WODA: a platform for geospatial open data augmentation using hybrid human-machine workflows

Gianfranco Cecconi[1], Elena Simperl[1], Leslie Carr[1] and Nigel Shadbolt[2]

[1] University of Southampton {gc1a13||e.simperl||lac}@soton.ac.uk
[2] University of Oxford nigel.shadbolt@jesus.ox.ac.uk

Abstract. Until a few years ago, the creation and maintenance of geospatial data inevitably meant deploying expert personnel to survey the locations, with substantial cost and effort. As more open data is published by governments and organisations, new opportunities arise to create data products and services from integrating pre-existing sources with ad hoc additions.

Human contribution is still central to the creation of original data, and automation cannot address all the challenges, such as replacing us in most surveying tasks. However, we have now the opportunity to create hybrid human-machine workflows in which re-using open data, computation and crowdsourcing enable and enhance each other's output for the creation of new data products and services.

In this paper we present WODA (Workflows for Open Data Augmentation): an experimental platform implementing this hybrid model. In the attempt to maximise cost-effectiveness, scalability and availability of the technology component, we design WODA to rely only on open source software and the native features of mainstream services including CrowdFlower as the crowdsourcing platform and Google Maps for mapping and aerial imagery.

WODA is then applied to a real world use case: the creation of a dataset of all valid UK addresses, starting from available open data, augmenting it through computational inference and complementing it by using paid microtask crowdsourcing.

Experimental evaluation shows the feasibility, effectiveness and challenges of the approach.

Keywords: Crowdsourcing, Geographic Information, hybrid human-machine systems, open data

1 Introduction

Geospatial data is a critical component of many applications. Its value is universally recognised as instrumental to economic development, yet little is readily available and subject to high costs or restrictive licensing.

The availability of geospatial data under open licensing for the public good is under threat. Its production and maintenance is expensive and National Mapping and Cadastre Agencies (NMCAs) worldwide - the traditional suppliers of such data - are cutting investment into producing and updating cartography and other assets [5]. In the U.S., for example, the Geological Survey (USGS) no longer updates maps on a regular basis.

A vision is emerging in which geospatial data is gathered from a pool of heterogeneous open sources including government and private sector, and consolidated into new data products and services as needed. In the US, the National Research Council promotes such a vision[3].

This new paradigm requires process and technological change. As a useful tool for geospatial researchers and practitioners involved in this transformation, in this paper we present "WODA" (Workflows for Open Data Augmentation): an open source platform aimed at implementing hybrid human-machine workflows that cost-effectively integrate pre-existing open data, computation and original contributions by human participants - typically through crowdsourcing.

We then evaluate WODA by dealing with a specific use case: the creation of the list of all valid addresses in the UK, or the "Open Legal Address File", which has received a lot of attention in the country as an example of simple, though critical, geographic data that is available only as a commercial product.

2 Background and related work

2.1 Crowdsourcing geospatial information

Geographical Information Systems (GIS) are systems designed to capture, manipulate, analyse and visualise geographical data.

Crowdsourcing geographical data was made possible in the early 2000's by the availability of GIS to the wider public. As web technology matured, publicly available, web-based GIS systems emerged including OpenStreetMap[4] (OSM), Google Maps[5] and Wikimapia[6]. In what was by some called "the democratisation of GIS" [2] laypeople could for the first time contribute to the systems, augmenting pre-existing data and creating new. This was around the same time the term "crowdsourcing" was used first.[7]

Crowdsourcing geospatial data became the subject of extensive research. Because of the origins of the phenomenon, volunteered rather than paid participation to the systems was studied in particular, to the point of defining the

[3] M. S. Committee and N. R. C. Mapping Science Committee, Toward a coordinated spatial data infrastructure for the nation. 1993.

[4] Created in 2004, see `https://www.openstreetmap.org`.

[5] Launched in 2005, see `https://www.google.com/maps`. Google Earth is actually precedent to Google Maps and was launched in 2001, but at the time it was available as a downloadable desktop application only.

[6] Created in 2006, see `http://wikimapia.org/`.

[7] J. Howe, "The Rise of Crowdsourcing" Wired, 01-Jun-2006.

discipline itself. In 2007 Michael F. Goodchild coined the term "Volunteered Geographic Information" (Volunteered GI, or VGI) [8] and examined it as a new form of citizen science. Goodchild also was among the first to make the hypothesis that relying on the crowd could be more cost-effective at maintaining geospatial data than any of the traditional practices.

2.2 Challenges of crowdsourcing geospatial data

There are a number of recognised challenges in crowdsourcing geospatial data, most of which are in common with crowdsourcing in general. The most relevant to our research are summarised here.

Contributor credibility The trustworthiness and expertise of a contributor defines her "credibility". In cases where crowdsourced GIS systems are designed to be accessible to the layperson, the only variable remains trustworthiness, which depends on the contributor's motivation, which in turn "suggests greater or less potential for bias and deception" [6]. Gatekeeping and quality control are practices used to assure contributor credibility.

In our research we attempt at reduce bias by using paid crowdsourcing.

Data reliability Even assuming that the contributors are credible, it is necessary to assess the quality of the system's overall output.

The reliability of mainstreams VGI systems was assessed for example by Haklay in [10], who performed a comparative study of data offered by OSM and the British NMCA Ordnance Survey (OS). In 2008, only a few years after it was started, OSM offered a "reasonable accuracy" of about 6 meters and an overlap of up to 100% of roads in OS' data.

Further research in [11] suggests that the volume of volunteers involved in a project - even when working without a central coordination - can be considered as an intrinsic quality assurance measure.

In our research, paid crowdsourcing gave us more control over the volume of contributors we could engage, and enabled us to use robust statistical tools.

Data completeness Volunteered GI has a completeness issue, in the words of OSM creator Steve Coast: "Nobody wants to do council estates".[8] Loose organisation of the contributors and a range of socioeconomic barriers [10] may hinder achievement of sufficient coverage of those geographical areas volunteers are not motivated to contribute about.

Moreover, some applications may require the data to be produced within a given timeframe, not compatible with volunteer best effort, something that could be better controlled using paid crowdsourcing.

Domain knowledge Contributors to GI crowdsourcing projects need a sufficient understanding of the concepts that define the domain, such as what a road is, what a building etc. [1].

In our work we tried minimising this requirement, by simplifying the contributor task down to making simple observations from imagery.

[8] "The GISPro Interview with OSM founder Steve Coast" GIS Professional, no. 18, 2007.

2.3 Open data and open GI

Because research has focussed on systems that leverage volunteer contribution and are designed for the pubic good, GI is often implicitly associated to open data: "data that anyone can access, use and share"[9].

Government-funded NMCAs are natural owners of geospatial data and are commonly expected to release it in the open for the public good. In Great Britain, for example, since 2015 OS has opened a substantial volume of data that was previously available to the public as commercial products only, including datasets such as "Open Names" - a place-name index - and "Open Roads" - the generalised geometry and network connectivity of the road network - which we used in our platform as primary data sources.

The availability of such high quality and authoritative sources becomes a substantial enabler for the creation of new data and augmenting the existing.

All useful and needed geospatial data is not always made available in the open, though. Demand for open data can be suppressed, for example, due to failure in recognising open data-enabled business models, restrictive legislative and patent systems, or charging for access.[10] The OLAF problem, described below, is an example of this.

2.4 The Open Legal Address File (OLAF)

The main use case for our research is creating an Open Legal Address File[11]: a dataset that lists all known valid addresses and postcodes for the UK and is functionally equivalent to the "Postcode Address File" (PAF[12]).

PAF is the ownership of Royal Mail, the UK ex state-owned postal service. Law makes PAF available on "reasonable terms" to "any person who wishes to use it".[13] This, however, never translated into making the data open. PAF is available as a commercial product only, and was sold by the government as part of Royal Mail's privatisation in October 2013. OLAF aims to fill this gap in the UK open geospatial data offering.

Without going into the detail of official British Standard BS7666,[14] a practical operational definition of a "valid UK address" would be *the set of unambiguous information needed to instruct any delivery service operator to deliver a piece*

[9] See `http://theodi.org/faq`.

[10] N. Shadbolt, "A Cornerstone for Open Data: The Postcode Address File" Apr-2013. Online. Available: `http://theodi.org/blog/cornerstone-open-data-postcode-address-file`. Accessed: 09-May-2015.

[11] See `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/blob/gh-pages/docs/README.md#legal-address-files` for an explanation of the "legal address" term.

[12] PAF is a registered trademark by Royal Mail plc. For convenience we won't show the registered trademark sign "®" in this document every time we refer to it.

[13] See the Postal Service Act 2000, part VII, article 116.

[14] See `http://shop.bsigroup.com/ProductDetail/?pid=000000000030127201`.

of mail at a given public address (a private house, a commercial establishment etc.), from a consumer perspective.[15]

3 The WODA platform

3.1 The workflow

The generic workflow in Figure 1 uses primary and authoritative data sources as an input, and iteratively augments them using computation. Wherever the conditions for computation are not verified, crowdsourcing is used to create the missing data to enable it.

This section outlines the design principles of WODA: a re-usable platform to support implementations of such workflow. The crowdsourcing component is then explored in more detail.

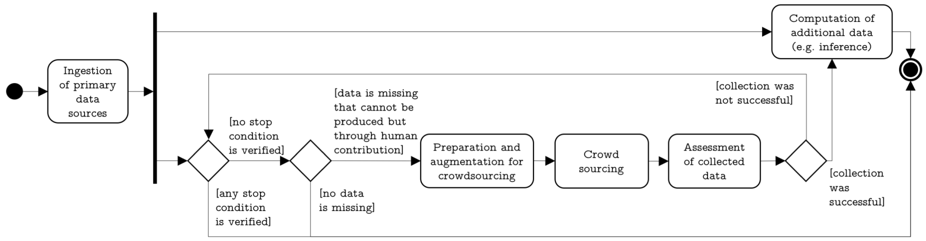

Computation of additional data (e.g. inference)
Ingestion of primary data sources
[data is missing that cannot be produced but through human contribution]
[collection was not successful]
[no stop condition is verified]
Preparation and augmentation for crowdsourcing
Crowd sourcing
Assessment of collected data
[collection was successful]
[any stop condition is verified]
[no data is missing]

Fig. 1. UML process diagram of the generic WODA-supported workflow

3.2 Overall design principles

WODA was designed following the principles below:

Human-machine workflows support WODA meant to automate as much as possible of the data creation / curation process, while effectively integrating human components, from administration to the crowdsourcing of part.

Data re-use As mentioned in chapter 2.3 we wanted to make the best possible use of pre-existing, high quality and authoritative open data sources.

Open source software It was assumed that open source software could be used to implement any required component of WODA. This also saves budget for the compensation of paid contributors.

Scalability and high availability WODA was designed to be highly scalable and available, suitable for real-world deployment.

[15] Find the reference address data model at `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/tree/gh-pages/docs#address-format`.

Versatility WODA had to be versatile and suitable to be applicable to a range of problems describable within the workflow above and leveraging different input data sources as needed. The crowdsourcing component is required to be flexible enough to cater for different models: paid contributors vs volunteers, approaches to results aggregation, quality assessment etc.

3.3 Crowdsourcing design principles

The crowdsourcing component, can be described against the dimensions presented by Simperl in [18].

What is outsourced The objective of the outsourcing activity is the production of original geospatial data or the curation (correction, validation etc.) of pre-existing data, wherever the activity can be only performed by human agents.

The activity translates into surveying the locations or examining imagery thereof and record observations (e.g. "how many trees can be seen from longitude x and latitude y?"), or amend previous recordings (e.g. "can you confirm that there is a hospital in Vicarage Rd, Watford?").

To avoid the cost of performing a physical survey of the locations, participants examine publicly available imagery. This is not uncommon, e.g. OSM currently uses aerial imagery from Microsoft Bing to let its contributors edit street topology.

Progress in machine learning, computer vision etc. hints at a future in which the human component may be made redundant. To this day, however, automation can only complement rather than replace humans (e.g. in [7] or [16]).

Who is the crowd In VGI, contributors are often associated to the locations they work on, and their work can be seen as the act of capturing the knowledge they own of a place. Occasionally, this can be instrumental to assure the completeness of the data, e.g. the function of public buildings can't be inferred from observing OSM's aerial imagery. Physical survey of the locations may be also required, using specialised equipment.

To contribute to WODA, specialised knowledge, skills and equipment are not required, nor is a connection to the places being surveyed.

How is the task outsourced In the current implementation, WODA focuses on outsourcing micro rather than macro tasks. Wherever possible, one's contribution is limited to a few minutes' work at the computer and does not require awareness of the larger system they are part of, in the attempt of removing complexity and barriers to participation.

Why do people contribute The motivation of VGI crowds is a complex subject, summarised e.g. in Coleman *et al.* in [3]: motivation spans from altruism to intellectual stimulation, social reward and "pride of place".

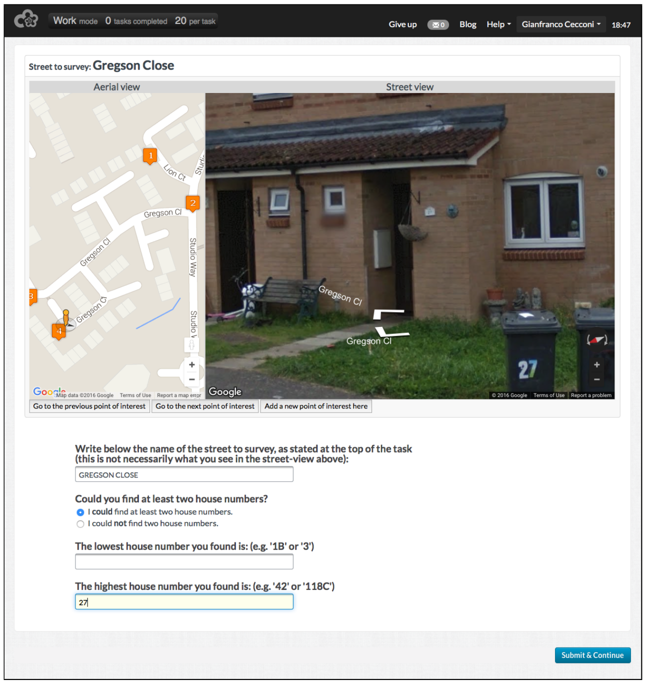
In the current implementation, WODA contributors are driven merely by financial interest. They are oblivious of the the context of the project, have no personal connection to the locations, and are likely unaware of and find no motivation in contributing to the "cause" of open data. This is typically the crowd that can be recruited through mainstream crowdwork platforms.

3.4 Implementation

Overview

From the **platform owner**'s point of view, WODA is made of several sets of tools to (i) ingest the primary data sources into a reference database, (ii) process the ingested data and derive new / enhanced data computationally, (iii) configure the crowdsourcing platform of choice, including the user interface participants use, (iv) support the execution of the crowdsourcing stage (load data to the crowdsourcing platform, collect its results etc.), and (v) consolidate primary, derived and crowdsourced data in one consistent dataset.

From the **participant**'s point of view, WODA presents itself simply as a task hosted on a crowdsourcing platform. The task web page offers interactive maps and panoramic views of one or more locations to survey, and a form contributors use to submit observations, as shown in Figure 2.



Street to survey: Gregson Close

Write below the name of the street to survey, as stated at the top of the task (this is not necessarily what you see in the street-view above):

GREGSON CLOSE

Could you find at least two house numbers?

● I **could** find at least two house numbers.
○ I could **not** find two house numbers.

The lowest house number you found is: (e.g. '1B' or '3')

The highest house number you found is: (e.g. '42' or '118C')

27

Fig. 2. The task section of the web page, offering the Google Maps and Street View panes and the form participants use to submit their observations.

On first access, instructions are provided at the top of the page , possibly alongside a video used to illustrate the more interactive features.[16]. The video

[16] See `https://raw.githubusercontent.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/gh-pages/docs/images/screenshot%20with%20instructions.png`.

specific to OLAF's deployment of WODA can be watched at `https://www.youtube.com/watch?v=pIiGJ6gMEY0`.

Components

CrowdFlower CrowdFlower was chosen as the crowdsourcing platform. It is a SaaS specialised in hosting data-centred microtasks for volunteer or paid participants. Clients who accept that their data could be re-published as open data may access the service under a plan that has no costs but for the compensation to the paid participants and a commission. The service is available both through a templating system called CML and through APIs.

When using CML, the options available to implement a crowdsourcing model - such as how the accuracy of the contributor is assessed or how results are aggregated - are limited by the functionality supported by the system.[17] In order not to introduce additional components in the system and maximise scalability and high availability, we relied on CML only and worked around its limitations by performing some of the data processing outside of CrowdFlower.

Google Maps Google Maps is a mapping SaaS, made available to the public for free. It offers satellite and aerial imagery, maps, interactive panoramic views of streets alongside the metadata through APIs. Many of its services can be embedded in third party websites and customised using client-side JavaScript, hence making it suitable to integration with CrowdFlower or other template-based systems.

YouTube YouTube is a video publishing platform by Google. Similarly to Google Maps, videos can be embedded in third party websites, hence is suitable for integration in CML. WODA uses YouTube to provide instructions to participants.

GitHub GitHub is a web-based Git repository hosting SaaS. Together with WODA's source and documentation, it is used also for hosting static assets, e.g. the icons used to mark points of interest in Google Maps.

PostgreSQL + PostGIS The PostgreSQL relational database running the PostGIS spatial extender was chosen as the main data repository. All geospatial data is converted from its source format to PostGIS' native geospatial types to be consistent across all sources and enabling geographic querying.

Scripting Bash, NodeJS and PostgreSQL scripting were used to glue all components together wherever automation is possible.

[17] For example, CrowdFlower calculates Workers' agreement as the quorum of participants who expressed the majority vote, and tasks can be stopped automatically when a target % is achieved. The same cannot be done when using alternative statistical measures of agreement, whose calculation must take place outside the system.

3.5 Legal implications

The design choices have legal implications to be taken into account when deploying beyond research.[18] It is outside of the scope of this paper to examine this in any detail, however some of the matters are described below.

Personal data and privacy implications. According to EU Data Protection Directive (95/46/EC), implemented in the UK by the Data Protection Act of 1998, geospatial data such as addresses can be considered as personal data even when it is not associated to information about who lives at the locations, as it is "information relating to an identified or identifiable natural person". The directive describes a framework of practices to comply with.

Imagery terms and conditions Google Maps' terms of service specify restrictions on producing "derivative works of the Content or any part thereof" and on creating "a database of places or other local listings information". It is advisable that - before deploying WODA for real-world applications - the mapping services provider is informed and licensing clarified. This is what OSM did when integrating Bing imagery.

4 Applying WODA to the Open Legal Address File problem

4.1 How primary open data sources shape the solution

The availability of high quality and authoritative open data is central to the definition of the solution. An assessment of available sources highlighted OS' "Open Names"[19], (OSON) as the key primary input.

OSON lists place names, roads numbers and postcodes in Great Britain, but not (i) which house names and numbers are in which road, and (ii) which house names and numbers are associated with which postcode. Therefore, OLAF can be seen as an augmentation of OSON, obtained by adding those missing components. Production of the missing data can be obtained through running the three complementary processes *p1*, *p2*[20] and *p3* described in figure 3.

Records of existing house numbers in other primary sources such as Land Registry's "Price Paid Data"[21] (LRPP) enables to further refinement of *p1* in four sub-processes *p1.1*, *p1.2*, *p1.3*[22] and *p1.4*. LRPP records every property own-

[18] See `https://peepbeep.wordpress.com/2015/12/18/what`.

[19] See `https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-names.html`.

[20] It is estimated 98% of addresses are characterised by a house number, so *p1* is more relevant to OLAF's completeness than *p2*.

[21] Land Registry is a non-ministerial UK Government department with the responsibility to register the ownership of land and property in England and Wales. See `https://www.gov.uk/government/collections/price-paid-data`.

[22] See `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/blob/gh-pages/docs/README.md#error-in-inferred-house-numbers` for an assessment of the impact of error.

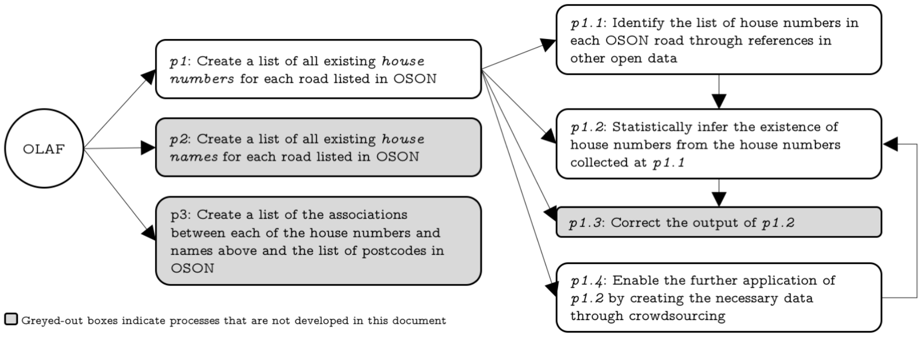

OLAF
p1: Create a list of all existing house numbers for each road listed in OSON
p2: Create a list of all existing house names for each road listed in OSON
p3: Create a list of the associations between each of the house numbers and names above and the list of postcodes in OSON
p1.1: Identify the list of house numbers in each OSON road through references in other open data
p1.2: Statistically infer the existence of house numbers from the house numbers collected at p1.1
p1.3: Correct the output of p1.2
p1.4: Enable the further application of p1.2 by creating the necessary data through crowdsourcing
Greyed-out boxes indicate processes that are not developed in this document

Fig. 3. A possible decomposition of OLAF

ership transfer in England and Wales since 1995, including their full addresses, from which the inference of addresses is possible.

The generic workflow described in section 3.1 can be used to integrate *p1.1*, *p1.2* and *p1.4*. Figure 4 shows the new workflow and the mapping against the three processes. The following describes in more detail its components.

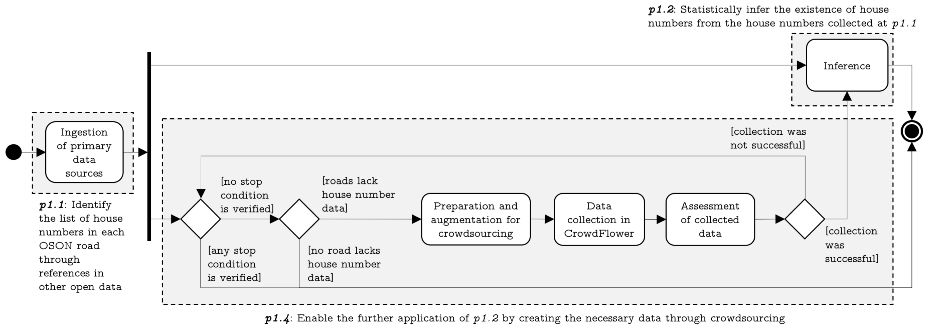

p1.2: Statistically infer the existence of house numbers from the house numbers collected at p1.1
Inference
Ingestion of primary data sources
p1.1: Identify the list of house numbers in each OSON road through references in other open data
[collection was not successful]
[no stop condition is verified]
[roads lack house number data]
Preparation and augmentation for crowdsourcing
Data collection in CrowdFlower
Assessment of collected data
[any stop condition is verified]
[no road lacks house number data]
[collection was successful]
p1.4: Enable the further application of p1.2 by creating the necessary data through crowdsourcing

Fig. 4. UML process diagram of the OLAF-specific workflow to create house numbers

4.2 Ingestion of primary data sources

Both OSON and LRPP contain references to streets, by their name. The main challenge of harvesting LRPP is to identify unambiguously the streets the house numbers are associated to, across the two datasets. Issues such as differences in spelling (e.g. “Downing Street” instead of “Downing St”) and association of the locations not to the same town, but to localities thereof (e.g. “Clapham” instead of “London”), need being managed with common data processing practices.

4.3 House number inference

House numbering convention Inferring house numbers is strictly dependent on the local numbering convention for the assignment of house numbers and names to buildings. In the UK, buildings are typically numbered sequentially starting from 1, at the extremity of the road closest to the town centre. Odd numbers are on the left-hand side, as seen from the town centre, while even number are on the right-hand side. House numbers can be suffixed by one or more letters: this is typical of larger buildings that at some point in time got divided into smaller dwellings.

Inference algorithms Once the numbering convention is known, inference can be used to create a large volume of missing house numbers from the observation of known house numbers. Algorithms 1 and 2 below have a very high probability of correctly inferring the existence of house numbers[23].

Algorithm 1: Inference of house numbers

```
Data: The list of known house numbers in a road
Result: The list of inferred house numbers in the same road
if the list includes at least one even and one odd number then
    infer all numbers between the lowest and the highest known numbers;
else
    if the list includes at least two even or two odd numbers then
        infer all even/odd numbers between the lowest and the highest
        numbers;
    end
end
```

Algorithm 2: Inference of house number with suffixes

```
Data: The list of known house numbers with suffixes in a road
Result: The list of inferred house numbers with suffixes in the same road
for each house number appearing in the list with at least two suffixes do
    infer all suffixes between the lowest and the highest known suffix, in
    alphabetical order;
end
```

Enabling inference for roads with no data Inference of house numbers is enabled by either of the following two conditions: (a) the knowledge of two or more different house numbers in the same street and (b) the knowledge of two or more suffixes for the same number. The former has the highest potential to generate new house numbers.

When dealing with roads for which LRPP provides no data, it is useful to focus crowdsourcing on creating the conditions that enable that potential. To

[23] See `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/blob/gh-pages/docs/README.md#exceptions-in-the-uk-house-numbering-system` for exceptions in the convention

infer the largest sets of numbers it is necessary to use as input to algorithm 1 the road's lowest and highest known house numbers[24].

4.4 Stop conditions

The stop conditions referred to in the diagram are every condition not strictly related to the function of the system, e.g. the availability of budget. In other words, the workflow is iterated as long as budget is available, even if the aimed outcome is not achieved.

For the experiments, the conditions were as either (*s.1*) consensus is reached on all roads, or (*s.2*) the number of repeat judgements by reliable Workers is higher than the number of first judgements, or (*s.3*) a total 300 USD budget is spent, whichever condition is verified first.

4.5 Preparation and augmentation for crowdsourcing

The data required by the crowdsourcing component may need preparation and augmentation before being used. In the case of OLAF, an additional primary data source is used to support the Workers in their tasks: OS' "Open Roads"[25] (OSOR).

OSOR is used to calculate the geographical coordinates of the extremities of the roads where the lowest and the highest house numbers are more likely to be found. These are offered as "points of interest", to support participants in their search. It is an example of how open data can be used not necessarily as a direct input into producing the output dataset, but to support the human participants in their function, too.

4.6 Data collection in CrowdFlower and assessment

The crowdsourcing component of WODA can be configured to create the house numbers needed to enable inference. Observing the imagery of a street to identify its features is not conceptually different than crowdsourced labelling and annotation applications that are extensively studied in literature.[26]

The following is a description of the task model and the approach in genera that were used for crowdsourcing house numbers, that is common to all experimental conditions that were tested.

Requester. The Requester desires to gather the lowest and the highest house numbers that can be observed in a specified street, as they can be intelligibly

[24] For simplicity, we did not consider (i) that it is also useful to know if the streets have both odd and even house numbers and (ii) the case where one house number only is known for a street, for which we could assume 1 to be the lowest house number.

[25] See `https://www.ordnancesurvey.co.uk/business-and-government/products/os-open-roads.html`.

[26] More detail is available at `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/tree/gh-pages/docs#finding-house-numbers-as-a-labelling-exercise`.

identified by browsing imagery, or the lack thereof. The Requester requires the help of human agents to carry out the tasks, that we will call Workers in the following.

Task. Each HIT (Human Intelligence Task) consists of browsing the pictures of a street until achieving reasonable certainty of having identified the lowest and the highest house numbers or the lack thereof.

Strategy. The strategy relies on traditional crowdsourcing techniques for image labelling.

Crowd → Worker. Each Worker provides judgement on a task by browsing the pictures and declaring if she has found the lowest and the highest house numbers or none. Multiple Workers are asked to identify the house numbers for the same street. The resulting data is chosen through majority voting.

Quality. Quality is defined by a combination of (a) credibility of the Workers in responding to tests questions, and (b) consensus in the data submitted through repeated surveys of the same road.

Probing Workers using conventional test questions - e.g. where the correspondence of the Worker submissions is checked vs the same data collected by the research team as described in [13] - would be a powerful tool to identify high vs low quality contributors, but is very expensive in OLAF's case. Spending a substantial part of the Worker's effort on test questions - e.g. making one out of three surveys a test - was not affordable.

As an alternative, though, simple test questions can be set up on data that is already available, in a way that is similar to classic anti-spamming techniques like CAPTCHAs as described in [4]. In OLAF's case the name of the street itself is used: Workers are asked to copy and paste or type the name of the street as part of their survey. Workers that do not achieve the target accuracy are excluded from further work.

Results aggregation. Repeated surveys of the same road are equivalent to the use of repeated judgement in conventional image labelling exercises. This practice is described extensively in literature and demonstrate that the results produced by a few expensive expert individuals are comparable to what emerges from involving multiple answers by crowds of non-expert Workers, e.g. in [19] and [17]. As the answers are inevitably noisy, different Workers were asked to survey the same road, and their responses are aggregated to decide what is the most likely and truthful observation.

Approaches to aggregation are an equally well studied subject, and a majority decision is a natural option (e.g. [14]). The detailed parameters and process of how consensus is defined and calculated are tuned for better performance and address issues specific to the context (e.g. in [12]).

In the case of OLAF, consensus is measured by using Fleiss' kappa statistics for inter-annotator agreement, as described for example in [15]. For those streets where the house numbers *were* found, a kappa of 0.6 on at least 5 surveys is sufficient consensus. For those streets where the house numbers were *not* found, a kappa of 0.8 on at least 10 surveys is required instead, as it is more likely that unreliable Workers agree in reporting that.

The number of 5 and 10 judgements is chosen because they are respectively the minimum number of judgements where 0.6 and 0.8 kappa can be achieved without the need of an unanimous agreement (4 vs 1 for 0.6 and 9 vs 1 for 0.8).

Rounds of 10 surveys per road are performed until consensus is reached on both its lowest and highest house numbers. Because of the nature of the task, new surveys are performed even if consensus is reached already on either of the two numbers.

Recruitment. We sourced all our Workers from CrowdFlower. For each experiment, we created one dedicated CrowdFlower job. We used identical settings for each experiment set, consisting of the following parameters: **Geography** Limited to the top 10 contributor countries in CrowdFlower where English is an official or officially recognised language.[27] **Skills** We chose Workers from the default CrowdFlower performance category (formerly named "level 2"), that accounts for 29% of the total population.[28] **Accuracy** As a test question, Workers were asked to copy and paste or type the name of the street as part of their submission in each task. Being the question this simple, the requested accuracy was 99%.[29] **Judgements** Each Worker is allowed to contribute to as many tasks as possible. Repeated judgements on the same street are allowed but ignored.[30] **Behaviour** Each Worker was paid for 1 task, and 1 task is made of 1 street to survey. **Reward / Time Limits** The reward was 0.20 US Dollars per task. Workers requiring less than 90 seconds per task were considered at high risk of being malicious and excluded to perform additional tasks. CrowdFlower imposes a time limit of 30 minutes maximum per task.

4.7 Implementation

The code implementing WODA and detailed documentation can be found on GitHub, starting from the instructions at `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/blob/gh-pages/README.md`.

5 Evaluation

5.1 Data

WODA was deployed for a specific geographic sample, that is the same five OS 1:10,000 raster tiles that were previously used in literature to analyse the performance of GI crowdsourcing and originally selected by Haklay in [10] to

[27] See `https://success.crowdflower.com/hc/en-us/articles/202703345-Crowd-Demographics`, the identification of the countries was done on 19 December 2015. The list of countries is: Bangladesh, Canada, India, Malaysia, Netherlands, Pakistan, Philippines, Sri Lanka, United Kingdom and United States of America.

[28] See `https://success.crowdflower.com/hc/en-us/articles/202703345-Crowd-Demographics`, the calculation was done on 19 December 2015.

[29] CrowdFlower does not allow the Requester to set target accuracy to 100%.

[30] This was a limitation of our configuration, explained later in this document.

assess OSM[31]. This is an area of 113 km^2 of Greater London that includes 3,982 named roads.

5.2 Evaluation metrics

The evaluation is aimed at observing WODA's effectiveness at producing house numbers for the sample. Given the premises described in chapter 1, a financial metric is used to measure performance, that is the average cost per road of using crowdsourcing for data production.

5.3 Results

Ingestion of primary data sources The data obtained by implementing the approach described in section 4 successfully populated house numbers from LRPP for 82% of the streets in OSON.

House number inference The conditions necessary to apply the inference algorithms, before using any crowdsourcing, were verified for 74% of roads. Applied to these, algorithms 1 and 2 generated 113k house numbers in addition to the already known 111k (+102%).

House number crowdsourcing Stop condition *s.2* - with 118 repeat judgments vs 117 first judgements - was verified at the end of three crowdsourcing rounds. The data for only 4 roads only was collected successfully, with an average cost of 9.00 USD per road.

To better examine any trends in Worker behaviour and consensus through iterations, three additional rounds were run, too.

More than 12% of Workers were caught submitting judgements earlier than the 90 seconds limit, and were excluded from contributing further. 19% failed the test of copying the name of the road in the form at least once, and thus were identified as not credible and their judgements ignored. The average time it took for credible Workers to complete the tasks was 6:14. 83% of them were always faster than 3:00, that means that they never watched the instructions video in full before submitting.

The extra three rounds showed how agreement on most roads not only failed to converge, but stalled or worsened with new iterations.[32]

6 Discussion

From a merely technical perspective, the experience of implementing WODA confirmed that it is feasible to implement a system that supports non-trivial human-machine hybrid workflows while relying only on the scalability and availability of third party services only.

[31] The tiles are: TQ37ne, TQ28ne, TQ29nw, TQ26se and TQ36nw.

[32] See `https://github.com/Digital-Contraptions-Imaginarium/OLAF-yr2_lab/tree/gh-pages/docs#fleiss-kappa-vs-iterations`.

Conversely, from a functional point of view, many of the chosen services' characteristics - and of the crowdsourcing platform in particular - ended up not to be compatible with the needs of our design. So, while we achieved avoiding custom Worker-facing software components by relying exclusively on CrowdFlower and Google Map's native features, we were also forced to implement many ad hoc workarounds. Several of these arrangements were far from ideal and affected negatively the performance of the overall system, to a point compromising its effectiveness and advantages.

For example, not having the possibility to rely on CrowdFlower's own metric (quorum vs Fleiss' kappa) did not allow us to calculate consensus as new judgements came into the system, but only offline, between crowdsourcing rounds. This forced us to collect more judgements than necessary. In turn, we also lost the option to use CrowdFlower's features that prevent Workers to judge the same item repeatedly, causing an overwhelming volume of unwanted judgements that stopped the experiment very early in respect to our expectations.

An equally pressing issue is the choice of crowdsourcing task, that failed to catalyse the contributors' agreement and produce results that are statistically credible. This suggests a substantial doubt on the effectiveness of crowdsourcing survey activities of this sort. We reckon that the key cause was a combination of (i) conventional cheating, (ii) the complexity and open-ended nature of the surveying task and (iii) some degree of sloppiness of participants even when in good faith. When aiming at such a high consensus target, as in our case, even one single point of disagreement can substantially negatively impact the metric.

Cheating in crowdsourcing does not require explanation and is common to most systems. The way the task was designed, it was easy for participants to just state that the surveyed street offered no house numbers.

The act of surveying a location - in person as in the interactive imagery - is not trivial. Some of its dynamics may seem intuitive, however not all participants would necessarily grasp them, even after watching the instructions video. Gottlieb *et al.* in [9] faced similar difficulty, e.g. in exploring crowdsourcing to geolocate places in videos.

In terms of sloppiness of the surveys, even in good faith many Workers may have been content when finding house numbers that looked "small enough" or "high enough", and interrupted their search there, despite what was stated in the instructions.

Finally, any discussion should still be filtered through a cost / benefit examination: what is a "reasonable cost" of producing one address with a target degree of confidence. The volume of roads we could enable inference for was too little to make such considerations.

7 Conclusions

We have presented WODA: a platform to integrate geospatial open data, computation and original human contribution to create original data, using human-machine hybrid workflows. To maximise the platform's scalability and availability, our design has relied as much as possible on the native features of mainstream SaaS services such as CrowdFlower and Google Maps. We have then implemented the platform to tackle components of one specific real life problem, that is the creation of OLAF: the list of all valid UK addresses. Where the conditions to enable computation were not verified, we deployed the platform to use paid microtask crowdsourcing to create the missing input data.

The evaluation of the platform has showed how critical the crowdsourcing component of the system is, particularly when it needs to support contributors in performing more complex activities than what is common in microtasking, such as surveying interactive imagery of locations.

Moreover, our experience with CrowdFlower has also showed how third party services' native features may be intrinsically inconsistent with the needs of a workflow, de facto forcing the system designer to write ad hoc software that uses their API instead. This adds complexity and points of failure to the overall solution that could be avoided instead, and possibly compromises scalability and availability.

Our plans for future work include exploring alternative configurations of preexisting open data, computation and human contribution where the crowdsourcing component can be designed to achieve a higher degree of success. We also intend to share our work with the community of geospatial practitioners in London at one of their upcoming regular meetings, to gather feedback and ideas for further development.

Another interesting direction of research is to both (a) investigate how to redesign WODA to make the best use of the crowdsourcing provider's native functionality, and (b) work with them to implement those missing features and/or work around the limitations we identified during this work.

Acknowledgements. This work is supported under SOCIAM: The Theory and Practice of Social Machines. The SOCIAM Project is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J017728/2 and comprises the Universities of Oxford, Southampton, and Edinburgh. The authors would also like to thank Dr. Sophie Stalla-Bourdillon and Prof. Muki Haklay for their involvement and advice.

References

1. Ballatore, A., Mooney, P.: Conceptualising the geographic world: the dimensions of negotiation in crowdsourced cartography. International Journal of Geographical Information Science pp. 1–18 (Aug 2015)
2. Butler, D.: The web-wide world. Nature 439(7078), 776–778 (2006)
3. Coleman, D.J., Georgiadou, Y., Labonte, J.: Volunteered geographic information: The nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research (2009)
4. Difallah, D.E., Demartini, G., Cudré-Mauroux, P.: Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In: CrowdSearch (2012)

5. Estes, J.E., Mooneyhan, D.W.: Of Maps and Myths. Photogrammetric Engineering and Remote Sensing 60(5), 517–524 (May 1994)
6. Flanagin, A.J., Metzger, M.J.: The credibility of volunteered geographic information. GeoJournal 72(3-4), 137–148 (2008)
7. Goetz, M., Zipf, A.: Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information. International Journal of 3-D Information Modeling (IJ3DIM) 1(2), 1–16 (Jan 1)
8. Goodchild, M.F.: Citizens as sensors: the world of volunteered geography. GeoJournal (2007)
9. Gottlieb, L., Choi, J., Kelm, P., Sikora, T., Friedland, G.: Pushing the limits of mechanical turk: qualifying the crowd for video geo-location. In: CrowdMM '12: Proceedings of the ACM multimedia 2012 workshop on Crowdsourcing for multimedia. p. 23. ACM Request Permissions, New York, New York, USA (Oct 2012)
10. Haklay, M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B Planning and Design (2010)
11. Haklay, M., Basiouka, S., Antoniou, V.: How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. The Cartographic Journal (2010)
12. Hirth, M., Hossfeld, T., Tran-Gia, P.: Cost-Optimal Validation Mechanisms and Cheat-Detection for Crowdsourcing Platforms. IEEE (2011)
13. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. ACM, New York, New York, USA (Apr 2008)
14. Le, J., Edmonds, A., Hester, V., Biewald, L.: Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In: SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (Jul 2010)
15. Nowak, S., Rüger, S.: How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: The 11th ACM International Conference on Multimedia Information Retrieval. pp. 557–566. ACM, New York, New York, USA (Mar 2010)
16. Schmid, F., Cai, C., Frommberger, L.: A new micro-mapping method for rapid VGI-ing of small geographic features. Geographic Information Science: 7th International Conference (Sep 2012)
17. Sheng, V.S., Provost, F., Ipeirotis, P.G.: Get another label? improving data quality and data mining using multiple, noisy labelers. In: 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 614–622. ACM (Aug 2008)
18. Simperl, E.: How to Use Crowdsourcing Effectively: Guidelines and Examples. Liber Quarterly 25(1), 18 (Aug 2015)
19. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In: Conference on Empirical Methods in Natural Language Processing. pp. 254–263. Association for Computational Linguistics (Oct 2008)

Consent to Publish

Lecture Notes in Computer Science

Springer

Title of the Book or Conference Name: 16th Int. Conference on Web Engineering (ICWE2016)

Volume Editor(s):

Title of the Contribution: WODA: a platform for geospatial open data augmentation using ...

Author(s) Name(s): G. Cecconi (1), E. Simperl (1), L. Carr (1), N. Shadbolt (2)

Corresponding Author's Name, Address, Affiliation and Email: (1) University of Southampton, University Road, Southampton, SO17 1BJ, UK, (2) University of Oxford, University Offices, Wellington Square, Oxford, OX1 2JD, UK. Emails: gc1a13@soton.ac.uk, e.simperl@soton.ac.uk, lac@ecs.soton.ac.uk, nigel.shadbolt@jesus.ox.ac.uk

When Author is more than one person the expression "Author" as used in this agreement will apply collectively unless otherwise indicated.

§ 1 Rights Granted

Author hereby grants and assigns to Springer International Publishing AG, Cham (hereinafter called Springer) the exclusive, sole, permanent, world-wide, transferable, sub-licensable and unlimited right to reproduce, publish, distribute, transmit, make available or otherwise communicate to the public, translate, publicly perform, archive, store, lease or lend and sell the Contribution or parts thereof individually or together with other works in any language, in all revisions and versions (including soft cover, book club and collected editions, anthologies, advance printing, reprints or print to order, microfilm editions, audiograms and videograms), in all forms and media of expression including in electronic form (including offline and online use, push or pull technologies, use in databases and networks for display, print and storing on any and all stationary or portable end-user devices, e.g. text readers, audio, video or interactive devices, and for use in multimedia or interactive versions as well as for the display or transmission of the Contribution or parts thereof in data networks or seach engines), in whole, in part or in abridged form, in each case as now known or developed in the future, including the right to grant further time-limited or permanent rights. For the purposes of use in electronic forms, Springer may adjust the Contribution to the respective form of use and include links or otherwise combine it with other works. For the avoidance of doubt, Springer has the right to permit others to use individual illustrations and may use the Contribution for advertising purposes.

The copyright of the Contribution will be held in the name of Springer. Springer may take, either in its own name or in that of copyright holder, any necessary steps to protect these rights against infringement by third parties. It will have the copyright notice inserted into all editions of the Contribution according to the provisions of the Universal Copyright Convention (UCC) and dutifully take care of all formalities in this connection in the name of the copyright holder.

§ 2 Regulations for Authors under Special Copyright Law

The parties acknowledge that there may be no basis for claim of copyright in the United States to a Contribution prepared by an officer or employee of the United States government as part of that person's official duties. If the Contribution was performed under a United States government contract, but Author is not a United States government employee, Springer grants the United States government royalty-free permission to reproduce all or part of the Contribution and to authorize others to do so for United States government purposes.

If the Contribution was prepared or published by or under the direction or control of Her Majesty (i.e., the constitutional monarch of the Commonwealth realm) or any Crown government department, the copyright in the Contribution shall, subject to any agreement with Author, belong to Her Majesty.

If the Contribution was created by an employee of the European Union or the European Atomic Energy Community (EU/Euratom) in the performance of their duties, the regulation 31/EEC, 11/EAEC (Staff Regulations) applies, and copyright in the Contribution shall, subject to the Publication Framework Agreement (EC Plug), belong to the European Union or the European Atomic Energy Community.

If Author is an officer or employee of the United States government, of the Crown, or of EU/Euratom, reference will be made to this status on the signature page.

§ 3 Rights Retained by Author

Author retains, in addition to uses permitted by law, the right to communicate the content of the Contribution to other scientists, to share the Contribution with them in manuscript form, to perform or present the Contribution or to use the content for non-commercial internal and educational purposes, provided the Springer publication is mentioned as the

original source of publication in any printed or electronic materials. Author retains the right to republish the Contribution in any collection consisting solely of Author's own works without charge subject to ensuring that the publication by Springer is properly credited and that the relevant copyright notice is repeated verbatim.
Author may self-archive an author-created version of his/her Contribution on his/her own website and/or the repository of Author's department or faculty. Author may also deposit this version on his/her funder's or funder's designated repository at the funder's request or as a result of a legal obligation. He/she may not use the publisher's PDF version, which is posted on SpringerLink and other Springer websites, for the purpose of self-archiving or deposit. Furthermore, Author may only post his/her own version, provided acknowledgment is given to the original source of publication and a link is inserted to the published article on Springer's website. The link must be provided by inserting the DOI number of the article in the following sentence: "The final publication is available at Springer via http://dx.doi.org/[insert DOI]". The DOI (Digital Object Identifier) can be found at the bottom of the first page of the published paper.
Prior versions of the Contribution published on non-commercial pre-print servers like ArXiv/CoRR and HAL can remain on these servers and/or can be updated with Author's accepted version. The final published version (in pdf or html/xml format) cannot be used for this purpose. Acknowledgment needs to be given to the final publication and a link must be inserted to the published Contribution on Springer's website, by inserting the DOI number of the article in the following sentence: "The final publication is available at Springer via http://dx.doi.org/[insert DOI]".
Author retains the right to use his/her Contribution for his/her further scientific career by including the final published paper in his/her dissertation or doctoral thesis provided acknowledgment is given to the original source of publication. Author also retains the right to use, without having to pay a fee and without having to inform the publisher, parts of the Contribution (e.g. illustrations) for inclusion in future work, and to publish a substantially revised version (at least 30% new content) elsewhere, provided that the original Springer Contribution is properly cited.

§ 4 Warranties

Author warrants that the Contribution is original except for such excerpts from copyrighted works (including illustrations, tables, animations and text quotations) as may be included with the permission of the copyright holder thereof, in which case(s) Author is required to obtain written permission to the extent necessary and to indicate the precise sources of the excerpts in the manuscript. Author is also requested to store the signed permission forms and to make them available to Springer if required.
Author warrants that he/she is entitled to grant the rights in accordance with Clause 1 "Rights Granted", that he/she has not assigned such rights to third parties, that the Contribution has not heretofore been published in whole or in part, that the Contribution contains no libelous statements and does not infringe on any copyright, trademark, patent, statutory right or proprietary right of others, including rights obtained through licenses; and that Author will indemnify Springer against any costs, expenses or damages for which Springer may become liable as a result of any breach of this warranty.

§ 5 Delivery of the Work and Publication

Author agrees to deliver to the responsible Volume Editor (for conferences, usually one of the Program Chairs), on a date to be agreed upon, the manuscript created according to the Springer Instructions for Authors. Springer will undertake the reproduction and distribution of the Contribution at its own expense and risk. After submission of the Consent to Publish form Signed by the Corresponding Author, changes of authorship, or in the order of the authors listed, will not be accepted by Springer.

§ 6 Author's Discount

Author is entitled to purchase for his/her personal use (directly from Springer) the Work or other books published by Springer at a discount of 33 1/3% off the list price as long as there is a contractual arrangement between Author and Springer and subject to applicable book price regulation. Resale of such copies or of free copies is not permitted.

§ 7 Governing Law and Jurisdiction

This agreement shall be governed by, and shall be construed in accordance with, the laws of Switzerland. The courts of Zug, Switzerland shall have the exclusive jurisdiction.

Corresponding Author signs for and accepts responsibility for releasing this material on behalf of any and all Co-authors.

Signature of Corresponding Author: **Date:**

Gianfranco C .... 17/1/2016 ....

□ I'm an employee of the US Government and transfer the rights to the extent transferable (Title 17 §105 U.S.C. applies)

□ I'm an employee of the Crown and copyright on the Contribution belongs to Her Majesty

□ I'm an employee of the EU or Euratom and copyright on the Contribution belongs to EU or Euratom