Open Geospatial Data,
Software and Standards

ORIGINAL ARTICLE

Open Access

Eliciting academic SDI requirements through a survey of user practices

Eduardo Silverio da Silva* and Silvana Philippi Camboim

Abstract

During coursework and research projects, several geospatial algorithms are produced and mentioned by authors in written documents. However, these products often remain unavailable after the end of the projects but could be reused by third parties, providing an improvement in spatial data infrastructure (SDI), reproducibility and open science. Because SDI relies on the sharing of geographic resources, this article focuses on the study of geospatial algorithms. There are studies concerning the use of academic spatial data infrastructure (ASDI) as a solution to academic resources, but these rarely comprise the publication of algorithms and are mainly aimed at improving systems through functional requirements without considering the requirements of academic users. This study was carried out with the purpose of supporting the sharing of algorithms in an ASDI (www.idea.ufpr.br) created at the Federal University of Paraná, Brazil. Thus, this study aims to characterize the behaviors of academic users regarding their use, storage, sharing and development of geospatial algorithms. For this purpose, a questionnaire was published and received 196 valid responses. The results showed that compared to other interviewees, academics use, develop and share fewer spatial algorithms and have more concerns about citations and fewer concerns about profit. However, these findings do not imply that these users are less productive but rather that their work is different and may or may not rely on the use of algorithms. Furthermore, the results showed more active sharing when authors work with their own algorithms, which could be due to increased security related to the license information, representing important information to be included in geoportals.

Keywords: Geospatial algorithms, Reproducibility, User practices, Sharing

Background

During coursework and research, many projects are developed within academia. Some of these projects involve the production of data and geospatial algorithms. In the context of this paper, geospatial algorithms can be understood as both scripts programmed to process spatial data locally and as a web processing services and tools that are already part of geoprocessing software.

Because the purpose of a spatial data infrastructure (SDI) is to promote the broad consumption of geographic resources [1, 2], this article focuses on the study of algorithms applied to spatial data so that they can interact with other resources in the system. These resources are usually mentioned in the written works by their authors but most often remain unavailable after the

end of the study because there is no publication requirement for most academic repositories [3, 4].

To be properly accessed on the internet, geospatial algorithms require specific tools that provide the main geographic information systems (GIS) operations: storage, recovery, search, visualization, analysis and the processing of spatial data [5]. These tools are not provided by current academic resources, which in the best cases, provide direct access to resources through a file server.

The search for the better management of geographic resources within the academic environment requires not only the publication of products but also making these products available in a standard and interoperable way to allow wide access to information from any platform and without loss of value in any aspect, which conforms to the principles of open science [6].

At first glance, the solution is to integrate the resources into an SDI. However, an academic context requires a user-centered approach (bottom-up development) and a

* Correspondence: eduardosilverio1990@gmail.com
Universidade Federal Do Paraná – UFPR, Avenida Coronel Francisco Heráclito dos Santos, 210 – Jardim das Américas – , Caixa Postal 19001, Curitiba, Paraná CEP 81.531-990, Brazil

study of needs [7] because the inclusion of resources should be voluntary to stimulate sharing and give rise to what we call academic spatial data infrastructure (ASDI) [4]. In the context of ASDI and this paper, a user is defined as any person who can consume the shared resources (end user) and any academic who can produce resources and contribute to the system (provider).

In the future, an ASDI could allow students, professors and researchers to share geospatial algorithms related to their written documents [4]. This tool could allow algorithms to be consumed and cited. In addition, other academics could download or execute these resources over the web for adaptation to their field of study. This tool could contribute to open science because it brings more transparency to published works and allows third parties to verify the reproducibility of geospatial algorithms. Several ASDI initiatives exist [3, 8–12]; however, there are few implementations concerning algorithms. In Brazil, at Federal University of Paraná (UFPR) there is an ASDI in development, called IDEA UFPR [13], which currently provide access and sharing of spatial data from research projects or academic studies. However, this platform is being designed to also promote the sharing and access of algorithms and related functionalities in the next/future versions.

The first reason for this gap is that in comparison to data, there is a greater complexity in cataloging these resources because they are often embedded in software, require the use of compilers, and rely on other code or data to be consumed. A second reason for this gap is related to the resistance to algorithm sharing, which, like resistance to data sharing, may be due to factors such as the need to adapt the code for general use and the need to provide technical support to the end user [14, 15].

Few studies present data about characteristics of users of spatial algorithms and even fewer studies exist regarding the characterization of academic user, so system implementation is usually guided by attendance to the functional requirements of systems [16]. Thus, the aim of this research is to describe the characteristics of academic user behaviors to support the implementation of ASDI by adopting a user-centered design approach.

Methodology

A questionnaire was designed to measure people's perceptions of their practices concerning geospatial algorithms. Some questions are about the algorithms of the respondents, which can be understood as algorithms programmed by the participants (e.g., a script to transform coordinates), and others are about algorithms produced by third parties, which can be understood as algorithms shared by other people or that are already part of software (e.g., a spatial join from any commercial software).

Therefore, the target participants of this study were people who had already had some contact with geographical information or any previous experience with geoprocessing.

Design of the Questionnaire

The questionnaire was divided into seven sections. The first two sections were a content statement and a personal information section, while the remaining sections concerned the following:

a) Set I - use of algorithms and geoprocessing software: designed to collect information about the frequency of use and reveal the most commonly used geoprocessing software and functions;

b) Set II - development of algorithms and software: designed to collect information about the frequency of development and the most commonly used languages and development environments;

c) Set III - storage and sharing of algorithms: designed to reveal the current storage location, if people are likely to share algorithms and how they usually share algorithms;

d) Set IV - difficulties in use and sharing: designed to discover whether people have any difficulty when using third-party algorithms and the main barriers concerning the sharing of algorithms;

e) Set V – page/portal for sharing algorithms: designed to collect information about user preferences related to the functionalities and management of an academic repository of geospatial algorithms.

Despite the large number of questions, the questions were designed to be short and objective [17] and always presented an answer option of "other". Although the data coming from this option are generally more difficult to analyze (more unexpected responses to read), this option allows the retrieval of responses that were previously unknown when designing the questionnaire.

Disclosure and validation

Before publication, a group of nine students from the Postgraduate Program in Geodetic Sciences was asked to answer the first version of the questionnaire. The activity was performed in person, allowing the interviewee to ask questions of the interviewer. As each person finished the activity, some questions regarding the general understanding and opinion of participants were asked:

a) What is your general opinion about the proposed questionnaire?

b) Did you have any difficulty in understanding any question? Which one and why?

c) Do you have any suggestions to improve the questionnaire?

After an analysis of the doubts, understanding of questions, opinions and suggestions of participants, some questions were reformulated, others were included, and repeated responses were added to the list of available options.

Furthermore, to obtain less subjective data, explanations were added to each response whenever a frequency question was requested. For instance, when asked about the frequency of use of algorithms, the responses provided were as follows:

a) Always: In more than 90% of works;
b) Frequently: Between 60% and 90% of works;
c) Sometimes: Between 40% and 60% of works;
d) Rarely: Between 10% and 40% of works;
e) Never: Less than 10% of works.

To reach a broader public, the questionnaire was translated from Portuguese to English and Spanish. The questionnaire was published online and propagated via sharing with personal contacts and several academic departments via email, through social media, via application groups and via the Open Source Geospatial Foundation (OSGEO) discussion forums whenever the forum subject was related to development of spatial algorithms.

After four months of accepting responses, the answers were collected for validation so that duplicated responses, responses that did not agree with the aforementioned content statement and responses from people who did not have any contact with geospatial data/algorithms were removed.

Definitions and analysis
First, to achieve a description of the academic user, some criteria for defining an academic were adopted. In this research, people whose current job was student (undergraduate or postgraduate) or professor (anyone in a teaching position) were considered academics. People whose current job was not related to those roles were considered professionals. Nonetheless, some participants whose current job was student/professor and another unrelated job were considered both academics and professionals.

Second, the responses were divided into the group of "academics" and the group "professionals". Hence, the responses were used to construct graphics in spreadsheets that were analyzed and compared between both groups with the goal of finding specific patterns that could differentiate academics from professionals in the context of the consumption, storage, sharing and production of geospatial algorithms.

Results and discussion
A total of 196 valid responses to the questionnaires were received (Fig. 1). The current jobs of the participants were diverse, and most responses came from Brazil because this was where the questionnaire could be sent to people and institutions known by the authors.

The responses were divided into a group of "professionals" with 79 people and a group of "academics" with 118 people. The purpose of this division was to identify the specific characteristics of academic behavior through a comparative analysis between the groups.

A limitation of this approach is that most academics were Brazilian (79.7%), which could have biased the results, creating a regional scenario. For instance, there is specific software for the transformation of coordinates between reference systems that are only suitable for Brazil, and thus, some responses could have been inflated.

However, the "other" response option in all questions may have mitigated this problem by allowing people from other countries to provide responses indicating different software that is more suitable for their locality.

In addition, ASDIs are relatively little used in Brazil [3], and research funding institutions in the country do not make specific recommendations on the need to publish geospatial information or algorithms.
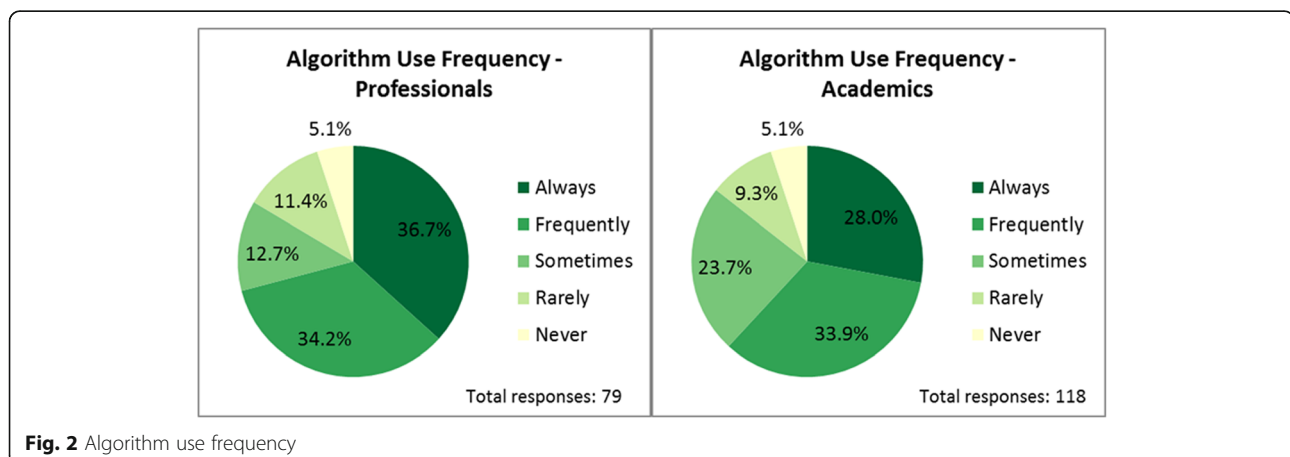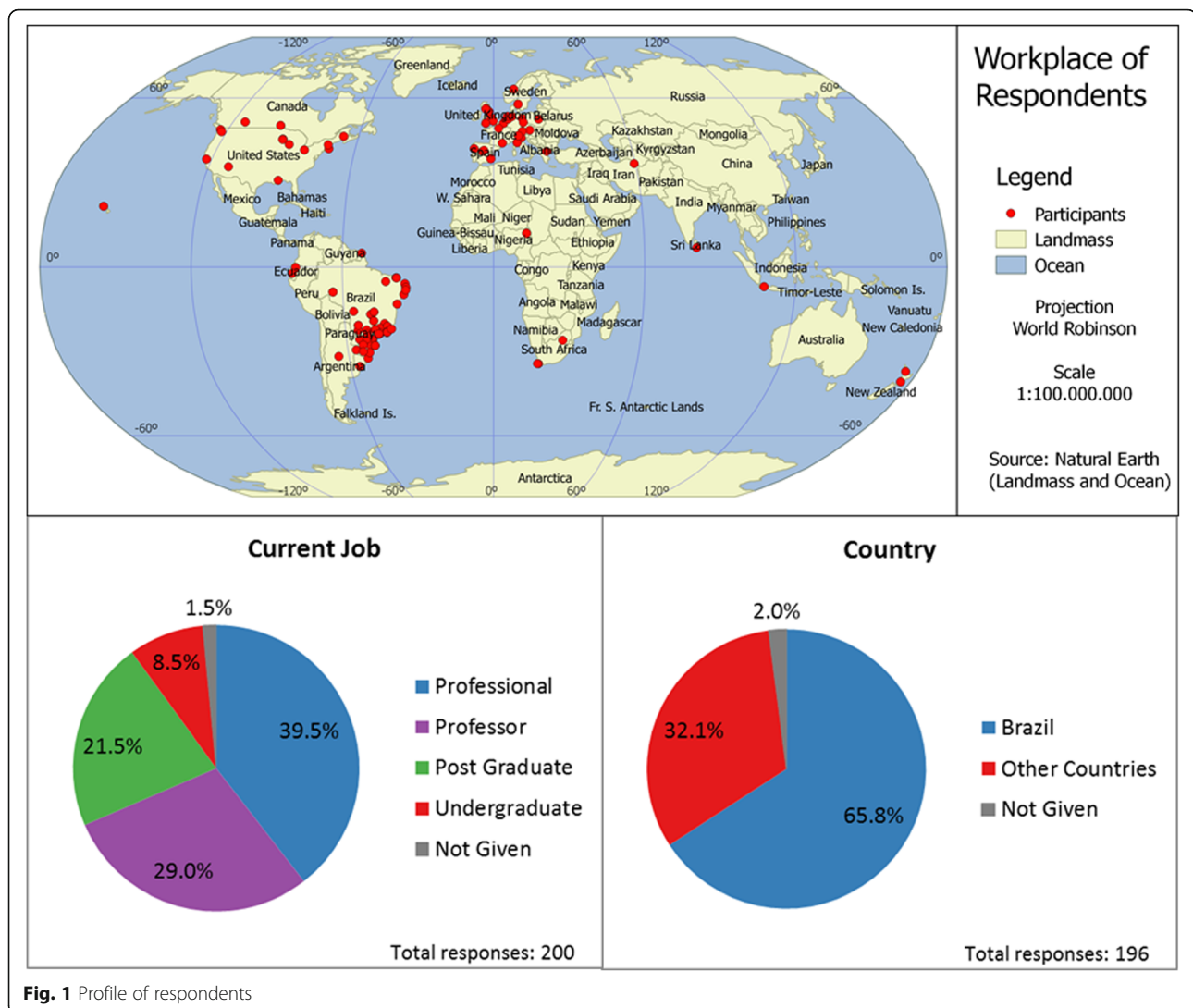
Practices of use
The group of professionals presented a higher frequency of use of geospatial algorithms (Fig. 2). The subsets of software more used among the group of professionals were computer aided design (CAD) and database management systems. However, academics were found to use geodesy and digital image processing (DPI) software more often.
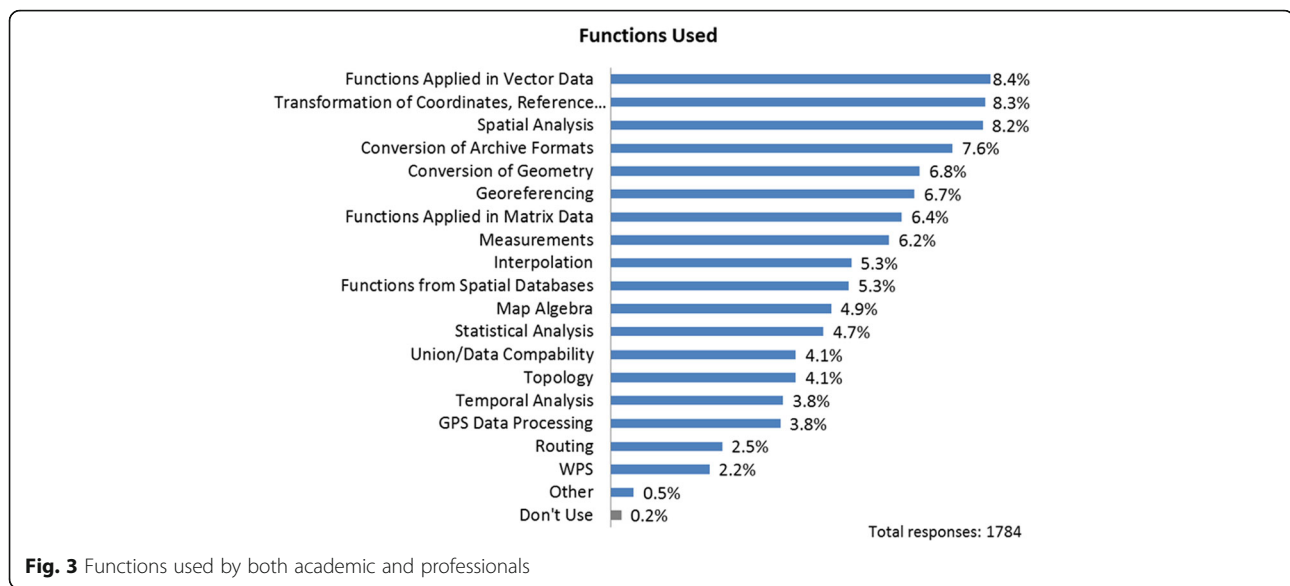
Both groups presented a diverse set of uses of geospatial algorithms (Fig. 3). These categories were listed from common functions used in GIS, although their subdivision may have influenced the respondents. In addition, there were low use and knowledge of web processing service (WPS) in both subsets.

The results indicate a low adoption of WPSs in day-to-day activities and that functions regarding vector data and the conversion and transformation of coordinates/reference systems should be prioritized for the broad use of an SDI. However, in an ASDI, more functions regarding geodesy and DPI software should be implemented.

Practices of development
A higher development rate was found in the professionals group (Fig. 4); this group also showed a greater use of programming languages, web map building tools and development environments. Both academics and professionals indicated high rates of Python and Structured Query Language (SQL) use as languages used in their professional activities (Fig. 5). A greater use of Freemat/Matlab as development environments was observed for academics (Fig. 6).

Fig. 1 Profile of respondents

Fig. 2 Algorithm use frequency

Fig. 3 Functions used by both academic and professionals

The differences in the frequencies of development could be due to the different type of work performed by people inside academia, where there are broad areas of study, that do not necessarily require the development of algorithms. In a company, it might be more common to use algorithms to optimize processes for better production, avoiding wasting time and, hence, increasing profit.
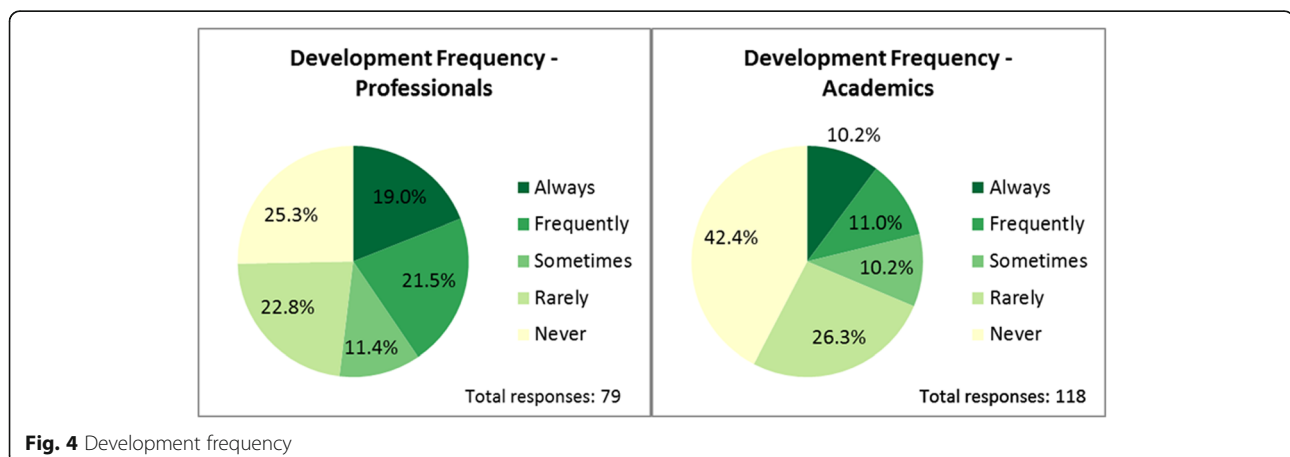
These results show that Python and SQL should be present in an SDI implementation, so alternatives such as the Python Web Processing Service (PyWPS) could be considered. However, the use of SQL may have been inflated because most respondents considered queries to databases as algorithms, which, in a pragmatic view, could be considered just a necessary step for using a database. On the other hand, in an ASDI, the Freemat/Matlab environment should be enabled for processing through the web.

Practices of storage and sharing

The responses for this set of questions showed that the main place where algorithms are stored is in local file systems. Additionally, there is a higher sharing rate by professionals, and this groups shows more sharing of third-party algorithms (Fig. 7). There were few mentions of WPS as a method of sharing, and for professionals, there is a high use of repositories, while academics often use shared folders and email as a methods for sharing.

The differences between the two groups might be explained by the previously verified low use and development rates of algorithms by academics, who thus have fewer algorithms to share. However, the increase in sharing of third-party code could be due a higher number of algorithms now in people's possession.

In addition, when comparing the percentage of people who always share algorithms (i.e., share to everyone
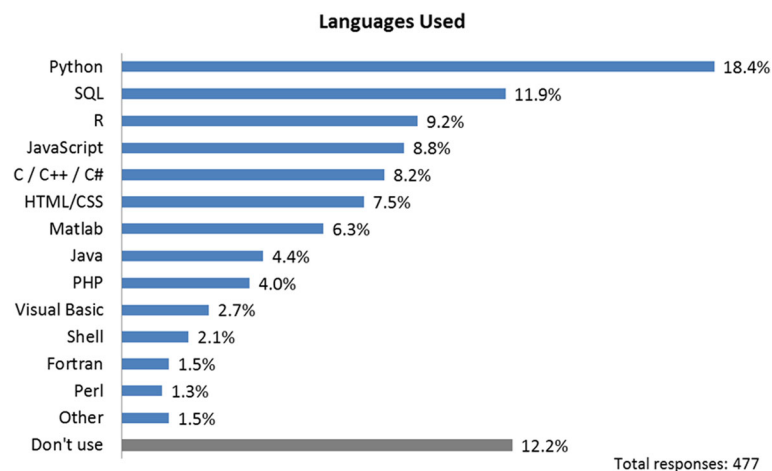

Fig. 4 Development frequency

Languages Used

Language	%
Python	18.4%
SQL	11.9%
R	9.2%
JavaScript	8.8%
C / C++ / C#	8.2%
HTML/CSS	7.5%
Matlab	6.3%
Java	4.4%
PHP	4.0%
Visual Basic	2.7%
Shell	2.1%
Fortran	1.5%
Perl	1.3%
Other	1.5%
Don't use	12.2%

Total responses: 477

Fig. 5 Languages used by both groups

through webpages), in the case of third-party code, there is a decrease in sharing. This finding could be related to the insecurity about algorithm licensing use, which reinforces the necessity of metadata in published resources.

Difficulties regarding finding, using and sharing algorithms

The responses to this set of questions showed that people still have some difficulties in finding the algorithms needed for their work (Fig. 8) because more than half of participants succeeded only sometimes (between 40% and 60% of times) in their search for algorithms. The main places where these searches occur are internet pages and plugin repositories, showing that WPSs and companies/institutions are unusual places for this task (Fig. 9).

The main barrier to the use of third-party algorithms identified by the participants was that codes are not adaptable to other fields of study. Furthermore, for the question regarding the difficulties of sharing, the most cited difficult was the necessity of providing technical support to users, and when dealing with third-party code, the identified difficulties were related to a lack of information and the need to contact the code's authors.

However, when dealing with self-produced algorithms, professionals and academics diverge on some points (Fig. 10). For academics, there is a high need to be cited, while for professionals, this need is small, but there is a greater need to generate profit with the code.

In addition, there are fewer academics claiming any difficulty when sharing, and this could be explained by the fact that there are fewer algorithms in their possession
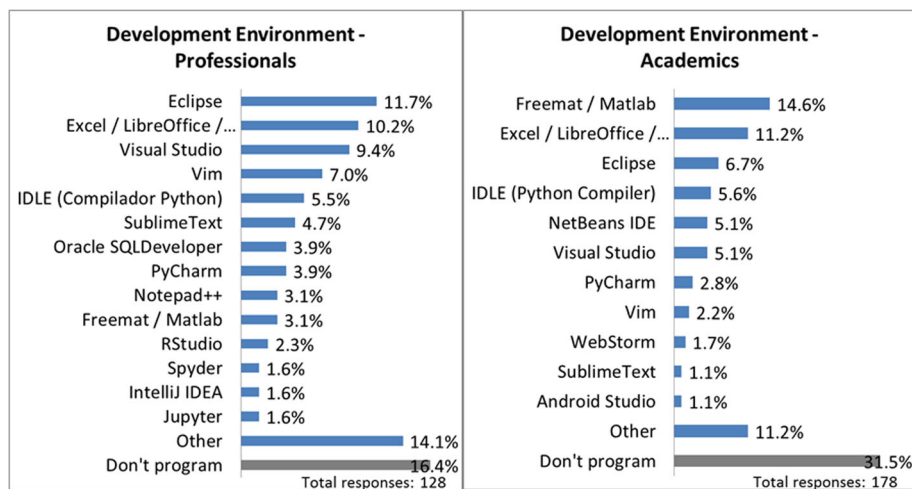
Development Environment - Professionals

Environment	%
Eclipse	11.7%
Excel / LibreOffice /…	10.2%
Visual Studio	9.4%
Vim	7.0%
IDLE (Compilador Python)	5.5%
SublimeText	4.7%
Oracle SQLDeveloper	3.9%
PyCharm	3.9%
Notepad++	3.1%
Freemat / Matlab	3.1%
RStudio	2.3%
Spyder	1.6%
IntelliJ IDEA	1.6%
Jupyter	1.6%
Other	14.1%
Don't program	16.4%

Total responses: 128

Development Environment - Academics

Environment	%
Freemat / Matlab	14.6%
Excel / LibreOffice /…	11.2%
Eclipse	6.7%
IDLE (Python Compiler)	5.6%
NetBeans IDE	5.1%
Visual Studio	5.1%
PyCharm	2.8%
Vim	2.2%
WebStorm	1.7%
SublimeText	1.1%
Android Studio	1.1%
Other	11.2%
Don't program	31.5%

Total responses: 178
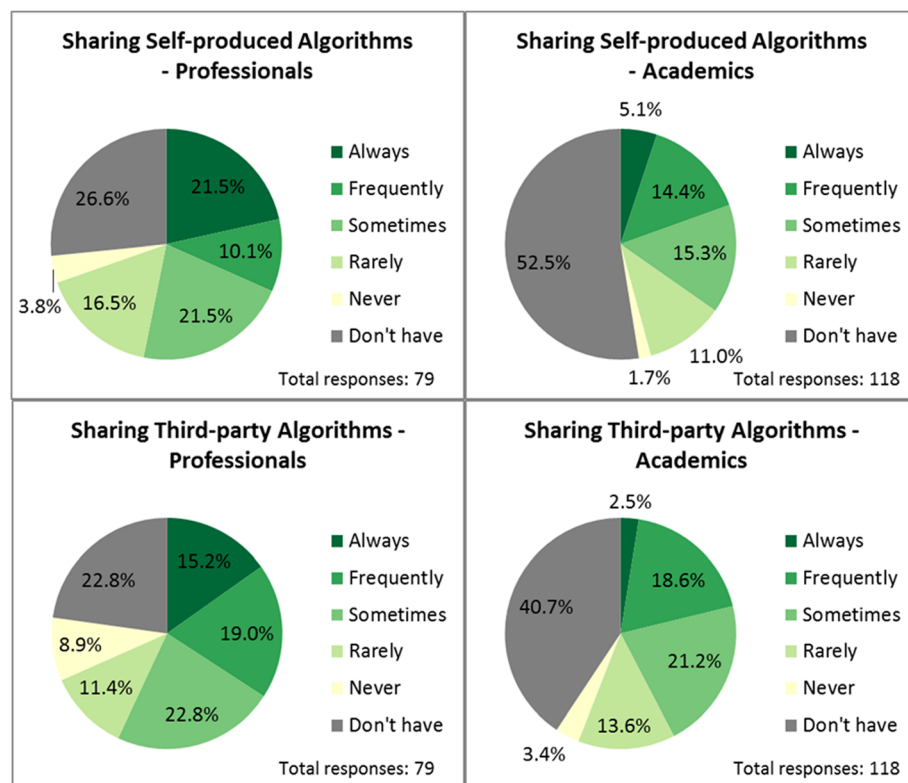
Fig. 6 Development environment

Fig. 7 Sharing self-produced and third-party algorithms

and thus a lower number of people who see problems in sharing is natural.

In an ASDI context, tools related to citing authors should be included, and in a general SDI context, some ways to provide technical support and allow the contacting of authors should be identified.

User views of sharing platforms

This set of questions returned similar results for both groups. From the perspective of the users, the more important functionalities provided by a specific platform for sharing algorithms are download, upload, purpose description and algorithm search (Fig. 11). Furthermore, for people both inside and outside academia, a platform wherein the authors can publish their own code in a decentralized system is the ideal scenario.

Software description ranked higher than hardware description. This result could be because advances in hardware have solved the main necessities required by algorithms (storage capacity, velocity and memory). Hence,
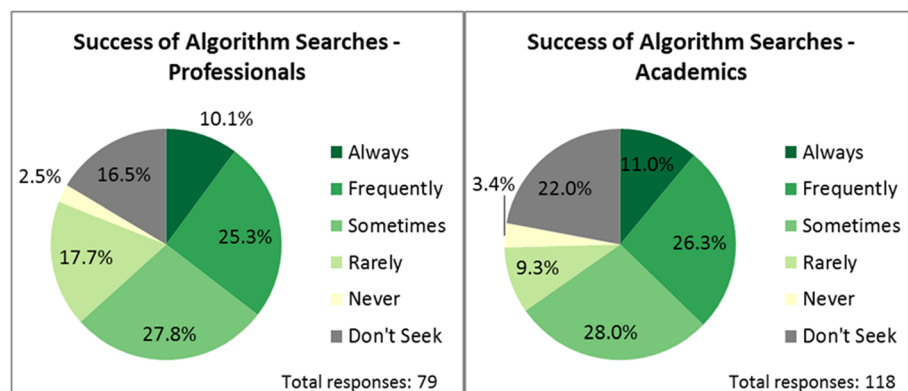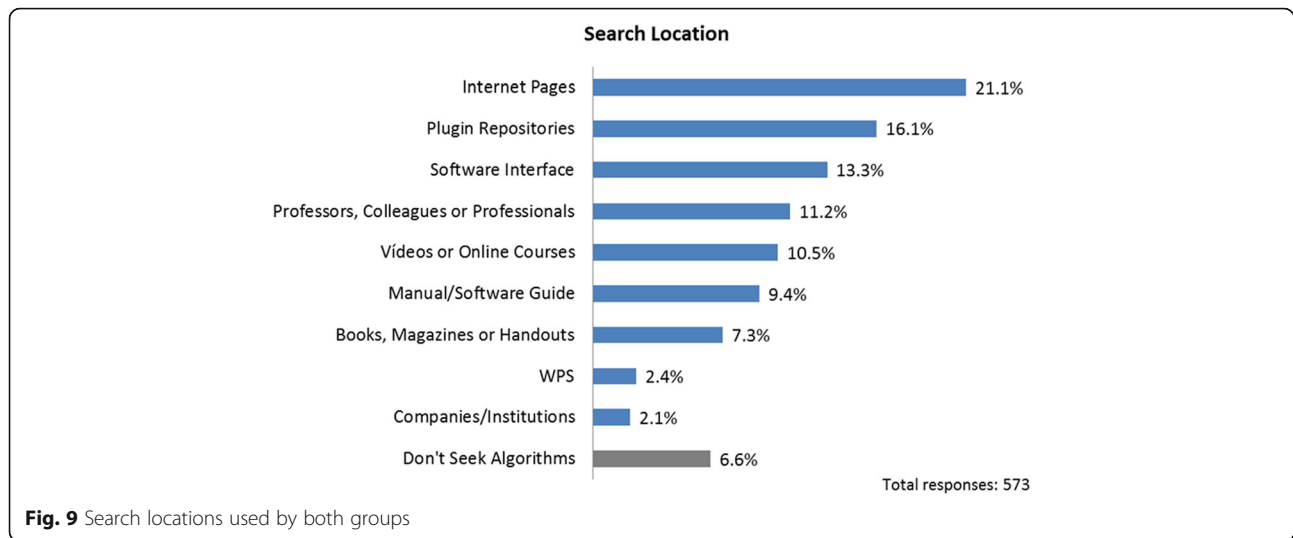

Fig. 8 Success of algorithms searches

**Fig. 9** Search locations used by both groups

users do not have the same level of concern for this factor as for software requirements.

Conclusions

The development of ASDIs through the availability of geospatial algorithms represents another step towards research reproducibility and open science because this action brings increased transparency to academia and allows resource verification by third parties.

Because the aim of this study was to elicit some requirements for ASDIs through a user-centered approach (bottom-up development), we have gathered information from possible users of this system, that is, people who have already worked with geospatial data and who are studying this subject.

By surveying user practices and analyzing the differences between academic and professional responses, we can infer that in an ASDI, there are recommendations
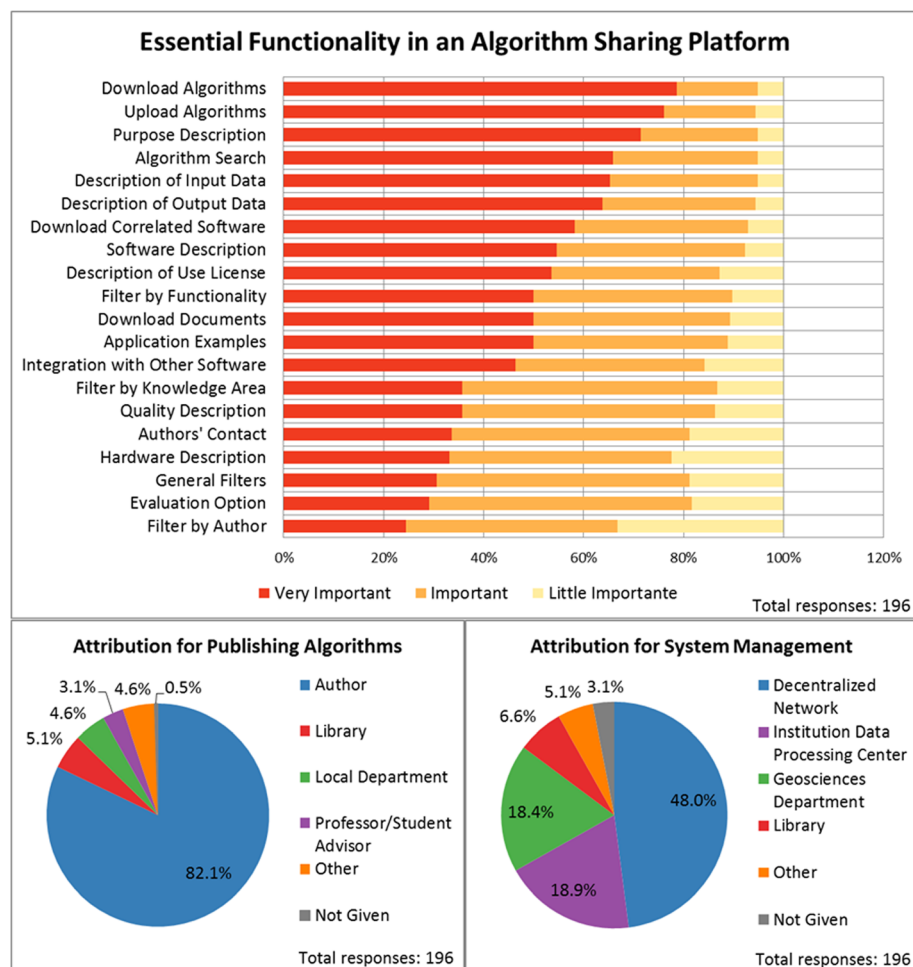

Fig. 10 Barriers to sharing self-produced algorithms
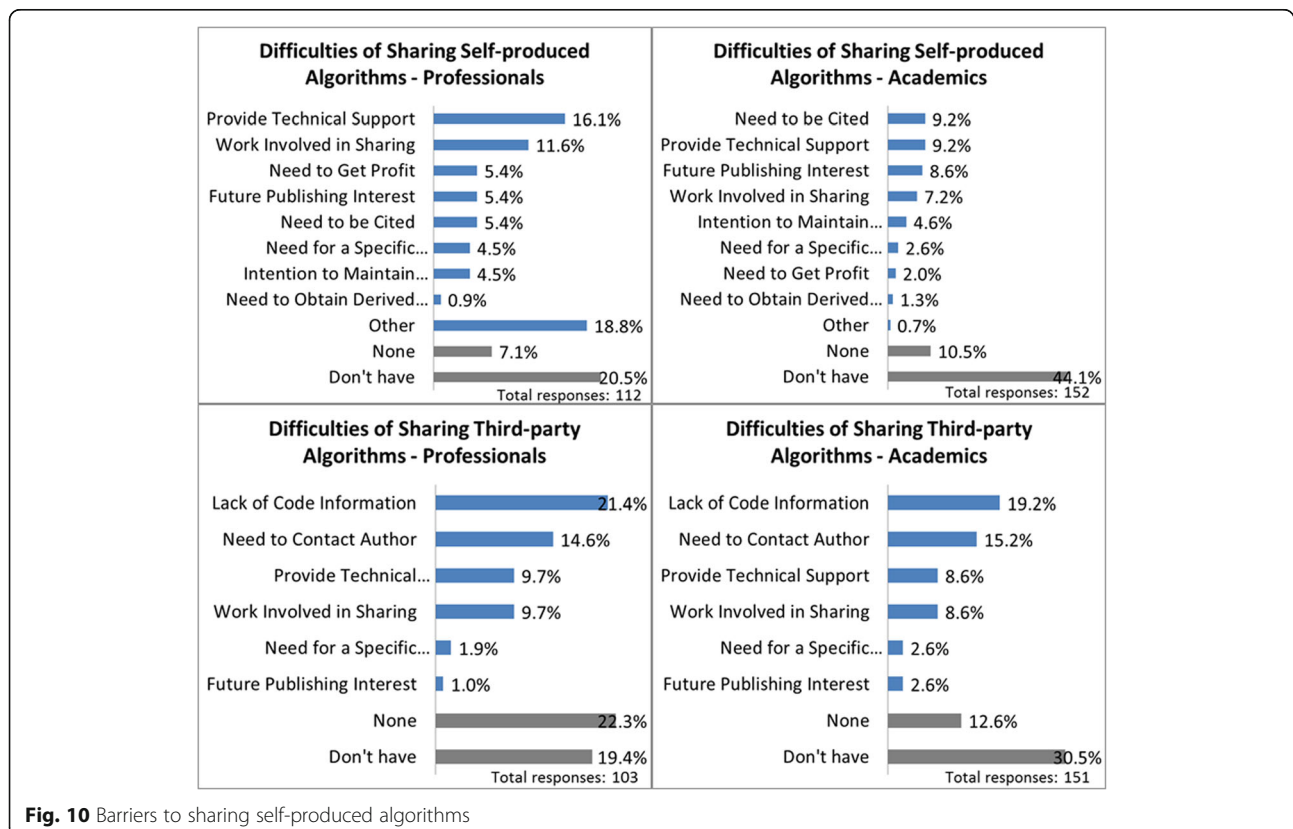
Fig. 11 Functionalities and attribution for publishing and system management

that apply for all types of users (including those who could also use this system) and other recommendations that are more suitable for academic users. These recommendations are presented below.

There is a consensus among all users that an ASDI should allow authors to publish their own resources and that the system management should be decentralized (user-center approach). Users should have the option to download, upload, read algorithm purpose descriptions and make algorithm searches. For an ASDI, it is also important to provide tools for citing authors because most geospatial algorithms are part of a research project.

An ASDI should provide algorithms applied to vector data, conversion tools and the transformation of coordinates and reference systems because these tools are used by the majority of both types of users. Additionally, it is important to include licensing metadata in the resources because this step could foster more security for users who share third-party algorithms (by providing information on how algorithms can be shared or modified).

Although universities generally lack specific staff to work on an ASDI, alternatives to providing technical support to the end user should be identified. The contact information of the algorithm authors can help the end user obtain information about its use, encouraging collaborative development and a mutual support community, since the lack of support was identified in the study as a significant barrier.

Although WPSs are unknown to most users, alternatives such as PyWPS could be tried because Python was the most popular language identified by the participants. For non-geospatial algorithms, there are alternative online repositories, such as Github or Bitbucket, which provide options to create projects online and versioning tools. However, for the case of geospatial algorithms shared via those repositories, the possibilities of executing these algorithms on geospatial data online and of searching for geospatial algorithms geographically by defining a region of interest on the map where the algorithm was applied requires further special developments to become possible.

In future works, researching alternatives to improve existing repositories with functionalities provided by ASDIs and finding alternatives for integrating both solutions are recommended.

Appendix

Questionnaire

The document containing the entire questionnaire cited in this research is available on Figshare and can be accessed and downloaded via this link: https://doi.org/10.6084/m9.figshare.6143105.v1

Abbreviations

ASDI: Academic Spatial Data Infrastructure; CAD: Computer Aided Design; DPI: Digital Image Processing; GIS: Geographic Information Systems; OSGEO: Open Source Geospatial Foundation; PyWPS: Python Web Processing Service; SDI: Spatial Data Infrastructure; SQL: Structured Query Language; WPS: Web Processing Service

Availability of data and materials

The datasets supporting the conclusions of this article are available in the figshare repository, in https://doi.org/10.6084/m9.figshare.5923762 and https://doi.org/10.6084/m9.figshare.6143105.v1.

Authors' contributions

ESS: design, disclosure and analysis of questionnaire; article general writing and formatting; design of images/map. SPC: references; improvements to the design and disclosure of questionnaire; writing part of introduction and conclusion; improvement of English and images/map; Edition of the manuscript. Both authors read and approved the final manuscript.

Competing interests

None of the authors have any financial or non-financial competing interests in this manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Bernard L, Craglia M. SDI - from spatial data infrastructure to service driven infrastructure. In: Research workshop on cross-learning between spatial data infrastructures and information infrastructure. Enschede; 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.456.160&rep=rep1&type=pdf. Accessed 4 July 2018.
2. Williamson I, Rajabifard A, Binns A. Challenges and issues for SDI development. International Journal of Spatial Data Infrastructures Research. 2006;1:24–35.
3. Brito Pl, Souza Fa, Camboim S, Giannotti Ma Primeiros Passos Para A Implementação De Uma Ide Universitária. Paper presented at V Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, Universidade Federal de Pernambuco, Recife, 12–14 2014.
4. Coetzee S, Steiniger S, Kobben B, Iwaniak A, Kaczmarek I, Rapant P, Cooper A, F-J B, Schoof G, Katumba S, Vatseva R, Sinvula K, Moellering H. The academic SDI—towards understanding spatial data infrastructures for research and education. In: Peterson M, editor. Advances in cartography and GIScience, vol. 1E. Washington: Springer; 2017. p. 99.
5. Maguire D. An overview and definition of GIS. In: Maguire D, Goodchild M, Rhind D, editors. Geographical information systems: principles and applications, vol. 2E. Minneapolis: Longman Scientific and Technical; 1991. p. 9.
6. European Comission. Open Innovation, Open Science, Open to the World - A Vision for Europe. Luxembourg: Publications Office of the European Union; 2016.
7. Harvey F, Iwaniak A, Coetzee S, Cooper AK. SDI past, present and future: a review and status assessment. In: Rajabifard A, Coleman D, editors. Spatially enabling government, industry and citizens - research and development perspectives: Needham: GSDI Association Press; 2012.
8. Burton A, Groenewegen D, Love C, Treloar A, Wilkinson R. Making research data available in Australia. IEEE Intell Syst. 2012;27:40–3. https://doi.org/10.1109/MIS.2012.57.
9. Fronza G. IDE Acadêmica: Construçao de uma Infraestrutura de Dados Espaciais Colaborativa. Dissertation: Universidade Federal do Paraná; 2016.
10. Hill E, Trimble L. Scholars GeoPortal: a new platform for geospatial data discovery, exploration and access in Ontario universities. International Association for Social Science Information Services and Technology. 2012;36:6–15.
11. Machado AA. IDE Acadêmica em Universidades Brasileiras Proposta para a Universidade Federal do Paraná (UFPR). Dissertation: Universidade Federal do Paraná; 2016.
12. Machado Aa, Silva Es, Fronza G, Campos Rg, Ferri Kc, Pisetta Ja, Camboim Sp (2016) PROJETO E IMPLEMENTAÇÃO DE UMA IDE ACADÊMICA NA UFPR. Paper presented at the IX Colóquio Brasileiro de Ciências Geodésicas, Universidade Federal do Paraná, Curitiba, 5–6 may 2016.
13. UFPR. IDEA UFPR. http://www.idea.ufpr.br. Accessed 17 Jun 2018.
14. SM E. Open code for open science? Nat Geosci. 2014;7:779–81.
15. Stodden V. Reproducible research - addressing the need for data and code sharing in computational science. Computing in Science & Engineering. 2010;12:8–12.
16. Iosifescu-Enescu I, Matthys C, Gkonos C, Iosifescu-Enescu C, Hurni L. Cloud-based architectures for auto-scalable web Geoportals towards the Cloudification of the GeoVITe Swiss academic Geoportal. ISPRS International Journal of Geo-Information. 6:192. https://doi.org/10.3390/ijgi6070192.
17. Boynton P M, Greenhalgh T. Selecting, designing, and developing your questionnaire. BMJ: British Medical Journal. 2004;328(7451):1312–5.