and the research process; identification, curation and preservation; last not least the function of data in the new context of open science. What is data? What is NOT data? What is functional and dysfunctional in the field of data management and data publishing? May be that at the end there are more open questions than answers. Yet, this is this just the beginning, and we'll have a whole week together to explore the field and improve our understanding.

Czech Literary Bibliography, Vojtěch Malínek, (Czech Academy of Sciences)

The aim of this paper is to give a short presentation about Czech literary bibliography research infrastructure, its activities in the last years and its plans for the future. The stress will be put on RETROBI software, developed as a result of project of digitisation of card catalogue of so called Retrospective bibliography of Czech literature 1770-1945. RETROBI software enables fulltext and semistructured searching in OCR representations of original catalogue cards and offers features for online editing and indexing of the data. Afterwards the possibilities of using CLB data for statistical and quantitative research in the field will be presented.
Turning the Polish Literary Bibliography into a Research Tool: Challenges, Standards, Interoperability, Maciej Maryl (Institute of Literary Research of the Polish Academy of Sciences)

I will discuss the research project aiming to transform a vast database of Polish Literary Bibliography (PBL) into a fully operational, digital research infrastructure for the study of Polish literature and culture of 20th century. The project entails retrodigitisation and transformation of the existing records into a coherent database as well as the development of data analysis tools for literary researchers. PBL is a specialized bibliography containing records about various types of materials concerning literature and literary scholarship (e.g. literary works, books, journals, magazines, articles, documents, dramas, movies, TV programs, conferences, awards, etc.), which are annotated in the unique semantic framework. In that respect it is similar to other national projects such as ABELL (Annual Bibliography of English Language and Literature). The online database contains records for 1988-2002 with printed volumes covering the period 1939-1987.

In my presentation I would like to focus on the following issues:

- Challenges: the methodological problems of dealing with data collected during a long stretch of 60 years, including the conversion of OCR'd scans into a database.
- Standards: choosing the right ontology for the data and mapping our resources onto it.
- Interoperability: plans to link the resources with LOD cloud and other bibliographies (hopefully with the Bibliography of Czech Literature too).

2:00 p.m. - 5:00 p.m. | Open critical Edition. The missing link between digital humanities and open access

Text Encoding Initiative, Marjorie Burghart (CNRS) & Emmanuelle Morlock (CNRS)

Transparency, interoperability, free and open access are values commonly shared by Digital Humanities projects. But the mere publication and display of content on the web is not enough to make a project part of Open Science. As a new way of doing science by allowing the users to process the underlying data of a publication with tools, instead of just perusing it, Open Science requires Open Data and Open Process on top of Open Access. In this session, we will explore how digital editions, historically a most important part of DH, can bridge the gap between Open Access and Open
Prague Winter School

Open Data Citation for Social Sciences and Humanities

Science. Using examples of digital editions based on the Text Encoding Initiative (TEI), we will see how to integrate the reflection on Open Data in an edition project as early as the conception phase, in close relationship with the theory about a text that is, in the end, a critical edition – a theory based on data.

5:30 p.m. - 6:30 p.m. | Public Presentation

OpenEdition: towards a European infrastructure for open access publication in humanities and social sciences, Pierre Mounier (EHESS, OpenEdition)

OpenEdition gathers four platforms for open access publications in humanities and social sciences: journals, books, scientific programs and academic blogs. Based in France, OpenEdition initiated specific programs in several European countries to be able to offer an international and multilingual infrastructure, currently disseminating online and open access more than half million academic documents coming from more than twenty countries in 14 languages. OpenEdition aims now at developing a distributed European wide infrastructure with 19 partners. Named OPERAS, this new initiative will foster cooperation at European scale and help humanities and social sciences join the common effort for the development of Open Science.
Tuesday 25th October

9:00 a.m. - 12:00 a.m. | Data Management Plan

*Research data management planning: a chance for Open Science. Methods and tutorials to create a Data Management Plan,* Marie Puren & Charles Riondet (INRIA)

With the growth of the Open Science movement in the past few years, researchers have been increasingly encouraged by their home institutions, their funders, and by the public, to share the data they produce. A new model of data sharing is emerging, and this issue is becoming more and more crucial for the scientific community and for national and international research policy. As shown by the OECD in 2007[1], public granting agencies hope that publicly funded research projects would give access to the data produced within their work, in order to provide new resources for economic development. And with the extension of the Open Research Data Pilot in Horizon 2020, H2020 beneficiaries have to make their research data “findable, accessible, interoperable and reusable (FAIR)[2]”, and are therefore asked to provide a Data Management Plan (or DMP) to this end.

More than a constraint, this new model of openness brings direct benefits for researchers. Sharing their data allows the researchers to organise and retrieve them effectively, to ensure their security, to collaborate with fellow researchers within the same discipline or from other disciplines, to reduce costs by avoiding duplication of data collection, to make easier validation of results, to increase the impact and visibility of their research outputs.

In this session, participants will get an overview of research data management principles and learn how to comply with these guidelines. We will define the purpose and the characteristics of a Data Management Plan (or DMP) and propose a method to write a DMP, with examples taken from the academic world. We will also mention the possible hindrances that researchers may encounter. Participants will finally be invited to write a Data Management Plan with a web-based tool, DMPonline. Those who are currently involved in a relevant research project are encouraged to work on their own research data.
The proliferation of datasets and services available online invites researchers to link them from their works. A link to a service, that makes it possible to explore the data in question yourself, might be more valuable than a picture. But these types of online resources and/or the infrastructures they live in are constantly evolving. Which effectively leads to dead links or links to a different version of the resource. PID systems can help with keeping the locations of resources up to date as well as store information about what the resource is.

Canonical Text Services, Matthew Munson (Leipzig University) & Christopher Blackwell (Furman University)

Canonical Text Services (CTS) is a protocol for identification and retrieval of passages of text by means of machine-actionable citations in URN form. CTS is not bibliographic database or commentary framework, but a protocol intended to serve use-cases like those. CTS consists of a specification for URN citations, and a specification for a service-protocol. CTS was created for the Homer Multitext to address that project's need to integrate (a) an open-ended diversity of texts, (b) many specific versions of the same text, some digital, some in print, and some in manuscript, many fragmentary, (c) at arbitrary levels of abstraction (“Iliad Book 2”) or specificity (“The third letter sigma at Iliad 1.2 on the Venetus A manuscript”), (d) with the assumption that technologies for storage, retrieval, and display will change completely during the project's lifetime.

CTS and its abstract data-model
CTS is based on a model of “text” as “an ordered hierarchy of citation-objects” (OHCO2), and an assumption that a canonically citable text exists in a bibliographic hierarchy of TextGroup >> Work >> Version >> Exemplar. Because it is based on an abstract model, CTS can be implemented in a variety of technologies. Current implementations use, as backends, relational databases, XML databases, and RDF databases. CTS texts can be implemented from, and delivered as, TEI-XML, JSON, Markdown, or plain texts in tabular (comma- or tab-delimited) form. A CTS URN can identify a passage of text at any level of specificity, from the notional (e.g. “Iliad Book 2”, referring to Book 2 in any version of the Iliad), to the highly specific, pointing to individual characters in specific versions of a text, whether it is a digital text or not.
This presentation will introduce CTS as a possible model for persistent identifiers in a large-scale, distributed digital library. The first part of the presentation will offer an overview of the protocol, the CTS URN citation scheme, and the CTS Service requests, with attention to applications and limitations of CTS. The second part will present how the Open Greek and Latin project of the University of Leipzig is implementing CTS and the tools it is making available for editors, publishers, and consumers of CTS texts.

Creation of Open Data Resources: Benefits of Cooperation, Kira Kovalenko (Russian Academy of Sciences & Austrian Academy of Sciences) & Eveline Wandl-Vogt (Austrian Academy of Sciences)

In the presentation we a going to tell about cooperation between Austrian Centre for Digital Humanities (Austrian Academy of Sciences) and Institute for Linguistic Studies (Russian Academy of Sciences). As a result of the collaboration, thee projects are going to be done: digital version of the Dictionary of Russian Dialects, electronic collection of Russian manuscript lexicons and a database of the Russian plant names (11-17 cc.). All the projects will be implemented with the use of cutting-edge technologies and will be available online.

Wednesday 26th of October

9:00 a.m. - 12:00 a.m. | Evaluation, acknowledgement and credit circulation
Simplifying license selection, Ondrej Kosarko (UFAL)

The necessity to share and preserve data and software is becoming more and more important. Without the data and the software, research cannot be reproduced and tested by the scientific community. Making data and software simply reusable and legally unequivocal requires choosing a license for data and software which is not a trivial task.
Open peer review & Open commentary: about an experiment, Julien Bordier (OpenEdition)

For five months, the OA journal Vertigo experimented both open peer review and open commentary devices within its scientific blog. While the first consisted strictly in opening a classical pre-publication review process, the second was inviting the whole “scientific community” to comment pre-publications in order to improve them before submission. In both of the two devices, every reviews, comments and annotations are accessible to everyone online, as the authors, reviewers and commentators names. The sociologist in charge of this project will present the details of the experiment, its main results – need of human mediation, technical possibilities and limitations – and will try to raise the questions and potentialities issued by new forms of reviewing in academic publishing.

1:00 p.m. - 4:00 p.m. | Case Studies

European Network for Research Evaluation in the Social Sciences and Humanities, Ioana Galleron (ENRESSH)

Evaluation has always been perceived as a difficult area for the SSH for a number of reasons. One of the problems is the fact that the most common procedures have been fine-tuned to the so-called hard sciences and as such are ill adapted the SSH disciplines. While abundant information exists about research practices, disciplinary biases and dissemination traditions in STEM fields, the situation is at least contrasted between Nordic and Southern countries with regards to the monitoring of the research production/outputs in SSH disciplines. This presentation will briefly introduce the COST Action CA15137, dedicated to the creation of a network of evaluators for the SSH disciplines, then will focus on the needs and the challenges of data collection for an informed peer evaluation of the SSH.

Network of Dutch war sources: pursuits and goals, Tessa Free (Network of Dutch War Sources)

In Holland, there are around four hundred organizations keeping a collection from or about the Netherlands in the Second World War. The program ‘Network of Dutch war sources’ (Netwerk Oorlogsbronnen) intends to make the geographically scattered sources digital findable and usable. We do that by engaging in or leading small projects with several participating organizations. For example creating a second world war – thesaurus, implemented in collection management systems. The use of OCR
and NER-techniques to make millions of documents accessible on a document level. And adding persistent geographical codes to sources to enhance searching with a focus on location.
The Network of Dutch War sources is a program of ‘NIOD institute for war holocaust and genocide’. See www.oorlogsbronnen.nl for more information about the program.

Open access meets productivity: “Scholarship, see effect of being an efficient source”, Adele Valeria Messina (University of Calabria)

“How do we use an EBSCO database?” and “How, without difficulty, can an article be found?” will be some of the concerns of this contribution.
The primary talking point of it is about the efficiency of Open Access in social sciences and humanities. The talk will therefore introduce the issue of a case study: “the method of online academic reviews and the alleged delay of post-Holocaust Sociology”. The presentation will confer about this method, halfway between hemerographia and metasociology, and about the measurement of some important indexes, such as “the speed of publication” of research and “the scientific impact” of it on the academic public. It argues that open access and usability of data need to be understood as more than simply a kind of digital research.

How a set of visualisation practices answered my own research questions and how free sources advanced the same research will be addressed.
In order to promote a thoughtful discussion on importance of open data to the social sciences, the talk will, specifically, examine how EBSCOhost databases and open access to full-text, allowed to measure the productivity (how many written works the scholar has produced), the visibility (how many times the name of the authors appear in articles and reviews on EBSCO), and also the degree of appreciation of post-Holocaust sociological works (calculated based on the number of citations that the academic environment has reserved for them).

This sheds a great deal of light on the question of how open access actually enforced the scholarly research.
But there will be other considerations as well: (1) access to older international publications; (2) quantitative diffusion of publications; (3) evaluation of these measurements unearthing previously ignored or marginal subjects of investigation.
From perusing online academic reviews unknown papers emerged, unpublished reports and this fact cleared up doubts related to the question of the alleged delay of sociology.
Thanks to online research methods this presentation will demonstrate the means adopted to write history and to do history through the reviews. Finally, the ideas behind this work will contribute to the digitization of unknown documents and manuscripts related to modern and contemporary history and critical thought. The talk will focus on the necessity of their reusage and employment into current research. It will also discuss about one of its central goal: to host unnoticed texts in open access through an open access thematic platform. Obviously, the project will support the circulation and connecting of data and will be linked with well versed institutions. This can happen best when there are energetic institutional means for the researchers: when they, to all intents and purposes, try to claim what they want is when digital democracy becomes challenging, as it is now, in Europe.

Case Studies on digital content reuse in the context of Europeana Cloud, Eliza Papaki (Athena Research Centre)

The use of digital content has, over the past couple of decades, become almost the norm for many researchers within the Humanities and Social Sciences. Curation of both digitised legacy data and born-digital content, however, makes it imperative that items are managed at an individual level in order for larger collections of data to be trusted and useful. Europeana is shifting focus from being a discovery portal of over 30 million digitised items to a platform that allows third parties to develop tools based on its content. In order to gather information about the potential use of existing collections in Europeana, research was conducted into developing an empirically-based, comprehensive list of User Requirements. Investigations included current data reuse within the sector, the quality of the content itself and identification of topics with which Europeana can be of most use.

In our investigations through the Europeana Cloud project, we took the approach from both users and providers. Topics were selected for trial using Europeana’s current content, and other potential resources, both of which were subjected to questioning: how useful was the data to them?; what tools or services could be used with it; what were the failings of the content and how might that be overcome? In this
presentation, two of these topics have been selected as case studies: Conflict-related Population Displacement; and Children’s Literature.
Thursday 27th of October

9:00 a.m. - 12:00 a.m. | Data Journals & editorialization of open data

*DBpedia Demo: Basic exploration of a RDF graph with simple SPARQL queries, Emmanuelle Morlock (CNRS)*

*Do we still need peer-review? Datajournals as a way of reconsidering our evaluation culture and our understanding of research, Anne Baillot (INRIA)*

Never had scholars had to write so many applications and so many reviews as nowadays. Peer-review has been institutionalized as the central regulation mechanism of the two core activities that are formulating a research question and its workflow on the one hand and criticizing its results on the other hand. Still, most scholars are deeply unsatisfied by a system in which they feel like they never really get to “do” research, but are rather stuck in a vicious circle of unproductive evaluations.

While evaluation is perceived by scholars as more and more disconnected from research itself, the datajournal model developed by DARIAH in the context of the episcience platform is aiming at re-harmonizing research and evaluation, allowing to integrate peer reviews as a contribution to the research and development process of an online resource, in a continuous (virtuous) feedback loop.

In this session, I will begin with a short introduction on the role of journals in the research workflow, on the different forms of peer review and their impact, and on the general principles of datajournals. In the main part of the session, the participants will be given the opportunity to work with the episcience sandbox dedicated to the datajournal model. They will be asked to outline a datajournal prototype in a field of research they can define themselves (this can be realised individually or in small groups and will take about 15mn of creative brainstorming). Each group will then present their idea in 2-3 minutes in which they should tackle in particular the workflow specificities connected to the data/discipline they have chosen as an example.

The session will allow to address both benefits and challenges of datajournals, aiming more widely at initiating a constructive dialogue on publication and evaluation structures in the digital age.

This session is followed by a workshop about setting up a data journal, coordinated by Anne Baillot (INRIA) & Marie Puren (INRIA)
The development of open access in humanities and social sciences faces a major challenge: sustainability. Whereas in STM disciplines, the new dissemination paradigm means shifting from reader-pays model to author-pays model, the infamous "APC", in humanities and social sciences that type of reconfiguration is simply not possible for many good reasons. Moreover, the dissemination of knowledge in those disciplines is mostly done through books and not solely in journals, which entails additional complications. Therefore, those who want to develop open access in SSH have to find their own solution, that fits with their specific ecosystem. Whether it be based on donations, grants, subscriptions, in kind contributions, crowdfunding or "freemium" model, there are many ongoing experimentations under development. A landscape of the different models and the main trends on the topic will be presented.

Sdvig press is a non-profit academic publishing platform dedicated to supporting the dissemination and linking of knowledge in the Humanities between Eastern, Central and Western Europe. One of its central mission is to give visibility and provide structured access to large corpora of texts from Russia, Poland, the Czech Republic or the Baltic states, relating in particular to important epistemological paradigms of the Humanities such as structuralism, phenomenology, or critical theory. This objective implies not only the high-quality digitisation of textual sources, but also their translation at least into English. Given the still obscure nature of these corpora, however, it is hard to find anything more than punctual funding, and there is no prospect of even mild commercial success to finance these tasks in the systematic, long-term perspective that they require.

The solution explored by sdvig press is the development of thematic platforms that integrate Central and Eastern European corpora into better defined, more visible and more international contexts. We are developing three prototypes, of which the Open Commons of Phenomenology is the most advanced (the other two are Structuralica and Pacem). Each of these platforms, at first, is conceived mainly as an exhaustive
bio-bibliographical repository providing structured access (ideally) to all sources and references in its thematic field. Access to contents is wholly unrestricted, but a number of tools (advance search, lists, visualisations, etc.) are made available only to libraries through a subscription.

The short-term logic of this model is that even a very extensive bibliographical repository is limited in scope (e.g. Phenomenology will include about 250k references, including chapters and articles) and can be produced relatively quickly (1-2 years). Subscriptions from topic-specific libraries or institutes (at least 400 in the case of Phenomenology) can then provide a six-figure yearly budget for the further digitisation, curation and translation of the corpus. Furthermore, by being closely embedded with its research community from the start, the platform also benefits from digital labor from relevant institutions, in the form of blogs, bibliographical inputs, etc.

In the long-term, the constant development of the platform and its increasing role as a crucial infrastructural hub for its research community means that its functionalities can extend beyond the role of a repository, serving as a publication service, a social network and a technical lab for hosting and implementing DH projects. The repository thus becomes a service provided to (and by) researchers – and paid by libraries.

Friday 28th of October

9:00 a.m. - 12:00 a.m. | Infrastructure & platform

Contrasting platforms and infrastructures as configurations for data sharing, Jean Christophe Plantin (London School of Economics and Political Science)

This talk will discuss the impact on scholarship when data sharing is increasingly organized by social media platforms. It does so by contrasting these entities with existing data infrastructures that acquire, curate, and process these data for archiving and further dissemination. The analysis of routine, procedures, and everyday work of data processing staff at a social science data archive will provide elements to detail the “regime of care” that define infrastructure towards research data, and how it contrasts with digital intermediaries in organizing data circulation.
Prague Winter School

Open Data Citation for Social Sciences and Humanities

*Humá-Num: a French infrastructure for open research data in humanities, Nicolas Larrousse (Humá-Num)*

In the field of Humanities and Social Sciences, the production of digital or scanned data has increased considerably in recent years. These data, which are usually very expensive to produce, are often lost at the end of the project. They are therefore rarely reused, due to a lack of financial, human and technical resources of the communities that produced them. This talk will present the general approach, both technical and educational, used by Humá-Num infrastructure to address these issues.

1:30 p.m. - 4:30 p.m. | Social impact

*Infrastructure for Global Philology, Gregory Crane (University of Leipzig)*

This presentation discusses core services and use cases for an infrastructure that seeks to support work on any historical language by speakers of as many modern languages as possible.

Further information: [https://datacite.hypotheses.org](https://datacite.hypotheses.org)
Contact: [nathanael.cretin@openedition.org](mailto:nathanael.cretin@openedition.org)