Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings

Volume 17 *Boston, USA* Article 14

2017

Processing Conservation Indicators with Open Source Tools: Lessons Learned from the Digital Observatory for Protected Areas

Lucy Bastin
European Commission, Joint Research Centre (JRC), Directorate D: Sustainable Resources, Knowledge for Sustainable Development and Food Security, Via E. Fermi 2749, I-21027 Ispra (VA), Italy

Andrea Mandrici
European Commission, Joint Research Centre (JRC), Directorate D: Sustainable Resources, Knowledge for Sustainable Development and Food Security, Via E. Fermi 2749, I-21027 Ispra (VA), Italy

Luca Battistella
European Commission, Joint Research Centre (JRC), Directorate D: Sustainable Resources, Knowledge for Sustainable Development and Food Security, Via E. Fermi 2749, I-21027 Ispra (VA), Italy

Grégoire Dubois
European Commission, Joint Research Centre (JRC), Directorate D: Sustainable Resources, Knowledge for Sustainable Development and Food Security, Via E. Fermi 2749, I-21027 Ispra (VA), Italy

Processing Conservation Indicators with Open Source Tools: Lessons Learned from the Digital Observatory for Protected Areas

Optional Cover Page Acknowledgements
The technical efforts described here are part of a team effort involving many developers, past and present. We would like to acknowledge the work and open-source contributions of Michael Schulz, Jon Skoien, Javier Martinez-Lopez, Mariagrazia Graziano, Santiago Saura, Martino Boni, Andrew Cottam, Michele Conti and Will Temperley.

Processing Conservation Indicators with Open Source Tools: Lessons Learned from the Digital Observatory for Protected Areas

Lucy Bastin[a,b,*], Andrea Mandrici[a], Luca Battistella[a], Gregoire Dubois[a]

[a] *European Commission, Joint Research Centre (JRC), Directorate D: Sustainable Resources, Knowledge for Sustainable Development and Food Security, Via E. Fermi 2749, I-21027 Ispra (VA), Italy.*

[b] *Department of Engineering and Applied Science, Aston University, B4 7ET, Birmingham, UK*

Abstract: The European Commission has a commitment to open data and the support of open source software and standards. We present lessons learnt while populating and supporting the web and map services that underly the Joint Research Centre's Digital Observatory for Protected Areas. Challenges include: large datasets with highly complex geometries; topological inconsistencies, compounded by reprojection for equal-area calculations; multiple different representations of the same geographical entities, for example coastlines; licensing requirement to continuously update indicators to respond to monthly changes in the authoritative data. In order to compute and publish an array of indicators, we used a range of open source tools including GRASS, R, python, GDAL, PostGIS, geometry libraries for Hadoop, Geoserver, Geonode, and Mapserver. In addition we assessed the value of the commercial ArcGIS Pro software and the Google Earth Engine platform. We describe the lessons that we learnt in building and documenting a usable and repeatable workflow,highlighting weak spots and workarounds., Code for our processing workflows will be shared via github and key process flows will be shared via a VRE to allow reproducible research while complying with data redistribution restrictions from the data providers. Our final goal is to move the entire processing chain to open source tools and share it as a versioned resource.

*Corresponding author
Email address: lucy.bastin@ec.europa.eu (Lucy Bastin)

1. Introduction

Assessing protected areas (PAs) for biodiversity conservation requires evaluation of characteristics such as the areas' connectivity and species assemblages (including the presence of threatened species), the uniqueness of their ecosystems, and the threats to which they are exposed. In this context, the Joint Research Centre of the European Commission has developed DOPA (the Digital Observatory for Protected Areas) a collection of Web tools supported by map, data and modelling services, designed to make biodiversity data readily accessible to policymakers, managers, researchers and other users (Dubois et al. 2016). The global indicators of DOPA support and frame the regional conservation picture for another set of tools developed by the JRC in the context of the Biodiversity and Protected Areas Management (BIOPAMA) programme, in particular the BIOPAMA Regional Reference Information System (http://www.biopama.org/observatories/access_the_rris). BIOPAMA builds capacity in Africa, Caribbean and Pacific to improve decision-making related to PA management. Given the importance of these decisions, the currency, consistency and quality of the data produced and shared by JRC is key in order to maintain a reliable service which is comprehensible and useful to a variety of stakeholders and decision makers.

As the technical staff responsible for populating the databases and Web services of DOPA, we realized that other individuals and institutions are undoubtedly facing the same challenges as us, with the same datasets. We have summarized some elements of our story in this manuscript, in order to start a conversation about data sharing and processing which we hope can lead to more effective collaborations in the future.

2. Our Technical Priorities

2.1. Robustness and transparency of methods

DOPA uses peer-reviewed methods to calculate a range of indicators on PA state, pressures and threats (Dubois et al. 2016); for example:

(1) A species irreplaceability index (SII) is computed according to the method of Le Saout et al. 2013 which scores each PA according to its relative coverage of suitable range for selected species. The SII score may relate to a specific taxon and/or threat group, or an overall score; it does not incorporate any consideration of complementarity in the network.

(2) The connectivity indicator developed by Saura et al. 2017 allows a user to distinguish between the proportion of a country or ecoregion which is formally protected and the amount of that protection which is adequately connected to allow movement of animals across the landscape, considering a range of different dispersal capacities.

2.2. Reliability and openness of datasets and workflows

Our goal is to use openly available datasets with clearly documented provenance, including some high-quality datasets developed in-house at JRC. The recently-published Global Surface Water product (Pekel et al. 2016) was analyzed to record area and net 32-year change of permanent / seasonal water within each PA. To evaluate human pressures on PAs, we use the Global Human Settlement map (Pesaresi et al. 2016), and in future we hope to exploit the outputs of projects such as ROADLESS (http://forobs.jrc.ec.europa.eu/roadless/) which detect and delineate the unofficial logging roads which can lead to further forest encroachment. All DOPA

indicators are made available using open OGC standards and publicly-accessible REST services (http://dopa-services.jrc.ec.europa.eu/services), and can be consumed by a wide range of web clients such as DOPA Explorer (http://dopa-explorer.jrc.ec.europa.eu/dopa_explorer/). The European Commission has a commitment to open data and open source software. Wherever possible, we use free and open source tools and support the use of these tools by developing and sharing code in GitHub (https://github.com/doctorluz/py-utils/blob/master/PostGISTableToAvro.py).

3. Challenges

DOPA analyses and aggregates global datasets encapsulating highly complex phenomena, which naturally contain certain generalizations and omissions. Issues such as data accuracy and completeness are outside the scope of this paper, since we assume that these datasets are the best possible representations of the global state of protection and of potential species ranges. Figure 1 shows the post-processing steps for one example dataset (UNEP-WCMC's World Database on Protected Areas (WDPA) (https://www.protectedplanet.net/)), to provide some context for the discussions below.

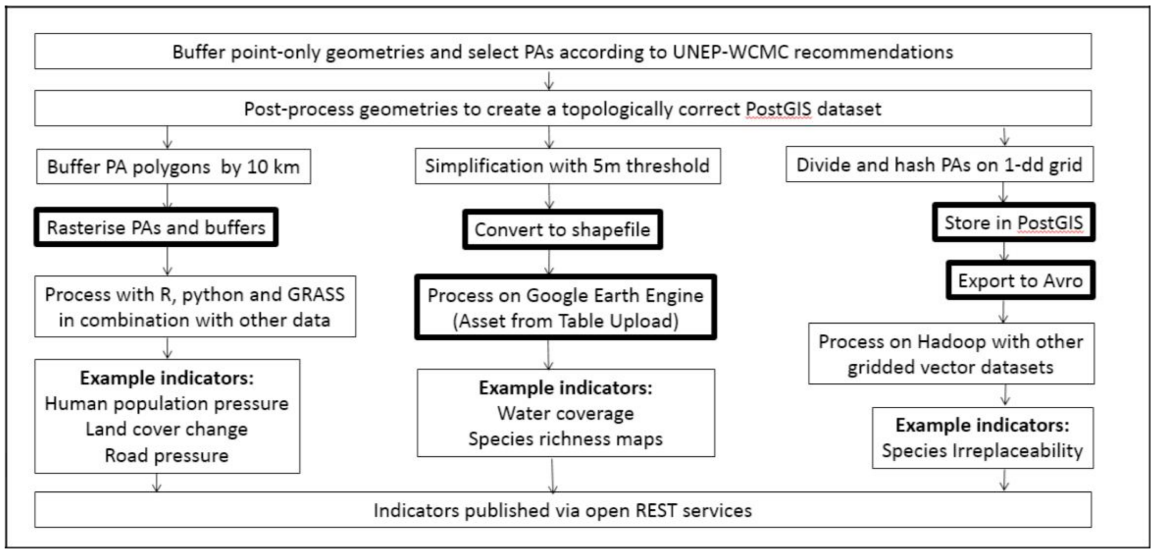

Figure 1: Processing steps for the protected area polygons. Boxes in bold indicate value-added datasets which, legal agreements permitting, could be usefully shared with other researchers to ensure consistency and reduce duplicated effort.

3.1. Staying up-to-date

The WDPA, documenting over 200,000 PAs globally, is updated every month, and information is published on the countries where PAs have been updated (https://www.protectedplanet.net/c/monthly-updates/2016/november-update-of-the-wdpa). The ideal would be to compute new metrics only for those PAs which have been added or altered, except in the case of complementarity-based indicators where the whole set of PAs will need to be reprocessed. However, at the time of writing, it is not possible to obtain a 'change-only' update. An example of good practice on this topic is the GAUL dataset of administrative boundaries, where changes are tracked using unique IDs (http://www.fao.org/geonetwork/srv/en/metadata.show?currTab=simpleid=12691).

3.2. Legal restrictions

While the datasets used for DOPA are open for public use, redistribution of data and derived products is more tightly controlled. For example, PAs can be displayed using our Mapserver or Geoserver Web Map Services, but not exposed via a Web Feature Service; and currently, we cannot share post-processed data with partners in order to ensure consistency. Since many analysts around the globe are using these datasets, it is certain that much data preparation work is being duplicated, but the curating agencies do not have the resource or mandate for this post-processing. In some cases, generation of derived products requires prior approval, a restriction which limits scientific research by third parties. Our current approach to this challenge is described in section 5.1.

3.3. Different representations of the same thing: the example of coastlines

For DOPA processing, we use datasets representing the countries of the world (GAUL, FAO https://www.protectedplanet.net/), Exclusive Economic Zones (http://www.marineregions.org/eezsearch.php), terrestrial ecoregions (Olson et al. 2001) (https://www.worldwildlife.org/biomes), marine and pelagic ecoregions (https://www.worldwildlife.org/publications/marine-ecoregions-of-the-world-a-bioregionalization-of-coastal-and-shelf-areas), coastal protected areas, and species ranges intended to delineate terrestrial or marine specialization. These coastline representations rarely or never match, leading to numerous sliver polygons, and to apparent inaccuracies such as narwhals being found on land. Nevertheless, it is crucial to derive a harmonized and consistent base data layer, since DOPA has important responsibilities in supporting regular reporting initiatives such as the Protected Planet report (UNEP-WCMC and IUCN 2016) and the country reports (https://www.cbd.int/reports/) for the Convention on Biological Diversity. The CBD is currently using the indicators provided through the DOPA Explorer to support the preparation of country dossiers for Aichi Targets 11 and 12 (https://www.cbd.int/sp/targets/), and has also encouraged its Parties to consult DOPA in the revision of their NBSAPs (National Biodiversity Strategies and Action Plans https://www.cbd.int/nbsap/). Typical metrics which must be computed in this context are: the area of an ecoregion which falls under a country's responsibility; the amount of protection for each terrestrial or marine ecoregion within a country; the relative contribution that a country is making to the protection of an ecoregion worldwide; and the number of different ecoregions which fall within a particular protected area. The last metric gives a very generalized estimate of habitat diversity, which is then supplemented by detailed biophysical modelling and assessment of marine floor geomorphological features. However, even the relatively simple overlay of countries, EEZs, ecoregions and protected areas required for the generalized estimates can be computationally demanding. For purposes of transparency, reproducibility and data consistency, it is necessary to build, from available open data sources, a high-quality, topologically correct base layer encapsulating legal country boundaries permissible for use by the European Commission, EEZs and ecoregions. The complex procedure, which currently involves the use of a wide range of software to handle the different steps, is described in Section 5.2.

3.4. Technical and interoperability challenges

The datasets that we integrate use a variety of data models and standards. For example, the topological models of PostGIS and ESRI differ in terms of legal coordinate ordering for internal rings, and ESRI shapefiles, which are frequently used as an exchange format, comply with a Simple Feature data model rather than the Topological model required

by GRASS. Reprojection to equal-area projections such as Mollweide to assess real-world extent can introduce self-intersections, even where polygons were originally topologically correct. Some datasets span the range from -180 to 180 decimal degrees, while others (e.g. http://www.marineregions.org/download_file.php?fn=v9_20161021_HR_0_360) can be obtained in a 0-360 degree version.

Even when a workflow is documented, operations are often specified without details of the specific software or algorithm used. For example, geodesic buffers are recommended in the processing guidance for the WDPA (http://wdpa.s3.amazonaws.com/WDPA_Manual/English/WDPA_Manual_1_4_EN_FINAL.pdf) and are also used in DOPA to compare conditions inside and outside PAs. However, different software can yield very different results. For PAs which cross the dateline, ArcGIS produces a geometrically-correct buffer in two segments, while both PostGIS and the Hadoop spatial framework (https://github.com/Esri/spatial-framework-for-hadoop) create a self-intersecting polygon that spans the entire globe. Thus, in order to ensure consistency, we translate such polygons to a 'safe' longitude before buffering them, then translate the buffer back to its original location and, if necessary, split the buffer on the dateline. The whole workflow, including this extra step, is described in an open access document (https://dopa.wikispaces.com/WDPA+protected+Area+boundaries), and summarized in Section 5.1.

3.5. 'Big Data' problems

Given the global nature of DOPA and its ambitious aims, the input datasets are necessarily large and extremely topologically complex. For a subset of our metrics, only PAs over 50 square km are processed, (this amounts to 98% of global protected area) but even this reduced dataset has over 20,000 polygons. The most complex PA (Vindellven in Sweden) has over 862 thousand vertices, and the species ranges for the IUCN Red List are equally challenging, particularly for many bird species and marine mammals, which span the globe while incorporating complex coastlines. Rasterization of the data simplifies computation, but also brings new challenges and tradeoffs: for example, when calculating zonal statistics for large regions and small (<=30m) pixels, on a distributed platform such as Hadoop or Google Earth Engine, tasks must be carefully partitioned across space and results re-aggregated from smaller regions, to minimise communication between multiple nodes. Large PAs close to the poles are especially problematic, especially if processed in a non-sinusoidal projection, since the number of pixels to be aggregated rises far beyond the actual area for which information is being derived. Vector processing and data sharing with Google Earth Engine is also fast advancing: Fusion Tables could each contain a maximum of 1 million vertices, but have been very recently superseded by the more flexible Table Upload tool (https://developers.google.com/earth-engine/importing).

4. Potential Solutions

As a platform for storing, cleaning, processing and managing vector data, and for underpinning the REST service interface that delivers DOPA indicators to Web clients, PostGIS (http://www.postgis.net/) has been crucial to our work. At the time of writing, we are setting up a more powerful instance of PostGIS 9.6, which will allow us to exploit multi-threading capabilities and evaluate where PostGIS Raster can be of value (for example, speeding up overlay of relatively small raster polygons on tiled global datasets). Other tools such as GRASS (https://grass.osgeo.org/), with its well-established functions for vector and raster processing, and python and R spatial libraries, have been heavily used, along with fundamental libraries such as gdal (http://www.gdal.org/). For certain vector analyses we have

made use of Hive on a Hadoop cluster (see section 5.3) and for pixel-wise computations requiring access to large archives of EO data or global derived products, and for easy prototyping of the MapReduce approach, we have benefitted from the capabilities of Google Earth Engine (https://earthengine.google.com). As an organization with heavy network security, it is difficult to install specific modules and libraries on shared processing servers, and docker (https://www.docker.com) has been extremely useful as a means to set up and share self-contained working environments with specific collections of software and models.

We also exploit established best practice in the geospatial field, by partitioning the data in a smart way so that it is more tractable for processing while also being as reusable as possible. For relatively static datasets we tested data-dependent griddings, which as closely as possible equalized vertex number in cells of varying resolution. However, a common grid is required for more dynamic datasets and for sharing with other researchers. Currently we have settled on a 1-decimal degree tiling (described in Section 5.3) but equal-area grids such as the ISEA hexagonal grid supplied in IUCN's Red List Toolbox (http://www.iucnredlist.org/technical-documents/red-list-training/iucnspatialresources) are also good candidates. In many cases, this is a matter more of cultural than of technical interoperability. Given the good progress of OGC's Discrete Global Grid Standards Working Group towards an accepted standard and implementation pilots, we will follow the developments and demonstrators of this community, in order to be ready to adopt useful tools and data structures which emerge to support interoperability.

5. Example Case Studies

5.1. Sharing a processing workflow: Preparing the WDPA for use on several platforms

Our processing workflows (e.g. Figure 1) are described in open-access documents (http://www.marineregions.org/eezsearch.php), but, as described above, this is not sufficient to guarantee a reproducible result. Therefore the code will also be shared on GitHub and published as a Data Paper. The long-term goal is to share the processing itself through a Virtual Research Environment. This processing workflow can now be almost entirely executed using PostGIS (steps 1-8) and ogr2ogr (steps 12-13) after the initial import of the provider's feature geodatabase or shapefile (Bastin and Mandrici 2017). The one section of the workflow which remains intractable using PostGIS, ArcGIS Pro or GRASS is the generation of external buffers for every PA (steps 8-11). This necessitates the use of Hadoop for this step at present, but the first test for our emerging, more powerful database instance which will exploit the multi-threading capabilities of PostGIS 9.6, will be to evaluate whether this step can be brought back into the PostGIS processing chain.

The relevant WDPA version (IUCN and UNEP-WCMC 2017) is downloaded in a lat/long EPSG:4326 projection, and processed as follows:

(1) Protected areas with point-only geometries and a reported area >0 are given a circular geodesic point buffer with the reported area.
(2) Polygon self-intersections are corrected using the ST_MakeValid function of PostGIS, and point and line geometries resulting from the correction are discarded.
(3) Polygons digitized with only two coordinates (these mainly consist of sunken US ships which provide useful marine habitat) are converted to lines and buffered by a very small distance to create valid polygons.
(4) Features at the dateline whose geometries fall outside legal coordinates are split, and the resulting part of the geometry are each translated to the correct side of the dateline.

(5) Where a PA is divided into several parcels representing zones (separate records in the WDPA), these are aggregated to form a single union, so that WDPA ID is the unique identifier for the dataset. Any internal slivers and topological inconsistencies which result are repaired and removed. In these cases, (around 450 of the >200,00 PAs) parcel names, IDs etc are concatenated in a comma-separated string to allow a single record for each PA.

(6) *(January 2017 version only), to deal with erroneous duplicate PA IDs in the original dataset, the ID for Polesye Valley of River Bug, Belarus should be changed to 555624313 (the ID value for March 2017).*

(7) As recommended by WCMC (http://wdpa.s3.amazonaws.com/WDPA_Manual/English/WDPA_Manual_1_4_EN_FINAL.pdf), the data is filtered to remove all features with a status of "not reported" or "proposed", and all features designated as UNESCO Man and Biosphere Reserves.

For a subset of protected areas exceeding 50 square km in area, 10-kilometre buffers are produced in order to compare conditions inside and outside the PA, as follows:

(8) Polygons where any part of the PA or its buffer would cross the dateline are translated horizontally by 180 degrees to ensure correct buffering (This is because of issues with many buffering functions at the dateline, including the ESRI libraries available for use on Hadoop).

(9) All polygons are processed using Hive on a Hadoop cluster to produce an external geodesic buffer of 10 km.

(10) Polygons which were translated in step 8, along with their buffers, are split as appropriate and moved back to their correct positions on the dateline.

(11) Buffer areas are masked to eliminate areas which, although in the exterior buffer of a protected area, are covered by protection from another PA - i.e., to include only unprotected regions.

Additional processing for raster products

(12) The buffered PAs are rasterised individually to produce geotiffs with a background value of 0, a pixel value of 1 in the buffer and 2 in the PA itself. The resolution of these rasters is c. 100m at the equator, in a projection of EPSG:4326.

(13) For masking purposes (see step 11) and for use on Google Earth Engine for computing country-level protection of water and forest, a binary raster containing all PAs with polygon geometries is generated at a matching resolution to step 12.

The code for all above steps will be shared on github, and wrapped as python or R modules to allow easier configuration and progress tracking. The latter implementation will be hosted on a VRE by the end of 2017, to allow other users and research partners to generate and document data consistent with our processing methods. *For information on the progress of this initiative, please contact the authors on the email addresses provided.*

5.2. Building a consistent, well-documented base layer: Countries, EEZs and ecoregions

DOPA disseminates indicators based on intersections of the WDPA dataset with other data including population, landcover, etc. However, in order to correctly attribute protection efforts to the relevant countries and governments, and to assess the coverage of the world's habitats by protection, a reliable base layer encapsulating the combination of political and habitat boundaries is required. The generation of a consistent dataset which covers all areas of the globe with no overlaps has been challenging, and is documented below. Where possible, open source solutions have been used, but a few steps have required the memory management of ArcGIS

Pro. We are actively seeking means to move these steps into the open-source domain in order to avoid restrictions on reproducibility.

5.2.1. Country boundaries

This dataset is built from a combination of GAUL country boundaries and EEZ exclusive economic zones (referenced earlier). The two datasets come with different coastlines, different attribute structures; (they have only one common field: ISO3), and small differences within the content (for example, sovereignty updates).

Normalization of these two datasets involved (in brackets are the tools used at each stage):

(1) Intersection: where there is overlap, GAUL wins (QGIS, PostGIS, ArcGIS Pro).
(2) Gap filling: where there is a gap between the two coastline, this is filled and attributed to the containing EEZ (ArcMap 10.3).
(3) Unification of the attribute structure (PostGIS).
(4) Identification of Disputed, Joint Managed areas, and attribution to multiple ISO3 codes (PostgreSQL).
(5) Identification of the areas not covered by any of the two dataset as ABNJ - Area Beyond National Jurisdiction (SAGA).
(6) Identification of updated information according to the ISO 3166-1 standard (PostgreSQL).
(7) Dissolve to remove non-unique keys, where present (PostGIS).
(8) Geometric simplification (10 meters, preserving topology) (PostGIS).
(9) JOIN with ancillary information (ISO2 and ISO numeric/UN_M49) to include additional codes.
(10) JOIN with ancillary information to get country grouping information for example, continental and regional groupings which are of relevance for conservation funding and policy decisions (PostgreSQL).

The result is a dataset (Figure 2, left) with updated geometries and structure, which maintains the original source codes (ADM0 for GAUL, MRGID for EEZ).

5.2.2. Ecoregion boundaries

Sources for this dataset are:

(1) **Terrestrial Ecoregions of the World**: biogeographic regionalization of the Earth's terrestrial biodiversity. (TEOW) (Olson et al. 2001).
(2) **Marine Ecoregions and Pelagic Provinces of the World**: biogeographic classification of the world's coastal, continental shelf, and surface pelagic waters (The Nature Conservancy (TNC) 2012). This dataset combines two separately published datasets:
  (a) **Marine Ecoregions Of the World** (MEOW) (Spalding et al. 2007) and
  (b) **Pelagic Provinces Of the World** (PPOW) (Spalding et al. 2012)

The two datasets differ in their attribute structure. Normalization involved:

(1) Intersection: where there is overlap (MEOW/PPOW comes without coastline), TEOW wins (ArcGIS Pro).
(2) Unification of the attribute structure (PostgreSQL).

The result is a tessellation of the world with no gaps, where every polygon can be allocated

to a terrestrial or marine ecoregion, or a pelagic province (Figure 2, right).

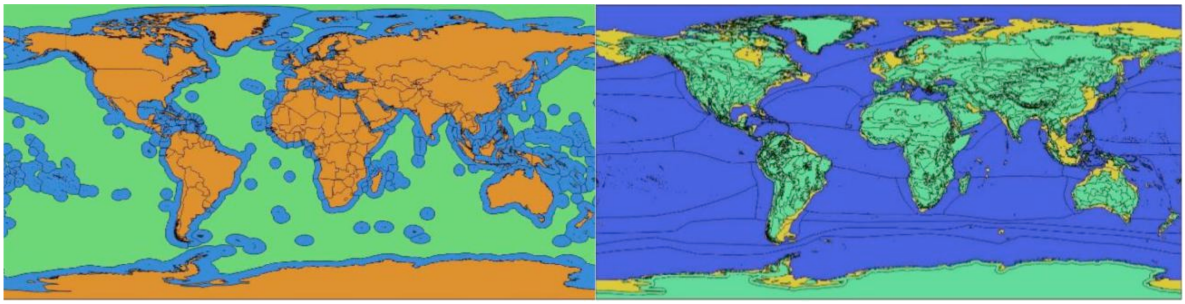

Figure 2: The final base layers for countries/EEZs (left) and ecoregions (right).

5.2.3. Unified base layer

To reduce the number of processing loops in computing indicators for the WDPA, the two above datasets are merged. The target is to achieve, with a single processing cycle, information on:

(1) Country
(2) Ecoregion
(3) Terrestrial/marine breakdown of the country and many further statistics related to the above as:
    (a) number of PAs per country
    (b) number of marine/terrestrial/coastal PAs per Country
    (c) area/percentage of a country's ecoregions which are protected
    (d) number of PAs per ecoregion
    (e) etc.

Merging of the two datasets involves:

(1) Dissolve of the country/EEZ on country codes into larger polygons bounded by the country's EEZ extent. In other words, the coastlines visible in Figure 2 are discarded. (PostgreSQL, ArcGIS Pro).
(2) Intersect (UNION) of the geometries and attributes of the result from the above step, and ecoregions. This means that, for the purposes of this analysis, the junction between land and sea is determined by the TEOW dataset rather than the GAUL geometries (ArcGIS Pro). The consequences of this decision are illustrated in Figure 3, and further discussed below.

Each polygon contained in the resulting dataset contains information on ecoregions (TEOW, MEOW and PPOW original codes), country (ISO3/ISO2/UN_M49), and presence of land or sea (as defined by the TEOW coastline, which is lower resolution than the GAUL coastline, but is still adequate to this particular analysis, given the data resolutions of the phenomena that are being summarized at a global scale).

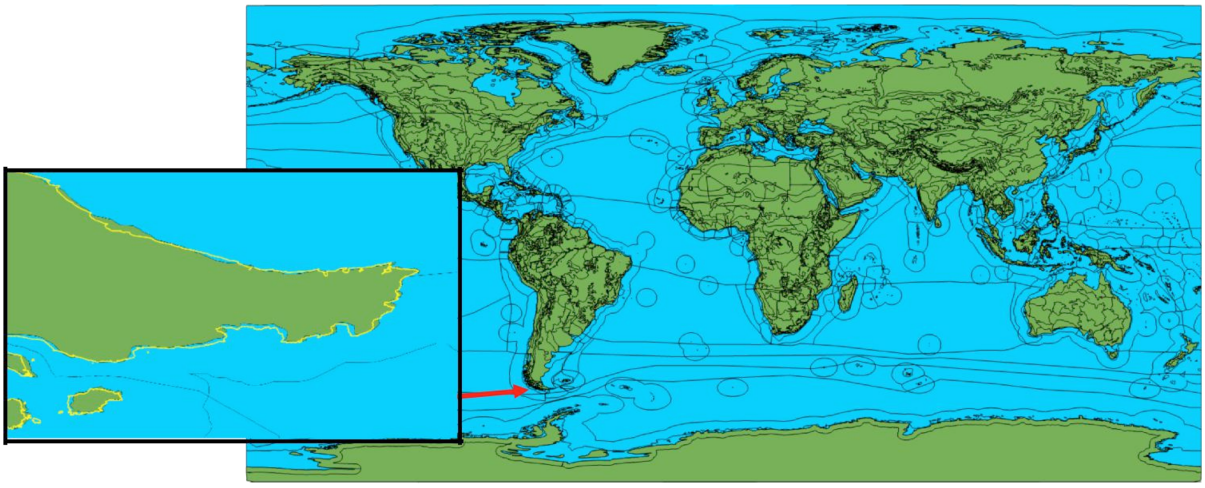

Figure 3: Unified Base layer, with example of discrepancy between GAUL (yellow) and TEOW (black) coastlines.

5.3. Multi-way intersections species ranges and protected areas

The SII described in section 2.1 requires a demanding multi-way intersection between over 200,000 protected areas and over 27,000 species ranges. This operation is achieved by a segmentation of the data on a one-decimal degree grid to form a list of either simple squares (where the grid cell is completely contained) or smaller polygons. When applied to all input datasets, this approach allows a straightforward SQL query using the Hadoop ESRI spatial framework, where an ST_INTERSECTION operation is only necessary for those cells which are not completely contained by either polygon. A bespoke Java class written to support this procedure can be found on GitHub (Lars Francke, 2014 https://github.com/lfrancke/jrc/blob/master/src/main/java/jrc/CellCalculator.java).

Even when the task is thus simplified, there are still occasional errors in the first (union) step of the procedure, and these are much harder to debug once the data is distributed across the Hadoop cluster. Therefore the gridding and union steps were moved off the cluster, as shown in Figure 1. This has the added benefit of making helpfully-gridded and indexed data available for use on other platforms.

6. Lessons Learnt

The wide range of demanding analyses behind DOPA mean that different tools are suited to each step of the workflow, and this has been challenging in terms of provenance documentation and data models / formats. However, careful preparation of the data can make it more easily reusable across a variety of platforms. We have learnt that at present, some steps of our processing still need to be carried out using commercial software (for example, large-scale intersections in ArcGIS Pro), but are very close to testing more powerful and better-supported instances of open source alternatives on JRC's new JEODPP Big Data infrastructure.

Most users of environmental datasets are trying to do reproducible and accountable science, but different post-processing workarounds and tools can lead to published results which are not repeatable or comparable. To work more effectively, we would ideally share value-added data processed to an agreed standard and format. Since legal restrictions currently forbid this type of redistribution, the next best solution is to share the processing workflow, including

code and environmental settings or parameters. Either the whole environment can be shared (for example using docker or vagrant) or the processing itself can be published as a service (for example, DOPA's marine geomorphological metrics are computed in collaboration with the BlueBridge (https://bluebridge.d4science.org/explore) project, which uses the D4Science VRE infrastructure).

Processing power alone has not helped us to break down and overcome these challenges: far more fruitful has been the process of thinking through and classifying the problems faced, and we hope to share this active reflection process with other researchers in future.

References

Bastin, L., Mandrici, A., 2017. WDPA processing workflow description. Online; accessed May 31, 2017.
URL https://dopa.wikispaces.com/WDPA+protected+Area+boundaries

Dubois, G., Bastin, L., Bertzky, B., Mandrici, A., Conti, M., Saura, S., Cottam, A., Battistella, L., Martinez-Lopez, J., Boni, M., 2016. Integrating multiple spatial datasets to assess protected areas: Lessons learnt from the digital observatory for protected areas (dopa). ISPRS International Journal of Geo-Information 5, 242.

IUCN, UNEP-WCMC, 2017. The World Database on Protected Areas (WDPA). Cambridge, UK: UNEP-WCMC. Online; accessed May 31, 2017.
URL www.protectedplanet.net

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'Amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., Kassem, K. R., 2001. Terrestrial ecoregions of the world: a new map of life on earth. Bioscience 51(11), 933–938.

Pekel, J. F., Cottam, A., Gorelick, N., Belward, A. S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418422.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A. J., Freire, S., Halkia, S., Julea, A. M., Kemper, T., Soille, P., Syrris, V., 2016. Operating procedure for the production of the Global Human Settlement Layer from landsat data of the epochs 1975, 1990, 2000, and 2014. Publications Office of the European Union, EUR 27741 EN.

Saout, S. L., Hoffmann, M., Shi, Y., Hughes, A., Bernard, C., Brooks, T. M., Bertzky, B., Butchart, S. H. M., Stuart, S. N., Badman, T., Rodrigues, A. S. L., 2013. Protected areas and effective biodiversity conservation. Science 342, 803–805.

Saura, S., Bastin, L., Battistella, L., Mandrici, A., Dubois, G., 2017. Protected areas in the worlds ecoregions: how well connected are they? Ecological Indicators 76, 144–158.

Spalding, M. D., Agostini, V. N., Rice, J., Grant, S. M., 2012. Pelagic provinces of the world): a biogeographic classification of the worlds surface pelagic waters. Ocean and Coastal Management 60, 19–30.

Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., 2007. Marine ecoregions of the world: A bioregionalization of coastal and shelf areas. Bioscience 57, 573583.

The Nature Conservancy (TNC), 2012. Marine Ecoregions and Pelagic Provinces of the World. GIS layers developed by The Nature Conservancy with multiple partners, combined from Spalding et al. (2007) and Spalding et al. (2012). Cambridge (UK): The Nature Conservancy. Online; accessed May 31, 2017.
URL http://data.unep-wcmc.org/datasets/38

UNEP-WCMC, IUCN, 2016. Protected planet report 2016. UNEP-WCMC and IUCN: Cambridge UK and Gland, Switzerland.