Proposal for Session for Interaction and Engagement

Title:

Data as Impact Lab – Hackathon on metascience

Organizer(s):

Gustaf Nelhans, Johan Eklund, and David Gunnarsson-Lorentzen, Swedish School of Library and Information Science, University of Borås

Key Participants: See above

Abstract:

Recent medial discussion on "fake-news" underlines the importance of evidence-based decision-making. To gather, analyze and interpret "facts" is, however, in our information-dense digital times, not always easy. Activities such as information seeking, knowledge building and evaluation in scholarly practice are often performed using bibliometric/informetric methods. The increased interest in bibliometrics also opens for new questions on how data sources are being used and what kind of challenges and/or possibilities that warrant further investigation.

In this session for interaction and engagement, we invite participants to explore both means of analyzing already available data sources using machine-learning technology, as well as to include new sets of data that could augment the different views of grasping research activities using algorithms. Such data could be both content-intensive (as text), time-sensitive (as events), contextual (in terms of links between different properties) or multi-modal, meaning that other sources, such as imagery, sound, and video – even material objects may constitute possible contributions as data as impact.

Description:

Recent medial discussion on "fake-news" underlines the importance of evidence-based decision-making. To gather, analyze and interpret "facts" is, however, in our information-dense digital times, not always easy. Activities such as information seeking, knowledge building and evaluation in scholarly practice are often performed using bibliometric/informetric methods. We also see a gained importance for the use of bibliometrics as a "quality" measure that is used for evaluation and performance-based funding. The increased interest in bibliometrics also opens for new questions on how data sources are being used and what kind of challenges and/or possibilities that warrant further investigation. For example, by visualizing networks and relations bibliometrics can be used as a way for grasping large communities, on a meta-level, as well as make it possible to follow discourses around a certain discipline and/or the development of common words that are used in a certain scientific field. In light of the growing interest and development of computer-assisted tools traditional bibliometrics recently has been amended with semantic analysis of text, AI and prospective studies of potential knowledge

contribution based on data models of the content of linked documents, rather than only measuring number of interactions or citations. Sources have also been amended with data from so-called alternative metrics, derived from social media, policy and professional documentation that is constantly made available.

In this session for interaction and engagement, we invite participants to explore both means of analyzing already available data sources using machine-learning technology, as well as to include new sets of data that could augment the different views of grasping research activities using algorithms. Such data could be both content-intensive (as text), time-sensitive (as events), contextual (in terms of links between different properties) or multi-modal, meaning that other sources, such as imagery, sound, and video – even material objects may constitute possible contributions as data as impact.

Given the above question and the need to problematize bibliometrics as a methodological tool as well as a means to gain more knowledge useful for development of research policy, a newly started research-infrastructure has been initiated at SSLIS, *Data as impact Lab*. This lab is a collaboration with external actors that intend to investigate the properties and traits in specific data and/or information as well a venue for researchers dealing with vast amounts of data, corpus or digital material.

Building on the notion borrowed from Hjørland (1992) that the subject of a document is defined by its knowledge potential, we argue that the manifest subject (s) of research impact can be identified in the actual use(s) of the actual research in specific instances. Such 'fluid' classification additionally offers the possibility to study topic drift or changes in usage over time or across usage settings.

Purpose and Intended Audience:

The purpose of this session for engagement and interaction is to explore and investigate available information resources further by using computer-assisted methods to analyze and understand impact of research in specific societal domains investigated. The intended audience includes a wide range of scholars within the field of library and information-science with backgrounds within for example; research policy studies, informetrics, information practices and/or media studies. The overarching goal with the activity is that the accessible information in its data format might be technical per se, but that it is important to interpret and understand the mechanisms behind the presented result in a certain tool, interface and/or visualization.

Proposed activities including agenda, ramp-up (development), and follow-through:

Instead of submitting an abstract for a talk, participants are invited to submit a data set, including metadata descriptions to actually use the content *or* if such data does not *yet* exist, submit a specification of requirements for such a data set with the potential to actually find or develop such data during the session.

This two-part session for interaction and engagement aims at identifying means of using data in a meaningful way to not only indicate, but (to a varying degree) recognize the actual impact of the indicator measured. Ultimately, these links could be single impact stories, identified by manually following the data, but, hopefully, would be attained using data analytic methods, artificial intelligence

We will proceed through two work packages, where the first focuses on the collection and aligning of different data sources for identifying research activities and events as well as creating a machine learning experiment where as many as possible of the submitted data sets will be included. Techniques that will be used will depend on the actual data sources brought together, but might include Topic modelling, Word embeddings and Deep learning among other machine learning approaches. In the second work package, the results of the experiments will be investigated through an evaluation of possible "indicators" that can be derived from the experiment exercise. As a final step, this open-ended approach will yield different outcomes that will be evaluated, discussed and interpreted by the participants.

Goals or Outcomes:

The goals of the two work packages are to investigate new means of employing machine learning techniques in research evaluation by incorporating data from different sources. Combining text-based data with linked data, instead of only quantifying simple metrics is expected to yield interesting indicators for a broader analysis of research activities from the policy domain. The evaluation of the outcomes will also investigate potential bias due to lack of data, privacy issues and the pros and cons of exposing more and more parts of the research activity to scrutiny.

Relevance to the iConference:

As information scientists, we are constantly involved in the negotiation of meaningful sources of data to investigate phenomena under study. But data has no meaning in themselves, but are only purposive when they are used in a specific situation and in relation to other data. We expect that the participants of this interactive session will identify new means of understanding 'data as impact'.

Duration:

2 * 90 minutes

Attendance:

We expect that between 10-20 participants will actively attend both sessions.

Special requirements:

No.