QROWD - Because Big Data Integration is Humanly Possible

Innovation action

D3.4 – Crowdsourcing Vocabulary and Licensing

Author/s	Luis-Daniel Ibáñez
Due date	DD.MM.YYYY
Version	0.1
Dissemination level	PU
Status	Final

Project co-funded under Horizon 2020 Research and Innovation EU programme, grant agreement no. 732194

Table of contents

ABSTRACT

A currently overlooked aspect in the practice of crowdsourcing is how to make crowdsourcing tasks repeatable and reproducible. This is important both for research purposes, and for organisations that run crowdsourcing tasks periodically or that share tasks between different departments or with partner organisations. In this deliverable we describe an initial version of the Crowd-Voc vocabulary, aimed at becoming a standard for describing crowdsourcing tasks in a machine-readable and Web-compliant way. Crowd-Voc lives in a GitHub repository, where we defined a series of conventions, inline with practice of W3C working groups, to drive its community-driven development. We expect to have major releases at M30 and M36.

EXECUTIVE SUMMARY

This deliverable has as primary audience any organisation designing and executing crowdsourcing tasks, in particular, for those playing the roles of designers and taskmaster, and for researchers in crowdsourcing.

We deliver an initial version of a vocabulary for describing crowdsourcing tasks, and an accompanying repository that will support its continuous development for the rest of the project and beyond. This version has been opened to research and practitioners communities to collect feedback and improvements, and to kickstart the adoption process.

The vocabulary will provide the means to describe the specification of crowdsourcing tasks, and optionally, the link between crowdworkers/contributors, and their specific contributions, paving the ground for a fair licensing model for crowdsourced data, with special attention to microtasks.

1. CROWD-VOC VOCABULARY

For the first version of the vocabulary, we chose to fix three stakeholders: designers and taskmasters of the described task; taskmasters that would like to reproduce or adapt a previously executed task; and crowdworkers.

Following the pattern established by several W3C working groups devoted to vocabulary development we defined initial use cases following the convention Name-Stakeholders-Description-Existing approaches to kickstart the discussion in the community. We expect further ones to be contributed by the community.

Use Case 1: Description of crowdsourcing task

Stakeholders: Task initiator and task replicator

Description: In order to replicate crowdsourcing tasks, Crowd-Voc needs to provide the means to describe them. We understand a task as the sequence of actions that a single human contributor is asked to perform to produce a judgement or annotation.
A single contributor can be asked to do several different actions, for example, if shown an image, a first action can be "Identify if there is cat in the image", and a second action "Indicate the color of the cat". The vocabulary needs to be flexible enough to accommodate descriptions of complex tasks for humans.

Existing Approaches (using other vocabularies): PROV can be extended to encode a workflow-like pattern. PROV was not required to specify the notion of a subactivity[1], W3C suggested instead the use of the *hasPart* relation from the dcterms vocabulary[2], letting applications be responsible for ensuring its consistent use with respect to PROV. PROV recommendation of using hasPart, one can model sub-tasks as other PROV:Activities that compose a task. This can be used in combination of other several ontologies that have been designed for scientific workflows like P-Plan[3] or OPMW[4].

Use case 2: Crowdsourcing experiment parameter description

Stakeholders: Task masters (crowdsourcing task initiators), and crowdsourcing task replicators

Description: A task replicator wants to replicate a Crowdsourcing task or experiment that generates annotations on an input dataset. (S)He needs to know (i) the number of workers/collaborators in the task (ii) their characteristics (if required for the task, e.g., gender, age, socio-economic background), (iii) the aggregation and

[1] Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, Simon Miles, *The rationale of PROV*, Journal of Web Semantics, Volume 35, Part 4, 2015, P235-257,
[2] https://www.w3.org/2001/sw/wiki/PROV-FAQ#How_can_I_define_a_sub_activity.3F
[3] http://www.opmw.org/model/p-plan/
[4] http://www.opmw.org/

quality metrics used to aggregate and finalise the results.

Existing approaches (using other vocabularies): A task can be modelled as a PROV:Activity that uses an input DCAT:Dataset and generates an output DCAT:Dataset with the annotations made by the crowd. Workers/collaborators in a task can be modelled as PROV:Agents.

Use case 3: Transparent tracking of contribution attribution

Stakeholders: Task masters (crowdsourcing task initiators), crowdworkers and contributors

Description: During a task execution, link the entities or cells of the resulting dataset or annotations to the contributors/workers that generated them.

This should be optional, depending on any pre-existing agreement between task initiator and crowdworkers, and the desired level of privacy of the crowdworker.

The crowdworker should be able to query what contributions (s)he did for a task.

Existing approaches (using other vocabularies): The agent-centered provenance section of PROV could be reused for most of this use case.

2. CONTINUOUS IMPROVEMENT PROCESS

Crowd-Voc is published in a GitHub repository[5], from which we leverage several affordances to implement a continuous improvement process for Crowd-Voc, inline with common practice on currently active ontology groups like W3C DataExchange Working Group[6] and the opencitydata.es collection[7].

Any person can open a GitHub issue with any of the following tags:

Use Case: For the proposal and discussion of use cases
A good use case description should be structured as follows:

- Use case name
- Stakeholders
- Description
- Existing approaches (using other vocabularies)
- Links to other issues (if any)
- Eventually, links to requirements

Requirement: For proposal and discussion of requirements
A requirement is a concise description of a feature of the vocabulary. Additionally

[5] https://github.com/QROWD/crowd-voc
[6] https://github.com/w3c/dxwg
[7] https://github.com/opencitydata

there should be a link to the use cases that require it.

Alignment: For the proposal and discussion of alignments with other vocabularies and ontologies, including property re-use

Classes / Properties For the proposal and discussion of class and/or property definitions within the ontology. Commonly, these issues should be linked to a requirement or use case.

GitHub issues support discussions and markup, serving as a forum for contributors to discuss about them. The QROWD consortium takes the final decision on accepting a use case or requirement to be implemented in Crowd-Voc.

3. TIMELINE

We plan to have two major releases of Crowd-Voc, one at M30 and another at month 36. We also plan to introduce the discussion in a workshop on one of the main conferences on crowdsourcing and Human Computation in 2019 (HCOMP or CHI), around M30.

4. LICENSING SCHEMES

The dataset that results from a crowdsourcing task must be licensed if it is to be used by third parties. A license describes what a party may or may not do with a dataset. Even in situations where a dataset is not explicitly assigned a license, there may be licensing restrictions on it that stem from restrictions on the input datasets, agreements made with other parties involved in the production, or even the terms offered to the crowd workers.

The datasets produced by the Qrowd project are no different; they will need to be licensed to the parties that are making use of them. Here, we discuss the objectives that we feel responsible licensing should aim to address, and then evaluate some potential licensing models.

We suggest that responsible licensing should seek to comply with licensing restrictions that are imposed by source datasets, be compatible with agreements between requester and workers, and take account of restrictions that relate to particular characteristics of the dataset (in particular, data protection), provide a fair balance between the rights and obligations of all interested parties, and promote appropriate and legitimate exploitation.

1. Maintain and comply with licensing restrictions that are imposed by source

datasets. (Source Obligations)

Source datasets may be encumbered by a range of legal restrictions, including licenses or contractual arrangements between the providing and receiving parties (for instance non-disclosure or non-compete agreements). If these legal restrictions persist into the resulting dataset, they must be accounted for in the way that dataset is licensed; both for the protection of the organisation licensing it and the licensees who might otherwise be unaware of potential liabilities of using the data; for instance those that might arise if the resulting dataset constituted a derivative work under copyright law. Novel sources of data, for instance "Data Donation"[8] may impose novel obligations - for instance a duty to maintain long-term availability for specific purposes such as medical research.

2. Be compatible with the implicit and explicit agreements between requester and workers. (Worker Obligations)

Although relationships between workers and requester in online crowd work scenarios are usually short-lived and impose no ongoing obligations on either party, that is not *necessarily* the case. In particular, where workers are also the citizens of a municipality and motivated in whole or in part by potential social benefits, this imposes at least an ethical duty on the requester to make efforts to ensure that social benefit is realised; in some circumstances, it's possible that it may also impose a legal obligation. In any case, the license(s) applied to the resulting data should be compatible with the agreements between workers and requester which could be created implicitly through (for example) the way that participation is explained during recruitment.

3. Comply with and enforce regulations that relate to particular characteristics of the dataset. (Regulations)

Some datasets, for instance those that contain personal data, are subject to additional legal restrictions that need to be taken into account when those datasets are licensed and processed. In particular, datasets containing personal data will – in most cases – be subject to purpose limitation, and transfers may need to be subject to a data processing agreement.

4. Provide a fair balance between the rights and obligations of all interested parties. (Fairness)

Datasets have increasingly complex provenance, and those involving crowd work are no exception. In fact, as alluded to above, the rights of the workers themselves can introduce additional interests and complexity. We suggest that licenses should seek to balance the rights and obligations of the different parties fairly, as well as doing so legally. Debates around the fairness of crowd work itself (e.g. Ettlinger 2016 [9]) as well as related developments such as the sharing economy, demonstrate that it

[8] https://www.oii.ox.ac.uk/research/projects/a-european-ethical-code-for-data-donation/

[9] Ettlinger, Nancy. 2016. "The Governance of Crowdsourcing: Rationalities of the New Exploitation." Environment and Planning A 48 (11): 2162–80. doi:10.1177/0308518X16656182.

does not necessarily follow that because practices (in particular labour practices) are legal they are also fair. Crowd work may be exploitative. Although licensing cannot in itself resolve all of these concerns, it can help to address them. Novel forms of licensing, for instance vesting ownership of datasets in a data trust that can benefit the workers who helped to create it, could have a particularly important role to play with regard to fairness.

5. Promote appropriate and legitimate use of data. (Exploitation)
We alluded to the importance of making efforts to exploit data for social good in relation to agreements with workers, above, but society also has a broader stake in datasets – particularly those created at public expense – that goes beyond the expectations of the parties who were involved in their creation. Licenses should therefore seek to promote exploitation of datasets that could unlock social value, where such exploitation is appropriate and legitimate. This requirement implies that a legally conservative approach, which could satisfy the other four conditions, should be balanced against the need to encourage exploitation and value creation.

The huge scope for variation between projects means that it’s impossible to suggest one licensing model that would work for all, or even a significant proportion, of datasets produced in whole or in part through crowd work.

We have briefly evaluated the implications of some common licensing models with respect to our five principles, and explain them next. We offer these observations as a starting point for deliberations over how a specific dataset should be licensed, rather than as definitive descriptions of how each model necessarily operates.

1. Proprietary Ownership

In some senses the ‘default’ form of licensing for many datasets, by ‘proprietary ownership’ we mean the treatment of datasets as the private property of a single (or small group) of organisations, with no (or very limited) access by third parties. The overwhelming majority of customer databases and market research data (for instance) would fall into this category.

Source Obligations & Worker Obligations: This arrangement provides plenty of scope to meet obligations imposed by the dataset’s sources or workers involved in its creation – except, of course, if those obligations include making the data more generally available! It would be improper, for instance, for a dataset that had been collected from citizens of a municipality ostensibly to improve public transport provision, if it was not shared with relevant organisations in some form.

Regulations: Again, this model provides plenty of scope to meet regulations such as data protection. The dataset can be included within an organisation’s broader risk management activities.

Fairness: The fairness of keeping data in proprietary ownership will depend on the dataset in question. In the case of a business's customer database, it is unlikely to be fair to the business or the customers (nor legal) to publish the data more widely; but the realities and priorities of commercial organisations might also limit the efforts that can be made to ensure overall fairness.

Exploitation: Proprietary ownership of a dataset is a major barrier to exploitation, unless that organisation is willing negotiate access with other parties. In many cases, it will not even be clear which datasets are held by an organisations, severely limiting discoverability.

2. Open Data

A number of licenses suitable for Open Data are available, and the UK's Open Data Institute (ODI) recommends[10] the use of a Creative Commons public domain (CC0), attribution (CC-by), or attribution & share-alike (CC-by-sa) license for open data sets.

Source Obligations & Worker Obligations: An Open Data license provides no scope to meet specific obligations to the owners of source materials or to workers, except in the special case that the obligation is itself to make the resulting dataset Open data (which is not uncommon).

Regulations: An Open Data license does not enforce specific regulations, so is not suitable (for example) for datasets containing personal data. A more thorough treatment of Open data and personal data is provided by Simperl, O'Hara and Gomer (2016)[11].

Fairness: Open data licenses give no scope to enforce additional restrictions on data use, so publishing as open data does not provide an opportunity to address unfairness in the overall data pipeline. However, it certainly does not follow that publishing open data is inherently unfair. In many cases, for instance in the case of government data, it may be fairer to stakeholders (e.g. citizens or tax-payers) to publish the data openly than to restrict access through a more closed arrangement.

Exploitation: Open data licenses make data freely available for re-use, and so provide a high level of exploitability. For some datasets (notably those derived from personal or other restricted data) the high availability of publishing as open data might need to be balanced against the utility that is lost through necessary pre-publication processing such as anonymisation.

[10] https://theodi.org/article/publishers-guide-to-open-data-licensing/

[11] Simperl, Elena, O'Hara, Kieron and Gomer, Richard. 2016. "Open Data and Privacy." European Data Portal Analytical Reports. Available: https://www.europeandataportal.eu/sites/default/files/open_data_and_privacy_v1_final_clean.pdf

3. Data Trusts

The UK's Open Data Institute defines a Data Trust as "a legal structure that provides independent third-party stewardship of data"[12]. As in other areas of law, a data trust is established to act in the interests of a defined beneficiary, or group of beneficiaries. Although still rare and novel, data trusts have the potential to act as dynamic data custodians, and to manage relationships with interested parties – including licensing the dataset for specific uses. Unlike a commercial organisation, in most jurisdictions a trust is required to act in the interests of its defined beneficiaries.

Source Obligations & Worker Obligations: A data trust should act within law, as trustees are subject to the usual criminal and civil sanctions. It does not necessarily follow, if the data trust was not the original producer of the data, that all contractual obligations would carry across, although arrangements could be made to do so.

Regulations: A data trust, as a dynamic entity, can respond to changing regulations, and also make case-by-case decisions about when and how data is released to third parties, making it suitable for datasets that could not be made freely available as Open Data for privacy or other concerns.

Fairness: A data trust could be established to deliver a fairer outcome for parties that are involved in the production of a dataset. For instance, workers could receive royalties when data is licensed by commercial entities, or patients could benefit from earlier access to experimental treatments developed using data they had donated.

Exploitation: Data trusts are likely to place some restrictions on how data is used, and increase the overheads of doing so by introducing an administrative burden to obtaining access. However, the restrictions and overheads introduced may be outweighed by the richness of the datasets that can be provided. Because a data trust is established specifically to hold datasets, those datasets are likely to be more discoverable than those held in proprietary ownership. Although the trust must be adequately resourced to remain effective throughout the useful life span of the dataset itself.

Using Crowd-Voc to Support Licensing Decisions

Although to date we have not formally modelled legal or ethical aspects of crowdsourcing, Crowd-Voc provides a practical tool to assist organisations in identifying where activities within a crowdsourcing activity are relevant to the five

[12] https://theodi.org/article/defining-a-data-trust/

identified licensing objectives. Aspirationally, we would like to develop an extension to Crowd-Voc to account for the objectives more formally. Here, we briefly consider how the five objectives could be modelled on top of a Crowd-Voc description in an ad-hoc manner that could be undertaken today, and as the basis of an aspirational formalisation.

By breaking a crowdsourcing activity down into more granular tasks, Crowd-Voc provides a framework for describing the tasks, data sources and stakeholders involved. This granular account of the overall activity provides a skeleton upon which legal and ethical aspects can be superimposed, in essence acting as a structured provocation for an in-depth discussion among stakeholders and task owners.

We propose that restrictions and obligations arising from (for example) contractual arrangements or copyright restrictions can be identified at the point at which they are introduced into the data flow.

For instance, obligations to a source dataset's owner can be noted at the point that the dataset is incorporated into the flow. Obligations to crowd workers can be noted at the point that those workers are involved.

Similarly, one can note where those obligations are removed or discharged - technically or legally. For instance, restrictions imposed on datasets containing personal data can be removed by an (effective) anonymisation process. An obligation to citizens to deliver public benefit could be discharged by a task that makes the dataset available the local municipality.