Editor's notes:
Sharing open data without risk, and with machine-actionable provenance metadata

Welcome to the fourth issue of 2020 and the last issue of volume 44 of the *IASSIST Quarterly* (IQ 44(4) 2020).

At a future time, there might be a special issue of *IASSIST Quarterly* on 'Corona data'. Right now, the numbers are rising as we are entering winter, but on the other hand vaccination is around the corner. I hope that only 2020 will be remembered as the year of the Coronavirus, and that 2021 will bring us better times.

Open data and the sharing of data, including public data, is a foundation of democracy – by making data available for all and not only for the current political leaders. We are talking about free data like we talk about free speech. Data archives around the world are contributing to making this possible, and this issue reveals the efforts made at the Czech Social Science Data Archive towards involvement of the users in data sharing. We have to obtain a balance, and I shall return here to the phrase 'as open as possible, as closed as necessary'. Open data must not compromise individuals and their right to privacy, and the need for protection of sensitive data. In this issue you will also find guidelines for assessing risks embedded in the data as well as remedies for de-identification and anonymization. Such instructions for safe sharing and depositing of datasets are central for data producers, researchers, and students. We must tolerate speech we disagree with – and then we might decide to take the time and effort to refute the statements. Likewise, we might experience data that call for scrutiny. For that purpose, the highest level of metadata is needed. I recall that many years ago, many hours were spent at the Danish Data Archives trying to figure out how a central variable in an important survey was constructed. The Structured Data Transformation Language will solve such issues, by providing access to provenance metadata harvested from transformation scripts of common statistical languages. Again, another accomplished contribution to the improvements in the sharing of quality data. Great thanks are due to all the contributors at data archives, library schools, research institutions, and many more.

Michaela Kudrnáčová and Ilona Trtíková are working at the Czech Social Science Data Archive. Michaela is a PhD student with a focus on research and methodology, and Ilona is data manager with expertise in retrieving and sharing research information. Their work at the data archive includes a great deal of communication with students and researchers on issues of data management and data analysis. The archive views its role in open science as creating a trusted and sustainable environment for social science data. In order to better respond to users' demands and needs, data were collected on their users from several sources including user registrations, a survey, and interviews. The article on 'Sustainability through the liaison with data archive users' thus includes some statistics on the users of the Czech Social Science Data Archive (CSDA) since its availability for online data storing and sharing through Nesstar. The presentation of the survey results includes distributions of purpose and frequencies of use of the CSDA. A very positive conclusion was the high degree to which users express their willingness to cooperate in the development of the CSDA functions and services. The investigations also revealed specific areas and functions that could be

 Rasmussen, Karsten Boye (2020) Editor's notes: Sharing open data without risk, and with machine-actionable provenance metadata, *IASSIST Quarterly* 44(4), pp. 1-2. DOI https://doi.org/10.29173/iq990

improved, and they are planning now to perform a short online survey of CSDA users every two years, supplemented with interviews.

While data sharing is encouraged, at the same time sharing data from surveys often implies a risk for identification of individuals. That situation is addressed by the submission titled 'Mathematics, risk, and messy survey data'. The authors provide researchers and data collectors with a better understanding of the theory and concepts of anonymization and risk assessment. The obvious first step in data anonymization is the removal of specific individual information, e.g., name, telephone numbers, and social media identifiers. However, demographic variables (quasi-identifiers) such as occupation and geography might also be sufficient to identify individuals, for example a small town's only doctor. Many possible such quasi-identifiers might exist in a dataset. However, deleting quasi-identifiers may seriously decrease the value of the dataset. One widely used technique for ensuring anonymity is k-anonymity, which assesses how many records in the dataset have the same combination of quasi-identifiers. The authors illustrate the utility and challenges of k-anonymity by walking readers through the anonymization of two datasets. The authors Kristi Thompson and Carolyn Sullivan are at Western University, Canada, where Kristi Thompson is the Research Data Management Librarian and Carolyn Sullivan is a student of Information and Media Studies.

The article 'Provenance metadata for statistical data: An introduction to Structured Data Transformation Language (SDTL)' presents a truly collective effort. The authors are George Alter, Darrell Donakowski, Jack Gager, Pascal Heus, Carson Hunter, Sanda Ionescu, Jeremy Iverson, H V Jagadish, Carl Lagoze, Jared Lyle, Alexander Mueller, Sigbjorn Revheim, Matthew A. Richardson, Ornulf Risnes, Karunakara Seelam, Dan Smith, Tom Smith, Jie Song, Yashas Jaydeep Vaidya, and Ole Voldsater representing the institutions University of Michigan, Metadata Technologies North America, Algenta Technologies, Norwegian Centre for Research Data, and NORC. The Continuous Capture of Metadata for Statistical Data Project (C2Metadata) created SDTL to capture provenance metadata from data transformation scripts found in statistical analysis software. The objective is to provide standardized, machine-actionable documentation to answer questions like 'Which original variables were used to construct this derived variable?' SDTL works with metadata standards, like DDI, to add file- and variable-level provenance to data catalogs and codebooks. Software created by the C2Metadata Project translates commands of five leading statistical packages (SPSS, Stata, SAS, R and Python) into SDTL. SDTL covers basic data transformation commands for assigning and recoding values, metadata commands for setting labels and data types, and file-level processes like merging and appending rows. The article describes the SDTL approach and discusses similarities and differences among statistical analysis software.

Enjoy reading the three articles.

Submissions of papers for the *IASSIST Quarterly* are always very welcome. We welcome input from IASSIST conferences or other conferences and workshops, from local presentations, or papers especially written for the *IQ*. When you are preparing such a presentation, give a thought to turning your one-time presentation into a lasting contribution. Doing that after the event also gives you the opportunity of improving your work after feedback. We encourage you to login or create an author profile at https://www.iassistquarterly.com (our Open Journal System application). We permit authors to have 'deep links' into the *IQ* as well as deposition of the paper in your local repository.

Chairing a conference session or workshop with the purpose of aggregating and integrating papers for a special issue *IQ* is also much appreciated as the information reaches many more people than the limited number of session participants and will be readily available on the *IASSIST Quarterly* website at https://www.iassistquarterly.com. Authors are very welcome to take a look at the instructions and layout:

https://www.iassistquarterly.com/index.php/iassist/about/submissions.

Authors can also contact me directly via e-mail: kbr@sam.sdu.dk. Should you be interested in compiling a special issue for the *IQ* as guest editor(s) I will also be delighted to hear from you.

Karsten Boye Rasmussen - December 2020