Licensing policies for collaborative data production and reuse: Open licensing and digital traces

Melanie Dulong de Rosnay

Institute of Communication Sciences
(CNRS - Paris Sorbonne – UPMC)

@melanieddr

Workshop on Collaborative Data Projects
Academia Sinica, Taipei, 8 December 2016

Access and reuse of data

- Open licensing
  - ARR + transfer of rights
  - CC -> PD
  - CC0 + PDM
  - Also for database
- Permission embedded in the registration process and/or the terms of use/contribution

Copyright embedded with personal data

- https://www.zooniverse.org/privacy
- Non exclusive grant

  “if you contribute to the Zooniverse, you grant the CSA and its collaborators, permission to use your contributions however we like to further this goal, trusting us to do the right thing with your data. However, you give us this permission non-exclusively, meaning that you yourself still own your contribution.

  We ask you to grant us these broad permissions, because they allow us to change the legal details by which we keep the data available; this is important because the legal environment can change and we need to be able to respond without obtaining permission from every single contributor”.

Privacy policy

“we collect additional data about you to support and improve the operation of the project. We also conduct experiments on the design of the website that we evaluate based on your reactions and behavior.

contributions you make to the Talk pages are widely available to others. Aside from the above, information is held as confidentially as is practical within our secured database.”

Profiling from open contributions

- Transportation or sleep patterns
- Political interests
- Big data
- The algorithmic processing of digital traces deduced from our online activities

  and those of our social networks

  helps platforms to make correlations and inferences

Digital traces

World Bank report (Schwab et al., 2011)

"personal data is defined as data (and metadata) created by and about people," distinguishing:

- Volunteered data created and explicitly shared by individuals, e.g., comments, contributions to social media.
- Observed data shared involuntarily captured by recording the actions of individuals, e.g., location data when using cell phones, browsing history, metadata.
- Inferred data about individuals based on analysis of volunteered or observed information, (e.g. deduction of age and social class on the basis of first name, education or solvability rating used for credit scores decision).

Beyond data protection

Conclusions drawn by algorithms based on inferred data

Can lead to arbitrary or exclusion decisions

are beyond the scope of data protection law, since they originate not from personally identifiable information, which is eligible for protection, but from digital traces resulting from algorithmic processing.

And “because predictive algorithms obfuscate the act of inference, (relying on privacy is) futile” (Mc Quillan, 2015), as it is not possible to know “when a certain form of information processing will produce predictive privacy harms” (Crawford & Schultz, 2014).

Crossing telephone metadata (Mayer & al., 2016), which receive less legal protection than content, with publicly available information from social networks allows to draw sensitive conclusions regarding, for example, medical conditions, cannabis usage or weapon ownership.

Even our “imagination” can leave traces, according to Latour (2007)

Informationelle Selbstbestimmung

Informational self-determination: a mode of personal data protection

1983 by the German Federal Constitutional Court

grounded on constitutional right to dignity

and right to free development of one's personality

provides individuals the right to behave freely in a society where we leave more and more digital traces

if individuals ignore which information about themselves is known, and by which parties, their freedom to initiate projects or decide without pressure is limited

not incorporated in European or international law

a positive contribution based on the nature of persons rather than on one's data ownership

Consequences

Examples of digital traces created or left by others

social media networks use individuals' web histories to personally target advertising.

IP addresses and geolocation history may be used to the same effect

MZ "secured a patent that, among other things, allows lenders to assess creditworthiness based on the credit ratings of people in a borrower's network." Even when regulations may not allow the use of social media to determine creditworthiness, in some cases "applicants must agree to have their mobile calling and texting patterns used in an analysis (WSJ).

Open access vs personal information

digital traces may not individually contain personal information and escape personal data regulations

the aggregation of even anonymized data can lead to re-identification (Sweeney & al., 2013; Dawson, 2014)

go behind the deceptive dichotomy between

- advocating for open access to information which does not contain personal information and

- controlling access to data which could reveal private information and lead to unfair decisions

Participatory platforms

Signatures on online petitions are a source of valuable information for indirect profiling.

“In parts of Africa a number of corporations, not coincidentally some with colonial legacies such as Unilever and Cadbury, transform participation by the poor into a new form of market research” (Arora, 2016), and investment decisions may be based on information gleaned from digital traces.

Researchers reusing information publicly available on Wikipedia (an API allows us to download the histories of contributions) will be able to shed light on the interests and political opinions of contributors.

In the field of citizen science, observations on natural and ecological phenomena will reveal time and location patterns of contributors.

Solutions?

- Privacy-by-design

Collaborative archiving

- Web90
- Memory hole
- Status of libraries & archive
- Legal deposit
- Wikipedia & similar projects