Connecting web-based mapping services with scientific data repositories: collaborative curation and retrieval of simulation data via a geospatial interface

Christian T. Jacobs  
Faculty of Engineering and the Environment, University of Southampton, United Kingdom

Alexandros Avdis  
Department of Earth Science and Engineering, Imperial College London, United Kingdom

Abstract

Increasing quantities of scientific data are becoming readily accessible via online repositories such as those provided by Figshare and Zenodo. Geoscientific simulations in particular generate large quantities of data, with several research groups studying many, often overlapping areas of the world. When studying a particular area, being able to keep track of one’s own simulations as well as those of collaborators can be challenging. This paper describes the design, implementation, and evaluation of a new tool for visually cataloguing and retrieving data associated with a given geographical location through a web-based Google Maps interface. Each data repository is pin-pointed on the map with a marker based on the geographical location that the dataset corresponds to. By clicking on the markers, users can quickly inspect the metadata of the repositories and download the associated data files. The crux of the approach lies in the ability to easily query and retrieve data from multiple sources via a common interface. While many advances are being made in terms of scientific data repositories, the development of this new tool has uncovered several issues and limitations of the current state-of-the-art which are discussed herein, along with some ideas for the future.

Submitted to the International Journal of Digital Curation on 15 May 2016

Correspondence should be addressed to Christian T. Jacobs, Aerodynamics and Flight Mechanics Research Group, Building 13 (Tizard), University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom. Email: C.T.Jacobs@soton.ac.uk

Copyright rests with the authors. This work is released under a Creative Commons Attribution 4.0 International Licence. For details please see http://creativecommons.org/licenses/by/4.0/

International Journal of Digital Curation  
20xx, Vol. 0, Iss. 0, 1–10.

http://dx.doi.org/10.2218/ijdc.v0i0.0  
DOI: 10.2218/ijdc.v0i0.0
Introduction

Advances in science are made by building on existing results and being able to reuse data from one or more sources. Growing quantities of scientific data are being made readily accessible to the public, or at least amongst teams of collaborators, through the Internet as a result of technological advances in digital curation, increased awareness of open data and reproducibility best practices, and changes in funding policies (Alsheikh-Ali et al., 2011; LeVeque et al., 2012; Stodden et al., 2013; Vines et al., 2013; McKiernan et al., Submitted). This is quite a change to the status quo in which data is normally kept solely on the researcher’s local computer, with only the key results being published in journal articles and conference proceedings (Whitlock et al., 2010; Kraker et al., 2011). The recent growth of online data repository services such as Figshare\(^1\) and Zenodo\(^2\) has certainly accelerated this trend through their ease-of-use, integration with University data management plans (Hahnel, 2014), and free/affordable storage space available to researchers.

Studies in the field of computational science often focus on simulating flow phenomena within particular geographical areas. Examples include the modelling of flow past marine power turbines in the Pentland Firth (Funke et al., 2014), traffic flow and pollutant dispersion patterns in Antwerp (Mensink et al., 2008), and inundation studies considering the Pacific coast of Japan following a tsunami event (Mori et al., 2011). Often an individual researcher will perform hundreds of simulations involving their desired area of study. While some of these simulations may not be worthy of being made available (for example, if a simulation were to fail, or the parameters were incorrect and the simulation had to be re-run), it is important to keep track of (i.e. ‘catalogue’) the ones that are. However, searching and retrieving the available datasets for a desired area can be non-trivial, particularly when the datasets are stored across multiple different hosting services and curated by different researchers. In addition, some researchers and their employers may wish to impose limited access to their data repositories, or even store the data on their own private servers because of sensitivity or commercial reasons. In such cases, having a means of bringing together different services and data sources, and presenting them in a user-friendly manner, is crucial for facilitating the re-use and exploitation of relevant and available data to the fullest extent possible. Furthermore, even when a particular area such as the Gulf of Mexico is studied, some studies may focus on a particular subarea such as along the coast of one of the Gulf States. The ability to separate studies based on the precise, non-ambiguous geographical location of interest, for example using latitude-longitude coordinates, is therefore extremely valuable to researchers.

This paper introduces a new data management tool called G-Spin (‘Geo-Spatial pins’), which handles the querying and retrieval of geospatial simulation data across multiple data repository services by integrating a Google Maps-based interface with the PyRDM research data management library (Jacobs et al., 2014; Jacobs et al., 2015). All data repositories complying with aspects of the Content Standard for Digital Geospatial Metadata (Federal Geographic Data Committee, 1998) are automatically pin-pointed on the map using markers based on the location that the dataset corresponds to, along with

---

\(^1\) Figshare: http://www.figshare.com

\(^2\) Zenodo: http://www.zenodo.org
basic metadata describing the dataset and a Digital Object Identifier (DOI) hyperlink (Paskin, 2005) to download the deposited files. Similar methods of presentation of tidal gauge data have been adopted by the National Oceanographic Database, for example (British Oceanographic Data Centre, 2015). However, G-Spin is not specific in the type of data it presents, and can therefore be used for a wider range of applications. Furthermore, the exact bounding box of the area that a dataset covers is also rendered if geospatial boundary data is included in the repository’s metadata, thereby helping researchers to find out more about the dataset without the need to download and visualise it themselves. Researchers can also search for a specific area, either by name or by latitude-longitude coordinates, and select which repository hosting services to query along with any requirements on the deposition (i.e. whether to search only private simulations curated by the user, or to search the full public repository database). Example datasets are used to present these key features of G-Spin; these datasets are curated and tagged with a geolocation (in the form of latitude-longitude coordinates) using a separate, automated publishing script which also exploits the functionality offered by the PyRDM library. Finally, some data management-related limitations and issues faced throughout G-Spin’s implementation and evaluation are discussed, along with suggestions for future work.

Design and Implementation

Overview

G-Spin adopts a web service-based approach in which the application’s code is executed on a web server. This minimises installation/setup time for large teams of researchers, and ensures compatibility across different operating systems since interaction is accomplished through the user’s web browser using a combination of standard HTML and JavaScript. The front-end (i.e. the user interface and presentation aspect) of G-Spin was implemented using Django, an open-source web development framework written in Python (Holovaty et al., 2007). This was chosen primarily because of its popularity, and ease of integration with the PyRDM research data management library (Jacobs et al., 2014) on which the back-end (i.e. the data access layer) of the application depends. The Google Maps Application Programming Interface (API) was chosen as the mapping service because it allows the straight-forward placement of markers, or ‘pins’, along with a wide range of other visualisations (Svennerberg, 2010).

The back-end of G-Spin features extensive use of the PyRDM library, which is able to interface with the Figshare, Zenodo and DSpace3-based repository hosting services. This facilitates the searching of multiple repository databases via a common interface. PyRDM also facilitates the publication of scientific data and software source code by introducing a significant amount of automation into the curation process (Jacobs et al., 2014). G-Spin currently only supports interaction with Figshare and Zenodo, but support for new services can be readily integrated.

The crux of G-Spin is its ability to query a range of supported repository hosting services such that users can find all simulation data for a particular area of interest. This type of workflow is illustrated in Figure 1 which also details the overall methodology

3 DSpace: http://www.dspace.org
employed by G-Spin. Each individual repository that satisfies user-defined search parameters and contains certain geospatial metadata (defined later) is marked on the Google Maps interface by a pin/marker. A different pin colour is assigned to each hosting service to identify each repository’s source; currently, all Figshare repositories are identified by red pins, all Zenodo repositories are identified by blue pins, and any other data sources (which in this case are unpublished, locally-stored test datasets) are identified by green pins. When a pin is clicked, the corresponding repository’s title, description, publication date and DOI are shown in the side panel. Clicking the DOI hyperlink loads the Figshare or Zenodo repository page through which the user can download the stored files if they wish to do so. If multiple repositories contain the same geolocation in their metadata, then they are placed in a single ‘spider’ pin which expands when clicked to show all the repositories within. The layout of the G-Spin interface is shown in Figure 2 along with some test datasets pin-pointed on the map.

The process of creating new repositories is handled separately to G-Spin, but still resides within the overall workflow in Figure 1. The repositories used for demonstration purposes here were created using an extended version of the automated publishing script described by Jacobs et al. (2014); this curates files required to reproduce a simulation such as its configuration file and any output data required for visualisation. If a geolocation is specified as an option at publish-time, the appropriate metadata fields (detailed in the next subsection) are created automatically. Note that if the geolocation is a city’s name or street address, then G-Spin will automatically try to resolve the required latitude-longitude coordinates using geocoding. Similar automated publishing tools can be created for the user’s particular simulation tool or other application. Alternatively, the data can be curated via the Figshare or Zenodo web interface, albeit with some caveats discussed later.

**Figure 1.** The typical workflow for users of G-Spin. The dashed lines represent functionality which is not available but may be readily implemented as an extension, such as the integration with other data sources on private servers (as long as a suitable API is available).
Geospatial metadata

For the purpose of locating and pin-pointing a repository on the map, the repository’s metadata must comply with certain (minimal) aspects based on the Content Standard for Digital Geospatial Metadata (Federal Geographic Data Committee, 1998); specifically, repositories must contain both latitude and longitude metadata fields. This minimal requirement allows a wider variety of data types to be presented to the user. For example, the computational mesh of solution points and cells, which tessellates the particular geographical area under study, may also be included. As mentioned by Jacobs et al. (2015), for the sake of reproducibility, recomputability and reuse, the mesh must also be made available along with the simulation setup files and outputs. However, it is often more beneficial and meaningful to keep the mesh repository as a separate entity from the simulation repositories, since multiple simulations frequently use one particular mesh and can simply link to that single mesh source, rather than duplicating it. G-Spin therefore also caters for repositories containing data other than simulation outputs, as long as they contain latitude-longitude metadata.

In addition to latitude-longitude coordinates, G-Spin supports the ‘bounds’ metadata detailed by the Content Standard for Digital Geospatial Metadata (Federal Geographic Data Committee, 1998) to render a ‘bounding box’ showing the geographical limits of the particular area that the repository’s data covers. An example is shown in Figure 3.
Specifically, the longitudes of the West and East-most points, and the latitudes of the North and South-most points must be provided. With this functionality users can see, for example, exactly which neighbourhood(s) of a city are being studied, and to what extent simulations performed by other researchers overlap the user’s study for the purpose of comparison and corroboration. Once again, such functionality also helps researchers to quickly decide whether a particular mesh or dataset can be reused or compared against.

![Figure 3](image)

**Figure 3.** A bounding box highlighting the bounds of the particular geographical area that the simulation, mesh, or other type of data considers. In this case, the mesh considers a portion of the coastline of Panama City Beach, Florida, USA, which may be useful for performing an inundation risk analysis concerning landslide-generated tsunamis in the Gulf of Mexico, for example.

Supporting these custom metadata fields required certain work-arounds in both G-Spin and the PyRDM-based automated publishing script, as a result of API limitations. At the time of writing, the Zenodo API\(^4\) does not yet accommodate custom metadata fields; the keywords metadata field was therefore used to store the latitude-longitude coordinates instead, although this is certainly not an ideal solution and will hopefully be revised once further API functionality becomes available. Figshare have recently released a new version (version 2.0) of their API\(^5\) that permits the creation of custom fields in the repository’s metadata, making the addition of latitude-longitude coordinates straight-forward. However, it is not yet possible to create these fields via the web interface.

---

\(^4\) Zenodo API documentation: [http://zenodo.org/dev](http://zenodo.org/dev)

\(^5\) Figshare API documentation: [https://docs.figshare.com](https://docs.figshare.com)
Querying

G-Spin includes a search/query functionality in which users can specify a geolocation, as well as options regarding the data sources (e.g. choosing only repositories hosted on Figshare, or repositories from all sources), and privacy (e.g. only search private repositories). The Google Map is then automatically refreshed with the search results. Searching private repositories requires valid authentication credentials to be provided by the user via the authentication page shown in Figure 4. The credentials for each service the user wishes to access are saved in a unique Django session (one per user) on the server-side. In the case of Figshare and Zenodo, the credentials take the form of an access ‘token’; for other services such as DSpace-based institutional repositories, a more traditional username and password may be required.

Figure 4. The G-Spin authentication page via which a user provides their personal access token for each service they wish to use.

Discussion and Ideas for the Future

The development of the G-Spin tool has highlighted several important issues and limitations concerning digital curation. Firstly, metadata fields describing a repository need to be flexible enough to allow a full description of a data object (i.e. the simulation data). Systems that restrict the allowed metadata to minimal and static fields potentially hinder the development of automated tools and limit the possibilities for data interrogation, often requiring undesirable work-arounds (e.g. in the case of Zenodo and the misuse of the keywords field to store latitude-longitude coordinates). A system flexible enough to handle different metadata standards is required if tools such as the one developed here are to present richer information to the end user scientists in a reliable way; otherwise, the extra metadata would need to be provided in a text file as part of the repository’s file set which may not be as easily searchable, standardised, and parsable by the tool.

At the same time, scientists need systems that are quick and straightforward to use, preferably with a substantial degree of automation wherever possible. The requirements concerning which metadata fields are mandatory should therefore not be too strict so as to not burden and put off researchers who might feel that publishing their data is a chore. Furthermore, while Figshare and Zenodo both offer different types of repository (e.g. one for datasets, one for figures, and one for source code) there is often a need to classify further and consider different subtypes of data (e.g. simulation visualisation files, mesh files, parameterisation data files) which is often going to be context/field-specific.
A relatively new development in Figshare’s API is that of ‘collections’ which may comprise multiple repositories (Hyndman, 2016). Such functionality can potentially be used to better organise scientific studies (e.g. to group together all simulations performed as part of a specific grant), thereby enriching the querying and data management process. It is hoped that future G-Spin and PyRDM developments will include support for such capabilities.

Finally, tools such as G-Spin have the potential to be interfaced with online diagnostic and/or visualisation tools to extend the typical user workflow further. For example, by clicking on a given repository marker/pin, extremely large datasets may be visualised via cloud computing services and rendered via the user’s browser without the user needing to download and visualise the data locally. As the challenge of Big Data becomes an increasingly important topic in computational science (Jagadish et al., 2014), such services are likely to become more abundant in the future.

Conclusion

The bringing together of different repository/deposition hosting services via a common interface offers users a much wider range of data to work with, whilst alleviating the burden of performing several separate database searches. Furthermore, by marking/cataloguing geoscientific simulations and related data by geolocation on a world map, users can more easily find the data sources they desire through a powerful visual interface.

Through the development of G-Spin, it is clear that the rich functionality already offered by scientific repository services has great potential to be exploited by scientific software as well as research data management tools. However, as noted, there are still some limitations with existing systems that will hopefully be addressed in the future.

It is hoped that tools such as G-Spin will encourage users to share their data with other researchers through publicly or privately-available repository hosting services, and facilitate data re-use and collaboration between different institutions and organisations.

Acknowledgements

The workflow diagram in this paper contains images from openclipart.org, used under the Creative Commons CC0 1.0 Universal (CC0 1.0) license. The design and implementation of G-Spin was carried out by CTJ. AA provided expertise on the areas of geoscientific interest considered in this work. CTJ wrote the manuscript.

References

Alsheikh-Ali, Alawi A., Waqas Qureshi, Mouaz H. Al-Mallah and John P. A. Ioannidis (2011). “Public Availability of Published Research Data in High-Impact Journals”. In: PLoS ONE 6.9, e24357. doi:10.1371/journal.pone.0024357.
British Oceanographic Data Centre (2015). All data series contained in the National Oceanographic Database (NODB). URL: https://www.bodc.ac.uk/data/online_delivery/nodb/.

Federal Geographic Data Committee (1998). Content Standard for Digital Geospatial Metadata. URL: https://www.fgdc.gov/standards/projects FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf.

Funke, S. W., P. E. Farrell and M. D. Piggott (2014). “Tidal turbine array optimisation using the adjoint approach”. In: Renewable Energy 63, pp. 658–673. doi:10.1016/j.renene.2013.09.031.

Hahnel, M. (2014). Loughborough University, figshare, Arkivum and Symplectic announce pioneering research data management solution. URL: https://figshare.com/blog/Loughborough_University_figshare_Arkivum_and_Symplectic_announce_pioneering_rese/136.

Holovaty, A. and J. K. Moss (2007). The Definitive Guide to Django: Web Development Done Right. Apress.

Hyndman, A. (2016). Unveiling figshare ‘Collections’ - a new way to group content. URL: https://figshare.com/blog/Unveiling_figshare_Collections_a_new_way_to_group_content/202.

Jacobs, C. T., A. Avdis, G. J. Gorman and M. D. Piggott (2014). “PyRDM: A Python-based library for automating the management and online publication of scientific software and data”. In: Journal of Open Research Software 2.1, e28. doi:10.5334/jors.bj.

Jacobs, C. T., A. Avdis, S. L. Mouradian and M. D. Piggott (2015). “Integrating Research Data Management into Geographical Information Systems”. In: Proceedings of the 5th International Workshop on Semantic Digital Archives. URL: http://hdl.handle.net/10044/1/28557.

Jagadish, H. V., Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan and Cyrus Shahabi (2014). “Big Data and Its Technical Challenges”. In: Communications of the ACM 57.7, pp. 86–94. doi:10.1145/2611567.

Kraker, Peter, Derick Leony, Wolfgang Reinhardt and Günter Beham (2011). “The case for an open science in technology enhanced learning”. In: International Journal of Technology Enhanced Learning 3.6, pp. 643–654. doi:10.1504/IJTEL.2011.045454.

LeVeque, R. J., I. M. Mitchell and V. Stodden (2012). “Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture”. In: Computing in Science & Engineering 14.4, pp. 13–17. doi:10.1109/MCSE.2012.38.

McKiernan, Erin C., Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A. Nosek, Karthik Ram, Courtney K. Soderberg, Jeffrey R. Spies, Kaitlin Thaney, Andrew Updegrove, Kara H. Woo and Tal Yarkoni (Submitted). “The benefits of open research: How sharing can help researchers succeed”. In: doi:10.6084/m9.figshare.1619902.
Mensink, C. and G. Cosemans (2008). “From traffic flow simulations to pollutant concentrations in street canyons and backyards”. In: *Environmental Modelling & Software* 23.3, pp. 288–295. doi:10.1016/j.envsoft.2007.06.005.

Mori, Nobuhito, Tomoyuki Takahashi, Tomohiro Yasuda and Hideaki Yanagisawa (2011). “Survey of 2011 Tohoku earthquake tsunami inundation and run-up”. In: *Geophysical Research Letters* 38.7. doi:10.1029/2011GL049210.

Paskin, N. (2005). “Digital Object Identifiers for scientific data”. In: *Data Science Journal* 4, pp. 12–20. doi:10.2481/dsj.4.12.

Stodden, V., D.H. Bailey, J. Borwein, R. J. LeVeque, W. Rider and W. Stein (2013). *Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics*. Tech. rep. Institute for Computational and Experimental Research in Mathematics (ICERM). URL: http://www.davidhbailey.com/dhbpapers/icerm-report.pdf.

Svennerberg, G. (2010). *Beginning Google Maps API 3*. Second edition. Berkeley, CA, USA: Apress. ISBN: 1430228024, 9781430228028.

Vines, Timothy H., Rose L. Andrew, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Nolan C. Kane, Jean-Sébastien Moore, Brook T. Moyers, Sébastien Renaut, Diana J. Rennison, Thor Veen and Sam Yeaman (2013). “Mandated data archiving greatly improves access to research data”. In: *The FASEB Journal* 27.4, pp. 1304–1308. doi:10.1096/fj.12-218164. eprint: http://www.fasebj.org/content/27/4/1304.full.pdf+html.

Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg and A. J. Moore (2010). “Data Archiving”. English. In: *The American Naturalist* 175.2, pp. 145–146. ISSN: 00030147. doi:10.1086/650340.