Documenting flooding areas calculation: a PROV approach

Monica De Martino* and Alfonso Quarati

Institute of Applied Mathematics and Information Technology,
National Research Council,
Genova, Italy
Email: demartino@ge.imati.cnr.it
Email: quarati@ge.imati.cnr.it
*Corresponding author

Sergio Rosim and
Laércio Massaru Namikawa

Instituto Nacional de Pesquisas Espaciais (INPE),
National Institute for Space Research,
São Paulo, Brazil
Email: sergio.rosim@inpe.br
Email: namikawa@inpe.br

Abstract: Flooding events related to waste-lake dam ruptures are one of the most threatening natural disasters in Brazil. They must be managed in advance by public institutions through the use of adequate hydrographic and environmental information. Although the Open Data paradigm offers an opportunity to share hydrographic data sets, their actual reuse is still low because of metadata quality. Our previous work highlighted a lack of detailed provenance information. The paper presents an Open Data approach to improve the release of hydrographic data sets. We discuss a methodology, based on W3C recommendations, for documenting the provenance of hydrographic data sets, considering the workflow activities related to the study of flood areas caused by the waste-lakes breakdowns. We provide an illustrative example that documents, through W3C PROV metadata model, the generation of flooding area maps by integrating land use classification, from Sentinel images, with hydrographic data sets produced by the Brazilian National Institute for Space Research.

Keywords: hydrography data sets; open data; reusability provenance workflow; metadata; W3C PROV.

Reference to this paper should be made as follows: De Martino, M., Quarati, A., Rosim, S. and Namikawa, L.M. (2021) 'Documenting flooding areas calculation: a PROV approach', *Int. J. Metadata Semantics and Ontologies*, Vol. 15, No. 1, pp.50–59.

Biographical notes: Monica De Martino has been a Researcher at the Institute for Applied Mathematics and Information Technology (IMATI) of the Italian National Research Council (CNR) in Genoa since 1992. She holds a degree in Mathematics. She has been scientifically responsible for many European Project leading the activity related to the Knowledge Technology for Geographic Information Management. Her research expertise and interest are in the fields of knowledge management including metadata management, semantics analysis, open and linked data, data quality and use. She co-authored more than 60 scientific publications on journals, conferences proceedings and reports.

Alfonso Quarati has been a Researcher at the Institute for Applied Mathematics and Information Technology (IMATI) of the Italian National Research Council (CNR) in Genoa since 1997. He has co-authored more than 70 scientific papers, published in journals, book chapters and conference proceedings, and participated in several EU and national funded projects. His former research activities firstly focused on e-learning technologies and then on distributed architectures. Recently, his research interests focus on methodologies for publishing open research data, and the assessment of open data quality and use.

Sergio Rosim holds a Bachelor's degree in Computer Science from the Federal University of São Carlos (1980), a Master's degree in Electronic and Computer Engineering from the Technological Institute of Aeronautics (1999) and a Doctorate in Applied Computing from the National Institute for Space Research (2008). He is a Senior Technologist at the National Institute for Space Research. He has experience in Computer Science, with Emphasis on Geo-Processing. Since 1997, he has been dedicated to the study and development of software that relates hydrological modelling to numerical models of land. Currently, he is responsible for the development project of the computational platform TerraHidro for the development of distributed hydrological models.

Laércio Massaru Namikawa holds a degree in Electronic Engineering from the University of Vale do Paraíba (1988), Master's degree in Applied Computing from the National Institute for Space Research (1995) and PhD degree in Geography from the State University of New York in Buffalo (2006). Currently, he is a Senior Technologist at the National Institute for Space Research. He has experience in the field of computer science, with an emphasis on environmental modelling and graphic processing, acting mainly on the following themes: dynamic spatial models, numerical modelling of terrain, irregular triangular grids, hydrological modelling, warning systems for natural disasters, colour and fusion in remote sensing.

This paper is a revised and expanded version of a paper entitled 'Hydrographic Datasets in Open Government Data Portals: Mitigation of Reusability Issues Through Provenance Documentation' presented at the 'MTSR 2019: Metadata and Semantic Research', Rome, Italy, 28–31 October 2019.

1 Introduction

Brazil is a country with a long history of minerals mining activity which causes important environmental pollution related to the waste-lakes that remain in place, even after the end of the activities. The rupture of a mineral waste-lake causes important and/or permanent damages, such as loss of human life, and damages that last for a long time. One of the most recent events is the Brumadinho waste dam burst on January 25, 2019, releasing a volume of more than 11 million cubic metres of tailing causing at least 270 deaths (Armada, 2019).

Tackling these natural menaces requires prompt responses from the governmental bodies both to prevent and recover from the consequences of such calamities. To be effective such reactions have to rely on accurate, up-to-date and real-time available spatial information such as surface data sets, Digital Elevation Model (DEM), hydrographic data sets (Rosim et al., 2018; Geiger and Von Lucke, 2019), and remote sensing data which detect the physical characteristics of large areas.

The Open Data (OD) movement is playing a relevant role in the geospatial sector, by introducing a paradigm shift in the supply and use of geodata that is provided for free, in a machine-readable format and with minimal restrictions on reuse (Coetzee et al., 2020; Johnson et al., 2017). Open Government Data (OGD) is data published by governments and public agencies on the Web, without restrictions for data sharing and reuse (Geiger and Von Lucke, 2019, 2012). Several OGD portals, from the local to the international scale, have been designed for releasing data sets openly and to making them traceable and re-usable. However, as we observed in Quarati and De Martino (2019); Quarati and Rafiaghelli (2020); Quarati et al. (2021), just publishing OD di-per-se does not necessarily grant their reuse.

Metadata, data describing data, are published along with open data sets to support data consumers to understand the meaning of data and enable their discovery. To this end, the quality of the metadata is fundamental to grant data sets' reuse by third-party applications providing benefits such as clarity, organisation, detailing, integrity, accessibility and meaning (Ribeiro, 2018). Therefore, a lack of metadata, or a naive version of it, can be prejudicial (Sadiq and Indulska, 2017; Safarov et al., 2017; Máchová and Lnénicka, 2017). Besides, good quality metadata may support the reproducibility of the computational processes underlying data production (Perez et al., 2017). The FAIR[1] data principles for scientific data management refer to a concise and measurable set of principles which may act as a guideline for those wishing to enhance the reusability of their data. Among them, the provision of provenance information is recommended to improve data reuse. According to the W3C Provenance Working Group[2], *provenance* information about a piece of data can be used to assess its quality, reliability or trustworthiness.

To preserve the meaning of data by describing data creation and transformation (*workflow provenance*) is particularly important in the case of hydrographic resources. The study of flooding events involves the integration of several intertwined activities, with the input of one activity that is the output of another, following a sequential workflow pattern (Máchová and Lnénicka, 2017; Van Der Aalst et al., 2003). Provenance documentation allows to provide answers to the key question "Where did a particular piece of data come from?", by providing information on inputs lineage, on the assumptions, parameters and tools used in data processing, on data producers, and so on.

In our previous work, De Martino et al. (2019) and Garoufallou et al. (2019), we performed an evaluation of the hydrographic data sets published in some relevant OGD Portals all over the world by analysing their compliance with reusability practices according to W3C recommendations and FAIR principles. Among issues concerning the availability of machine readable formats or the openness of data sets' licence, our study highlighted an overall lack of provenance information.

This paper aims at presenting a practice to support the exploitation of hydrographic data sets produced by INPE, providing provenance metadata, i.e. the processing workflow for the quantitative analysis of the flooding risk areas. In particular, we will focus on the documentation of the integration of INPE hydrographic data with Earth Observation open data provided by European Spatial Agency Copernicus

Programme[3] supported by the European Space Agency, to manage the natural disasters related to dam breaking in Brazil. Provenance metadata of the processing workflow is published in a machine-readable format for the realisation of further concrete applications. We also exemplify the adoption of standards to represent provenance metadata and the publishing of several hydrographic data sets according to OGD practices and their delivery through the OD management platform datahab.io.[4] Information about the workflow activities also serves to elucidate to other interested parties the scientist decisions according to parameters (e.g. density of drainage networks) supplied to the workflows or the characteristics of data input (e.g. DEM resolution, contour line value of flood areas) that may affect the outcomes of the workflow.

2 Motivation

Floods caused by dam rupture represent in Brazil one of the main catastrophic natural disasters causing relevant damages to protected areas, crops, livestock and properties. There are 769 mineral waste-lakes, 425 of which are part of the National Dam Safety Policy[5] (PNSB, in Portuguese) and 344 are not part of the PNSB. This indicates a situation of catastrophic environmental risk for the country. The breaking of the dam of a lake of mineral waste releases large quantities of waste into its surroundings, destroying the lives of people and animals and polluting the aquatic and terrestrial environment near and far from the lake. For example, Mariana's waste-lake dam broke on November 5, 2015, releasing a volume of more than 40 million cubic metres of waste that caused 19 deaths (Fernandes et al., 2016). The ”Doce” River had more than 600 km of pollution that caused the death of more than 11 tons of fish in just one month. Events that happened in Mariana in 2015 or Brumadinho in 2019 may occur in the other 769 mineral waste-lakes in Brazil causing even greater damage. One way of mitigating these possibilities is prevent such possible events by studying and understanding the causes that can contribute to the rupture of a dam and the social, economic and environmental consequences of such a rupture. It is essential to study this event and to perform a periodic classification of the status of the dams delineating the locations of inundation and water accumulation to prevent possible natural damages and outbreaks of waterborne diseases. Cooperation between the government body and the scientific communities is fundamental to predict such natural disaster. In particular, it needs sharing data and information in a transparent and easy to use way for further processing.

The National Institute for Space Research (INPE),[6] located in the city of Sao Jose dos Campos, Brazil, produces and provides hydrographic data sets aimed at supporting decision activities to governmental institutions and private organisations to cope with environmental issues. To cope with and prevent dramatic environmental events such as those that occurred at Mariana and Brumadinho, INPE is performing a study of waste-lake dams breakdown effects aimed at supplying both forecasting risks maps and detailed surface maps capable of tackling ex-post flood events. The designed methodology integrates INPE drainage data sets with Sentinel data provided by the European Copernicus Programme for Earth Observation, and it is characterised by these activities: (1) The study of the waste-lake hydrographic basin, i.e. the parcel of land that contributes to sediments, such as soil, vegetation remains and others, transported by the runoff of rainwater, to the waste lake. This study aims to identify elements that can, over time, cause the dam breaking; (2) The study of the impact area of the first waste wave flow, i.e. the area affected by the first wave if (or when) the dam breaks down; (3) Tracing of polluted rivers, i.e. of the natural path where the waste will flow from the breaking point of the lake, up to a watercourse, polluting the entire course up to the ocean.

In the paper, we focus on the second activity carried out by INPE, the study of delimiting the first waste wave effects on the surrounding area. We will show how to provide proper documentation of the data and steps involved in this workflow, thus enhancing the reuse of its outcomes.

3 Background

3.1 Geospatial open data and metadata

The OD movement has involved the geospatial sector for all of its existence (Coetzee et al., 2020). Already, in 2011 the OD for Resilience Initiative started to apply the OD practices (Open Government Working Group, 2007) to face vulnerability to natural hazards and the impacts of climate change. Currently, one of the most prominent geospatial OD portals is the Copernicus Open Access Hub[7] provided by the European Union's Copernicus Programme supported by the European Space Agency which delivers a growing volume of Sentinel satellite data in real-time for the monitoring of the Earth ecosystem. Although there has been significant progress in opening up this type of free data, it has not yet brought the expected effects as its use is still challenging (Umbrich et al., 2015; Beno et al., 2017). Janssen et al. (2012) argued that research is needed to address barriers by studying a greater understanding of the user perspective before Open Data systems are freely adopted. The main problem is to make the user able to understand the data correctly before any use. That information should be found in the metadata coupled with the data published. Thus, it is essential to provide data enriched with valuable metadata.

Among several metadata standardisation initiatives that have been established to support data exchange and understanding among different communities, the W3C Data on the Web Best Practices (W3C-DWBP)[8] related to the publication and usage of data on the Web, and the FAIR[9] Guiding Principles for scientific data management and stewardship, provide recommendations on data publication, accessibility, interoperability and reuse on the web. For example, W3C recommends the machine readability of data formats, and the FAIR reusability principle includes metrics such as the availability of usage openness licenses and the provision of metadata provenance. These principles and guidelines should be adopted by OGD portals to improve their reuse. For instance, more care should be placed on data format to avoid interoperability and integration issues. In particular, the use of a structured and machine-readable file format such as the Resource Description Framework (RDF)[10] is recommended for data interchange on the Web. Metadata should adopt the standard Data Catalogue Vocabulary (DCAT),[11] an RDF vocabulary designed by the W3C to facilitate metadata discoverability and to allow interoperability between data catalogues published on the web (Umbrich et al., 2015; Máchová and Lnénicka, 2017; Alemu and Garoufallou, 2020; Neumaier et al., 2016). The availability of these standards alone, however, does not guarantee the production of appropriate metadata and their association with the corresponding data sets. Several factors, such as lack of skills by metadata providers, or the lack of well-designed metadata editors, can hamper the productions of appropriate metadata, thus hindering the reuse of OGD data sets (Sadiq and Indulska, 2017).

3.2 Provenance metadata

W3C-DWBP recommends the provision of provenance metadata to describe the history of a data set to facilitate its reuse. For Ram and Liu (2007) the semantic of provenance consists of seven elements 'what', 'where', 'when', 'who', 'how', 'which' and 'why'. According to the W3C Provenance Working Group, "Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness". The use of provenance metadata is recommended both by FAIR in "R1.2. (meta) data are associated with their provenance" and W3C-DWBP in "Best Practice 5: Provide data provenance information". To be machine-readable, provenance should be expressed through shared ontologies like the Provenance Ontology (PROV-O)[12] realised by W3C (Sahoo et al., 2013; Moreau and Groth, 2013).

The documentation of workflows' provenance with shared standards and practices provides consumers with insights about how and why these data were obtained (Hartig and Zhao, 2010). The provenance of scientific results, i.e. how results were derived, what parameters influenced the derivation, what data sets were used as input to the experiment, helps reproducibility of the whole process (Gil et al., 2007). Workflows literature uses to distinguish prospective and retrospective provenance. The former provenance models workflow in an abstract and informative way. The latter models past workflow executions, informing about what task has been executed and how data artefacts were derived (Lim et al., 2010). Retrospective provenance does not depend on the presence of prospective (Freire et al., 2008). Notwithstanding through retrospective provenance, it is possible to capture the relevant details that occurred during workflow execution, thus making it understandable and reproducible. In this sense, data provenance documentation is crucial, especially in the public sector when data are used in support of policymaking (Perego, 2017). Several workflow management systems provide provenance collection mechanisms to capture retrospective provenance (Freire et al., 2008; Miksa and Rauber, 2017). However, such information is rarely supplied in a formal, machine-readable way, enabling a (semi)automated data processing workflow reproducibility (Perego, 2017).

This work aims at increasing the awareness of Open Data providers of the importance of supplying such data according to proper provenance documentation practices. To this end, we exemplify a W3C compliant solution to document the whole process involved in the study of the effects of the first wave of waste-lake ruptures.

3.3 Hydrographic OGD data sets reuse

The analysis of OGD portals metadata shows that despite the importance of access traceability, accountability, and accuracy of data none of the portals properly provides provenance information (Marcelo et al., 2016); usually, they are limited to the "Who" and "When" metadata such as publisher/organisation and the maintainer/ContactPoint names (Moreau et al., 2015).

In De Martino et al. (2019), we performed an analysis of currently available hydrographic data sets publishing practices. With the term of *hydrographic data sets* we refer to a set of relevant terrain descriptors such as drainage networks, basins, flood risk areas, watersheds and rivers. In the study we investigated the first nineteen OGD portals ranked by the Open Data Barometer[13] index, aimed at evaluating how leading governments have been performing during the last decade into the Open Data movement. The index outlines what needs to happen for the movement to progress forward. The report of the recent edition looks specifically at governments that have made concrete commitments to OD, either by adopting the Open Data Charter,[14] or by signing up to the G20 Anti-Corruption Open Data Principles.[15]

In the analysis, it was identified the hydrographic data sets available in the nineteen OD portals by keywords search. According to W3C-DWBP ("Best Practice 15: Reuse vocabularies, preferably standardised ones") and FAIR ("Reusable principle, R1. meta(data) are richly described with a plurality of accurate and relevant attributes"), the choice of the keywords should not be user-dependent. Controlled vocabularies (i.e. thesauri and code lists) encoded in Simple Knowledge Organisation System SKOS (Miles and Bechhofer, 2009) should to be used as a semantic layer which facilitates data search (Albertoni et al., 2018). Good quality vocabularies (Quarati et al., 2017) provide a key to disclosing the potential of OGD, by supplying common terms for marking up metadata and data consistently and coherently (Albertoni et al., 2018). We exploited the multilingual linked thesaurus framework LusTRE[16] which provides a unique point of access to several Environmental thesauri and code lists (e.g. GEMET,[17] EARTh (Albertoni et al., 2014), ThiST and AGROVOC[18]). They are encoded in SKOS and are published and linked according to the Linked Data Best Practices.[19] This allows cross-navigating between thesauri enlarging the space of concepts that can be browsed and used for data discovery.

Our analysis, carried out on May 2019, revealed that the hydrographic data sets amounted to 89,817, about 14% of the overall 654,454 data sets published in the considered OGD portals. We then analysed the compliance of those hydrographic data sets with respect to some reusability dimensions as recommended by FAIR and W3C-DWBP initiatives. We focused on three metrics: machine-readable standardised data formats (W3C-DWBP Best Practice 12), clear and accessible open usage licence (FAIR R1.1), and provision of detail provenance (FAIR R1.2). Only 52% of the sample's data sets are provided in a machine-readable format and very few in RDF. As to the existence of license information and the compliance to openness, as defined in the licenses list reviewed by the Open Definition,[20] almost all portals' data sets have associated an open licence (CC or OGL), with the exception of the US portal[21] which provide its revised open access and use licence.[22] Finally, our investigation pointed out that the provision of provenance metadata is limited, it sometimes includes some information about "Who" (i.e. authors or publisher contact information), and "When" (i.e. the date of data sets and metadata publication), but not information about "Why" and "How" the hydrographic data sets have been produced. The provision of proper documentation of the data sets' production workflow is needed to elucidate the process of data creation, and help users to clarify the design and experimental choices that yielded a hydrographic piece of data. These supplement data can facilitate data reuse in situations such as the prevention of flooding risk areas, caused by dams rapture.

4 Methods

The documentation of the information flows and of the specific steps and tools involved in a flooding risk areas calculation leverages on the one hand on the W3C PROV Data Model (Missier et al., 2013), and on the other on the proper methodology and tools developed by INPE).

4.1 PROV data model

PROV Data Model describes provenance in terms of relationships between three main types of concepts: *prov:Entity*, which represents (physical, digital, or other types of things); *prov:Activity*, which occur over time and can use and/or generate entities; and *prov:Agent*, which are responsible for activities occurring, entities existing, or another agent's activity. Relationships between these concepts describe the influence one has had on another. In particular, PROV-O specifies seven core properties to relate the aforementioned artifacts, for example, *prov:used* indicates that an activity used some entities; *prov:wasAttributedTo* indicates an entity was attributed to an agent, *prov:wasDerivedBy* indicates an entity was derived by another entity, *prov:wasGeneratedBy* indicates that activity generated an entity. The nature of the influence can be defined using qualified relations to describe the *prov:Role* of the entity, agent, or activity. Qualified relations include: *prov:Usage*, which defines the role of an entity used by an activity; and *prov:Association*, which defines the role of an agent in an activity, along with any *prov:Plan* the agent was following during the activity.

The paper exploits the aforementioned vocabularies to represents the workflow specifications and their execution.

4.2 Waste-lake dams breakdowns calculation workflow

The study of the first wave effects after a waste-lake dam breakdown aimed both as forecasting or recovery tool. It is based on a processing workflow characterised by three main steps: i) hydrographic data sets generation; ii) Sentinel image classification; and iii) hydrographic data sets integration with the Sentinel image classification. The data flow involves as input, the drainage network, the area of potential flooding calculated by a DEM, the classification of the satellite image, and it provides the integrated map of the elements of land use and coverage potentially affected by the flood event. All these data are produced by three tools developed by INPE a) TerraHidro,[23] is a distributed hydrology modelling platform which generates drainage networks and basins from a DEM; b) HAND, the Height Above the Nearest Drainage algorithm, is a low-cost solution in the absence of detailed hydrological and hydraulic data which predicts the location and spatial extent of potential inundation and c) SPRING[24] a state-of-the-art GIS and remote sensing image processing system. The steps of the flooding risk areas calculation workflow operate as follows:

1 *Hydrographic data sets generation.* The first step to generate a potential flooding areas map is the extraction of drainage network and basins from an SRTM-DEM by TerraHidro (Rosim et al., 2018). TerraHidro runs a sequence of computations to create a hydrologically coherent DEM, define flow paths and delineate the drainage channels. The second step generates the raster map of the location and spatial extent of potential inundation by HAND. It processes the grids of altimetry, the local drainage direction and the drainage network of the interested area (in our study, the one that starts at the waste-lake dam), and computes a raster map where each cell value is the difference of elevation between the surface value of a cell of the DEM and the river bed cell to which it drains. To visualise the HAND result, it is necessary to generate a slice with the desired contour lines. Each curve indicates a height of possible arrival of the wastes released by the first wave. A detailed description of HAND algorithm was presented in Rennó et al. (2008).

2 *Sentinel data classification.* To identify the elements of land use and cover, Sentinel 1 imagines provided by Copernicus programme are classified by functionalities supplied by SPRING. A classification operation consists of (i) a segmentation process using the region growth method and (ii) the application of an unsupervised classifier by region, named Isoseg[25] to group the segmented regions showing the similarities between the existing segments in the segmented image.

3 *Integration of Hydrographic Data set and Sentinel data classification.* The classified imagines and the potential flooding areas map are overlapped by SPRING to identify the flooding areas damages after a dam breakdown. This integration allows verifying the elements that risk of totally or partially covered by the waste of the first wave, which is usually the most intense and fast.

In the following section, we discuss the documentation of this workflow applied to the analysis of the effects of a real case of waste-lake dam breakdown through the PROV data model.

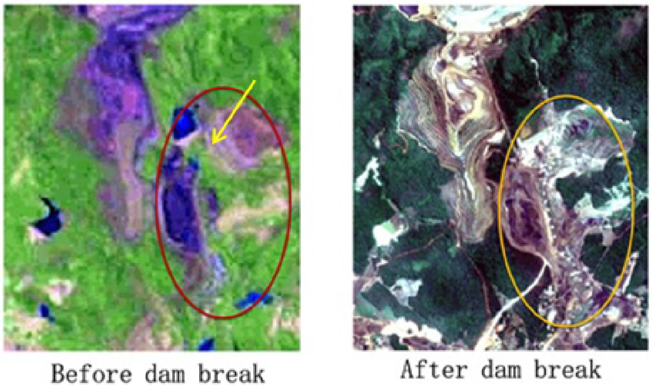
5 Workflow documentation

5.1 Illustrative use case

As illustrative use case, we considered the Brumadinho dam, which is located in the Minas Gerais region, in southeastern Brazil. The coordinate of this dam is latitude S 20° 07' 9.71" and longitude W 44° 07' 17.52". This is the region with the largest number of mineral tailings lakes in the country with 690 dams. Minas Gerais has an area of 586,528 km^2 and a population of 21,168,791 people.

Figure 1 refers to the Sentinel 2B images of the study area and before and after the dam rupture. They show the difference in the land cover before and after the Brumadinho dam rupture. The break occurred at the point indicated by the yellow arrow in the figure on the left. In the image on the right, we can see the vegetation areas covered by mineral waste. Satellite images and altimetric data are extremely helpful in carrying out studies in this area and the like containing lake waste.

Figure 1 Sentinel-2B satellite images of the Brumadinho area before and after the dam rupture: the first wave instantly reached the administrative area of the company responsible for the dam (yellow ellipse)

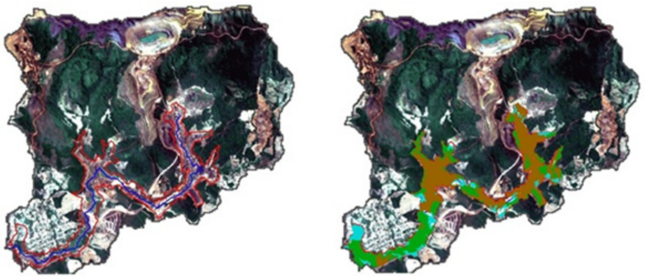


To analyse the impact of the dam breakdown on the surrounding area, we based on the SRTM altimetry data and the Sentinel image for the land use and land cover. SRTM, covering all Brazil country, allows a uniformity among various areas containing lakes of wastes.

The Sentinel-2B image used was an optical image of 10 m resolution, bands 2, 3, 4, respectively associated with B, R, G colours. First, a contrast was applied to each colour channel to enhance image comprehension. Then, these bands were used in the segmentation process, with similarity value 30 and a filter of minimum area of 100 m was applied. The definition of these values depends on the specialist's knowledge. At this point, the classification was performed using these bands and the segmented image. The evaluation of the classification result can induce the specialist to change the values for identifying similarity and minimum extension areas. For example, using a more restricted similarity value allows obtaining more regions.

Figure 2 shows on the left the Sentinel image used after the Brumadinho dam breakdown occurrence day. The red contour, computed by HAND, highlights the locations affected by the first waste wave considering 20 m slice, the black line shows the drainage network interested by the dam rupture. The figure on the right presents the classification of this area where the black, green and brown colours respectively represent the urban, vegetation and exposed soil areas. To make a more detailed check of the results of HAND, a hydrological model to simulate the dynamic behaviour of the tailings runoff can be used. This study will be performed in future research.

Figure 2 Bumadinho area: (left) Sentinel image GRB composition, drainage network interested from the dam (black), flood risk area (red); (right) Integration of Sentinel image classification with the map of flood area: urban area (black), exposed soil (brown), Vegetation (green)



We want to point out that to analyse the quality of the results the type of altimetry data (e.g. surface or relief) and the resolution of this data have to be considered. For instance, we used surface altimetry data with a resolution of 30 m. To improve the accuracy of the results higher resolution data (e.g. 5 m) could be used. The resolution and altimetry type data affect how far the work can go, in quantitative terms. However, as we meant to supply a qualitative analysis of the possible damage caused by the rupture of a mineral wastes dam, our goal has been to discriminate the elements that could have suffered economic and social damages with this event and not to quantify these damages. Therefore, we have decided to use medium resolution (i.e. 30 m) altimetry data that is freely available and widely used by the community working on elevation data. Moreover, the medium resolution allows verifying the feasibility of developing a computational workflow without field validation but only comparing the result qualitatively with Sentinel images. Furthermore, to improve accuracy should have required to incorporate other tools, such as a hydrodynamic model and high-resolution altimetry data, which are expensive, and need to carry out fieldwork, and are out of the scope of our work aimed to provide interested users with a free and agile method to determine the potential of destruction caused by the rupture of a dam.

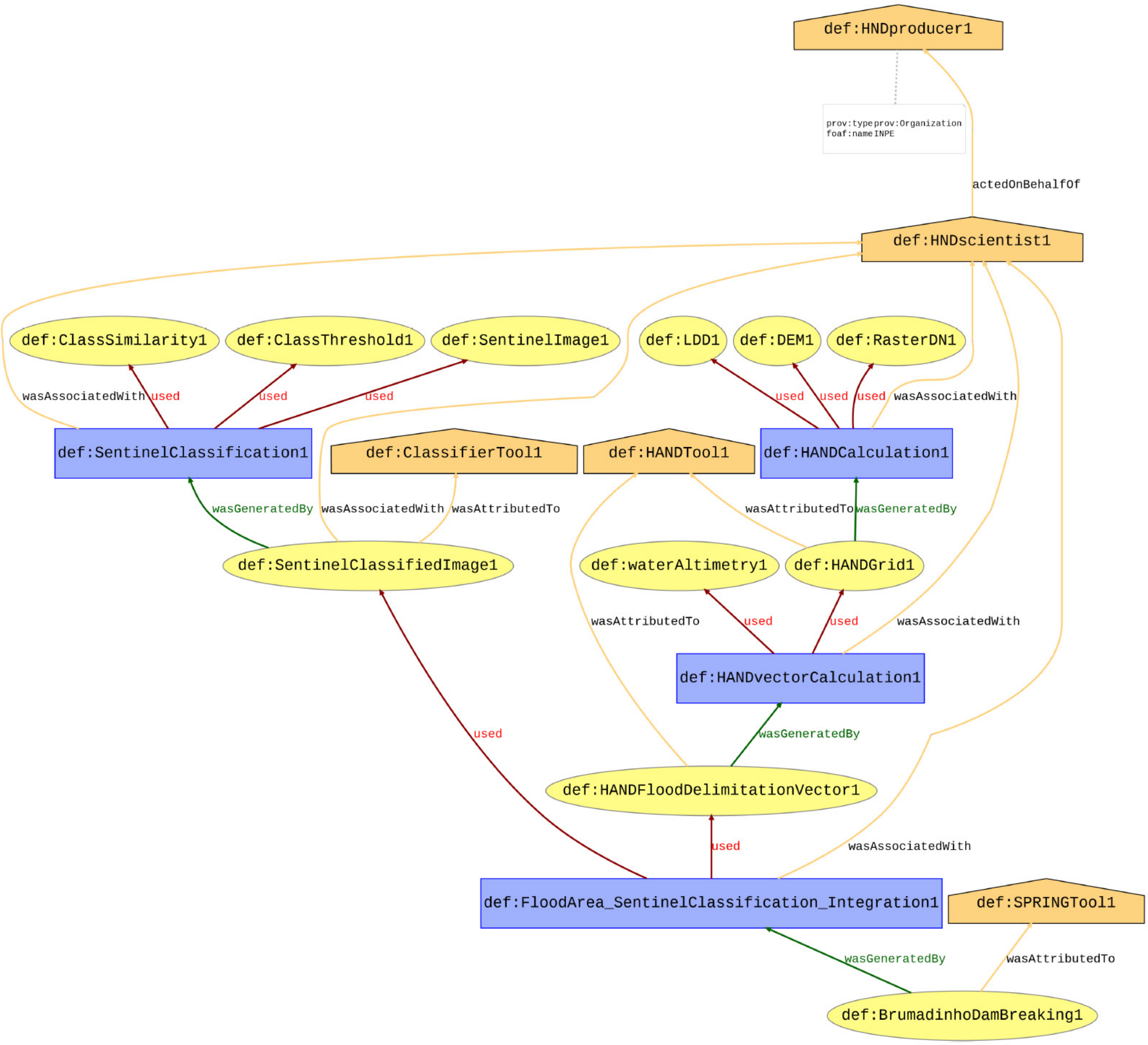
We made available on the web all the data sets produced in this use case as Open Data on the data portal datahub.io, accessible through the dereferenceable URI "Anonymised for double reviewing". The data file management is carried out by INPE and stored in a centralised repository.[26]

5.2 Modelling workflow provenance with PROV

Provenance data that documents workflow typically describes a graph, with steps and data as nodes and input and output connections as edges (Moreau et al., 2015). The provenance diagram in Figure 3 summarises the PROV-O elements involved in the process execution to study the first-wave effects of the Brumadinho dam breakdown. It is produced by mapping the PROV description provided in RDF-Turtle syntax to the dot graphic notation, by means of the ProvStore[27] online tool and subsequent rearrangements to make it better readable.

The diagram uses the graphical notation introduced in Sahoo et al. (2013) to depict the elements of the PROV model. The Entities (i.e. data sets used and produced, parameters) are depicted as yellow ovals, the Activities (i.e. the steps of the workflow) as black rectangles, and the Agents (i.e. tools, human actors and organisations) as orange pentagons. The graph's edge colours highlight the data flow (red and green) and the responsibility of the Agents (orange). Entities are laid out according to the ordering of their generation; the edges' arrows point "back into the past".[28] The three steps of the flooding risk areas calculation workflow introduced in Sub-section 4.2 are described by the four activities in the PROV graph.

Figure 3 Provenance diagram of the simulation of Brumadinho dam breaking from SRTM 30m workflow available on DataHub. Entities are depicted as yellow ovals, Activities as black rectangles, and Agents as orange pentagons



The hydrographic data sets generation phase involved two subsequent activities. The task *def:HANDCalculation1*, operated by agent *HANDscientist1*, calculates a vector map of contour lines of potential flooding areas from a raster file of the study basin. It relies on *def:HANDTool1* which ingests three data sets, i.e. a DEM (def:DEM1), a drainage network (def:RasterDN1) and a local drainage directions grid (def:LDD1) to produce a raster map in TIF format (def:HANDGrid1). From this map, a vector transformation is yielded by the *:HANDvectorCalculation1* task, based on a water altimetry value (*:waterAltimetry1*), that produces a vector (shape) file *def:HANDFloodDelimitationVector1* stored at INPE.[29] HAND calculation is affected by three parameters which may influence its reuse: the DEM resolution, the density of channels in the drainage network (provided by a stream initiation threshold value) and the different water altimetry tracks which determine the areas with a higher potential flood. For the sake of readability we do not report this information in the graph of Figure 3 but, to make HAND results effectively reusable, we detailed such information along with all other provenance information at https://old.datahub.io/dataset/brumadinho-dam-breaking.

The Sentinel Image Classification phase is carried out by the activity *def:SentinelClassification1*, operated by the agent *HANDscientist1* through the *def:ClassifierTool1* tool (i.e. the Isoseg classifier), that executes the classification of the input Sentinel of 10 m resolution (*def:SentinelImage1*). It produces the image of the study area (*def:SentinelClassifiedImage1*) classified by a segmentation process. The segmentation used the region growing method, which is based on a a similarity value (*def:ClassSimilarity1*) and a threshold (*def:ClassThreshold1*) to filter the minimum area.

The last workflow step integrating the hydrographic data sets and the Sentinel data classification is carried out by the activity *def:FloodArea_SentinelClassification_Integration*, generated by the agent *HANDscientist1* with the tool *def:SPRINGTool1*. It executes the integration of the hand vector isolines representing the (potential) flooding areas (*def:HANDFloodDelimitationVector1*) with the Sentinel classification map (*def:SentinelClassifiedImage1*). The output *def:BrumadinhoDamBreaking1* provides the analysis of the impact of the Brumadinho waste-lake dam breaking on the surrounding area from a SRTM 30 m.

6 Conclusion and future works

A critical usability issue of Geospatial Open Data concerns the lack of availability of data sets' provenance metadata, which currently includes only limited information referred mainly to data organisations and publishers. The paper presents a methodology to foster data sets' reuse by providing the metadata of the data sets' generating workflow. We exemplify this practice by providing the documentation for expressing the lineage of data sets produced by the qualitative analysis of flooding risk areas, studied to cope with Brazilian waste-lake dams breakdowns. Our contribution aims to make such processes intelligible and reproducible in other contexts of use. For example, it may be applied to replicate the study for the other numerous dams mineral waste-lakes in Brazil. We discuss the adoption of standards and W3C guidelines to represent provenance metadata and the publishing of hydrographic data sets according to OGD practices and their delivery through the OD management platform datahab.io. From our study we address the following recommendations to OD portal providers: (i) to adopt a common standard compliance approach to data description (i.e. mapping the metadata to standard DCAT), (ii) to provide proper documentation of the published open data considering at least a predefined list of options (e.g. file format, licence descriptions, provider) and (iii) to enrich metadata information with provenance metadata. By doing so, the portal may guarantee a quality level compatible with standards which, in return, ensure that hydrographic data can be used more effectively and thereby represent a significant step towards the prevention or the recovery from environmental catastrophes.

6.1 Implications of the study

To make provenance documentation easily accessible to data sets consumers, we have published the full PROV documents (Dam breaking provenance document RDF and the Diagram of workflow provenance) on datahub.io through the dereferenceable URI "Anonymised for double reviewing", along with other metadata and the data sets resources. These metadata provide the users interested in examining the effects of the waste-lake dam rupture with a precise indication of the process carried out together with all the detailed information on the parameters and data used. Some of these data are closely linked to the area affected by the catastrophic event. However, their disclosure through documentation is a concrete example of most of the salient procedural and decisional aspects to consider when replicating the same type of analysis in similar scenarios, with data relating to other regions and not limited to the study of mineral waste basins but for example to the case of artificial dams.

6.2 Future work

As for future work, we point out new strategies for linking geodata produced by INPE and provenance metadata generated in the run-time, i.e. exploring the possibilities to configure within Terrahidro tool the process of capturing provenance, and to execute the provenance graph reproducing the computation it represents.

Acknowledgement

The paper is a revised paper of the version presented in MTSR2019 (Garoufallou et al., 2019). We thank the Editors.

References

Albertoni, R., De Martino, M. and Quarati, A. (2018) 'Documenting context-based quality assessment of controlled vocabularies', *IEEE Transactions on Emerging Topics in Computing*, Vol. 1, pp.1–1.

Albertoni, R., De Martino, M., Di Franco, S., De Santis, V. and Plini, P. (2014) 'EARTh: an environmental application reference thesaurus in the linked open data cloud', *Semantic Web*, IOS Press, Vol. 5, No. 2, pp.165–171.

Albertoni, R., De Martino, M., Podestà, P., Abecker, A., Wàssner, R. and Schnitter, K. (2018) 'LusTRE: a framework of linked environmental thesauri for metadata management', *Earth Science Informatics Journal*, Vol. 11, No. 4, pp.525–544.

Alemu, G. and Garoufallou, E. (2020) 'The future of interlinked, interoperable and scalable metadata', *International Journal of Metadata, Semantics and Ontologies*, Vol. 14, No. 2, pp.81–87.

Armada, C. (2019) 'The environmental disasters of Mariana and Brumadinho and the Brazilian social environmental law state', *SSRN Electronic Journal*, pp.1–15. Doi: 10.2139/ssrn.3442624

Beno, M., Figl, K., Umbrich, J. and Polleres, A. (2017) 'Perception of key barriers in using and publishing open data', *JeDEM – EJournal of EDemocracy and Open Government*, Vol. 9, No. 2, pp.134–165.

Coetzee, S., Ivanova, I., Mitasova, H. and Brovelli, M.A. (2020) 'Open geospatial software and data: a review of the current state and a perspective into the future', *ISPRS International Journal of Geo-Information*, Vol. 9, pp.90.

De Martino, M., Rosim, S. and Quarati, A. (2019) 'Hydrographic datasets in open government data portals: mitigation of reusability issues through provenance documentation', in Garoufallou, E., Fallucchi, F. and William De Luca, E. (Eds): *Communications in Computer and Information Science Metadata and Semantic Research (MTSR'19)*, Springer, Cham. Doi: 78-3-030-36599-8_27.

Fernandes, G.W. et al. (2016) 'Deep into the mud: ecological and socio-economic impacts of the dam breach in Mariana, Brazil', *Natureza and Conservação*, Vol. 14, No. 2, pp.35–45.

Freire, J., Koop, D., Santos, E. and Silva, C.T. (2008) 'Provenance for computational tasks: a survey', *Computing in Science and Engineering*, Vol. 10, No. 3, pp.11–21.

Garoufallou, E., Fallucchi, F. and De Luca, E.W. (eds) (2019) 'Metadata and semantic research', *Proceedings of the 13th International Conference on Communications in Computer and Information Science (MTSR'19)*, Rome, Italy. Doi: 10.1007/978-3-030-36599-8.

Geiger, A.C.P. and Von Lucke, J. (2019) '*Open Data could have Helped us Learn From another Mining Dam Disaster*', *Scientific Data*, Vol. 6, pp.1–2.

Geiger, AC.P. and Von Lucke, J. (2012) '*Open government and (linked) (open) (government) (data)*', *JeDEM-eJournal of eDemocracy and Open Government*, Vol. 4, No. 2, pp.265–278.

Gil, Y., Deelman, E., Ellisman, M.H., Fahringer, T., Fox, G.C., Gannon, D., Goble, C., Livny, M., Moreau, L. and Myers, J. (2007) 'Examining the challenges of scientific workflows', *IEEE Computer*, Vol. 40, No. 12, pp.24–32.

Hartig, O. and Zhao, J. (2010) 'Publishing and consuming provenance metadata on the web of linked data', *Provenance and Annotation of Data and Processes – 3rd International Provenance and Annotation Workshop*, pp.78–90.

Janssen, M., Charalabidis, Y. and Zuiderwijk, A. (2012) 'Benefits, adoption barriers and myths of open data and open government', *Information Systems Management*, Vol. 29, No. 4, pp.258–268.

Johnson, PA, Sieber, RE, Scassa, T, Stephens, M. and Robinson, PJ. (2017) 'The cost(s) of geospatial open data', *Transactions in GIS*, Vol. 21, pp.434–445.

Lim, C., Lu, S., Chebotko, A. and Fotouhi, F. (2010) 'Prospective and retrospective provenance collection in scientific workflow environments', *IEEE International Conference on Services Computing*, IEEE Computer Society, pp.449–456.

Máchová, R. and Lnénicka, M. (2017) 'Evaluating the quality of open data portals on the national level', *Journal of Theoretical and Applied Electronic Commerce Research*, Vol. 12, No. 1, pp.21–41.

Marcelo, J.S.O., De Oliveira, H.R., Oliveira, L.A. and Lóscios, F. (2016) 'Open government data portals analysis: the Brazilian case', in Kim, Y. and Liu, M. (Eds): *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*, ACM, NY, USA, pp.415–424. Doi: 10.1145/2912160.2912163.

Miksa, T. and Rauber, A. (2017) 'Using ontologies for verification and validation of workflow-based experiments', *Journal of Web Semantics*, Vol. 43, pp.25–45.

Miles, A. and Bechhofer, S. (2009) *W3C Recommendation: Simple Knowledge Organization System Reference*. Available online at: http://www.w3.org/TR/skos-reference.

Missier, P., d Belhajjame, K. and Cheney, J. (2013) 'The W3C PROV family of specifications for modelling provenance metadata', *Proceedings of the 16th International Conference on Extending Database Technology*, pp.773–776.

Moreau, L. and Groth, P. (2013) 'Provenance: an introduction to PROV', *Synthesis Lectures on the Semantic Web: Theory and Technology*, Vol. 3, No. 4, pp.1–129.

Moreau, L., Groth, P., Cheney, J., Lebo, T. and Miles, S. (2015) 'The rationale of PROV', *Journal of Web Semantics*, Vol. 35, No. P4, pp.235–257.

Neumaier, S., Umbrich, J. and Polleres, A. (2016) 'Automated quality assessment of metadata across open data portals', *Journal of Data and Information Quality (JDIQ)*, Vol. 8, No. 2, pp.2–29.

Open Government Working Group (2007) *Eight principles of open government data. Open Government* Working Group. https://opengovdata.org/ (accessed on 15 September 2020).

Perego, A. (2017) *W3c dataset exchange working group (dxwg) use case working space: modeling data lineage [id12]*. Available online at: https: //www.w3.org/2017/dxwg/wiki/UseCaseWorkingSpacenID12

Perez, I.S., da Silva, R.F., Rynge, M., Deelman, E., Hernandez, M.S. and Corcho, O. (2017) 'Reproducibility of execution environments in computational science using semantics and clouds', *Future Generation Computer Systems*, Vol. 67, pp.354–367.

Quarati, A. and De Martino, M. (2019) 'Open government data usage: a brief overview', *IDEAS Proceedings of the 23rd International Database Engineering and Applications Symposium*, pp.229–236.

Quarati, A. and Rafiaghelli, J.E. (2020) 'Do researchers use open research data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case', *Journal of Information Science*, pp.1–40.

Quarati, A., Albertoni, R. and De Martino, M. (2017) 'Overall quality assessment of SKOS thesauri: an AHP-based approach', *Journal of Information Science*, Vol. 43, No. 6, pp.816–834.

Quarati, A., De Martino, M. and Rosim, S. (2021) 'Geospatial open data usage and metadata quality', *ISPRS International Journal of Geo-Information*, Vol. 10. Doi: 10.3390/ijgi10010030.

Ram, S. and Liu, J. (2007) 'Understanding the semantics of data provenance to support active conceptual modeling', Chen, P.P. and Wong, L.Y. (Eds): In book *Active Conceptual Modeling of Learning*, *Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling*, Springer, Berlin, Heidelberg, pp.17–29.

Rennfio, C.D., Nobre, A.D., Cuartas, L.A., Soares, J.V., Hodnett, M.G., Tomasella, J. and Waterloo, M. (2008) HAND, a new terrain descriptor using SRTM-DEM; mappingterra-firme rainforest environments in Amazonia', *Remote Sensing of Environment*, Vol. 112, pp.3469–3481.

Ribeiro, C. (2018) 'Promoting semantic annotation of research data by their creators: a use case with B2NOTE at the end of the RDM workflow', *Proceedings of the 11th International Conference on Metadata and Semantic Research (MTSR'17)*, Tallinn, Estonia.

Rosim, S., Namikawa, L.M., de Freitas Oliveira, J.R., De Martino, M. and Quarati, A. (2018) 'Workflow provenance metadata to enhance reuse of South America drainage datasets', *International Conference on eDemocracy and eGovernment*, pp.16–23.

Sadiq, S. and Indulska, M. (2017) 'Open Data: quality over quantity', *International Journal of Information Management*, Vol. 37, No. 3, pp.150–154.

Safarov, I., Meijer, A.J. and Grimmelikhuijsen, J. (2017) 'Utilization of open government data: a systematic literature review of types, conditions, effects and users', *Information Polity*, Vol. 22, pp.1–24.

Sahoo, S., Lebo, T. and McGuinness, D. (2013) *PROV-o: The PROV Ontology*. W3C, W3C Recommendation. Available online at: http://www.w3.org/TR/2013/REC-prov-o-20130430/

Umbrich, J., Neumaier, S. and Polleres, A. (2015) 'Quality assessment and evolution of open data portals', *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud*, Rome, pp.404–411. Doi: 10.1109/FiCloud.2015.82.

Van Der Aalst, W.M.P., Ter Hofstede, A.H.M., Kiepuszewski, B. and Barros, A.P. (2003) 'Workfow patterns', *Distributed and Parallel Databases*, Vol. 14, No. 1, pp.5–51.

Websites

1 www.go-fair.org/fair-principles/r1-2-metadata-associated-detailed-provenance/

2 www.w3.org/TR/prov-overview/

3 www.copernicus.eu

4 www.datahub.io

5 www.informea.org/en/legislation/law-no-12234-national-policy-dam-safety

6 http://www.inpe.br

7 scihub.copernicus.eu

8 www.w3.org/TR/dwbp

9 www.go-fair.org/fair-principles

10 www.w3.org/RDF

11 www.w3.org/TR/vocab-dcat-2

12 www.w3.org/TR/prov-o

13 www.opendatabarometer.org

14 www.opendatacharter.net/principles

15 www.g20.utoronto.ca/2015/G20-Anti-Corruption-Open-Data-Principles.pdf

16 http://linkeddata.ge.imati.cnr.it/

17 www.eionet.europa.eu/gemet

18 www.fao.org/agrovoc

19 www.w3.org/DesignIssues/LinkedData.html

20 www.opendefinition.org/licenses

21 www.data.gov

22 www.usa.gov/government-works

23 Anonymised for double reviewing

24 Anonymised for double reviewing

25 "Anonymised for double reviewing"

26 Anonymised for double reviewing"

27 https://openprovenance.org/store

28 www.w3.org/2011/prov/wiki/Diagrams

29 http://www.dpi.inpe.br/TerraHidro_Data/HAND/NASADEM_HAND_vector_2000_lin.shp