Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Expansive, informative datasets are vital in providing foundations and possibilities for scientific research and development across many fields of study. Assembly of grand datasets, however, frequently poses difficulty for the author and stakeholders alike, with a variety of considerations required throughout the collaboration efforts and development lifecycle. In this work, we discuss and analyse the challenges and opportunities we faced throughout the creation of a transnational, European agricultural dataset containing reference labels of cultivated crops. Together, this forms a succinct framework of important elements one should consider when forging a dataset of their own.

1 Introduction

With the progression of member states of the *European Union (EU)* to continually and openly publish administrative data, the opportunity to conduct research previously limited to a local-scale has advanced massively with the emergence of expansive, collaborative, and multi-national datasets. The ANOMDATASET initiative[1] harnessed the continually improving accessibility to *Common agriculture policy (CAP)* [3] data, demonstrating the feasibility of a transnational dataset and building out a framework of considerations one must account for when developing projects at a similar spatial coverage. While there are clear advantages of a large and diverse dataset for all breadths of research, the number of challenges imposed by administrations' legacy, and often manual, systems continues to make gathering data at this scale an ever-laborious task. Marini and Nicolardi [4], for instance, face these obstacles when they harmonised several sub-databases containing real estate information in order to analyse the social and economic changes within Europe. More generally, Connelly et al. [1] analysed the issues with administrative social science data while stressing the importance of its use.

In this paper, we distilled the challenges we faced while working on ANOMDATASET into a framework consisting of six distinct categories, which will be further explored in the sections following. The goal of this framework is to provide researchers and authorities with guidelines when approaching transnational, country-dependent, and collaborative open data projects.

While highlighting the challenges involved, we also want to encourage the community to take part in the development of large pan-European datasets by showcasing the advantages and benefits they are able to provide for a multitude of stakeholders.

[1] Dataset name and references hidden for blind review.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

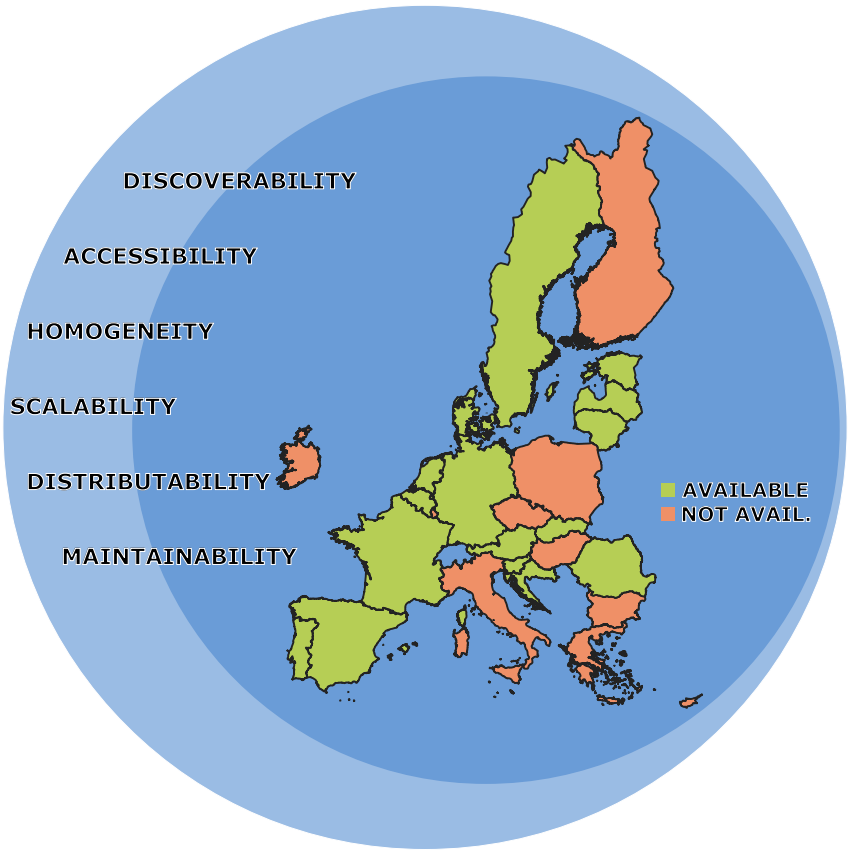

Figure 1: The ANOMDATASET dataset now includes data from 16 member states of the *European Union (EU)*. The countries coloured in green provide data and helped with the identification of the six biggest challenges faced when compiling collaborative, transnational datasets, which are listed on the left. At the time of the first ANOMDATASET release, data from the red countries was not available yet.

2 The Study Data

ANOMDATASET can be considered to be the first large-scale, pan-European dataset for cultivated crops, following the approaches of the LUCAS [2] and BreizhCrops [5] initiatives, combining and harmonising georeferenced agricultural parcel data with the corresponding crop. While originally motivated as a use-case for applying machine learning methods to Earth observation data, more specifically Copernicus Sentinel-2 satellite imagery, ANOMDATASET has now become an impactful project spanning across a number of disciplines, including biodiversity and agriculture. The data itself was obtained through the countries' agricultural ministries or paying agencies and contains the farmers' self-declarations, which are collected within the subsidy control of the *Common agriculture policy (CAP)*. At the time of this study's publication, 16 member states of the *EU* have contributed data, highlighted in green in Figure 1, which we collected over a span of 1.5 years.

As part of the collection effort, a new *Hierarchical Crop and Agriculture Taxonomy (HCAT)* was designed to both accommodate the high crop diversity and harmonise country-unique cultivation schemas and languages across Europe. The harmonised ANOMDATASET vector data is centrally

and publicly available via the project's GitHub repository[2], on Zenodo[3], and will be distributed over commonly utilised platforms, such as Google Earth Engine[4], GeoDB[5], CODE-DE[6], and EO-Lab[7].

3 Challenges

While compiling the ANOMDATASET dataset, we identified the following six challenges, forming the foundation of the aforementioned framework.

3.1 Discoverability

One would expect that "open" data would indicate the simplicity of locating the hosted data and determining the type and quantity of the available data; we found this was often not the case. As countries frequently choose not to use central data distribution points but instead their own national platforms, the method of actually finding these self-hosted repositories still requires incredibly manual investigatory work. Similar data is often hosted by different types of responsible agencies and authorities across Europe, raising the barrier when attempting to discover the available sources.

Even in the scenarios where the correct authority contact and platform is determined, one still encounters language barriers when navigating local websites, as they are often only hosted in their native language.

3.2 Accessibility

The ever-growing concerns surrounding security and storage of personal data continues to motivate authorities and platforms to restrict direct access to the data itself. This is especially true for data that can be linked back to individuals, notably farmers in the case of ANOMDATASET , and provides authorities the justification to obscure themselves behind GDPR walls, even if there are legally-just reasons for public access.

Unfortunately, the benefits of taking the step to openly release the data are not justifiably strong enough for the responsible parties to undertake such an effort. From the outside, there is no observably clear argument against working on the publication, and we can only hypothesise that it might be connected to lack of digitalisation efforts, internal politics, and availability of resources to appropriately prepare the data. Where resource has been found and appropriately applied, it has sometimes resulted in platforms necessitating user registration before being able to gain access, further adding to the manual effort.

3.3 Homogeneity

While lightly touched on previously, an important consideration is the variability in the data content released across countries. Regarding cultivated vegetation, each country groups their crop data into unique taxonomies and schemes, which further vary between legacy and modern formats. This includes mixtures of scientific plant nomenclature and terminology in national languages, but also the depth of granularity, ranging from precise biological species to the more generic plant families.

Not only does the data content vary across countries within the EU, but so do the data formats, most notably in fields where severally accepted formats compete within the community. For example, we observed that, despite being proprietary, Esri shape files still contribute to a large proportion of the geospatial data pool, amongst more modern formats, such as geopackage or webmapservice. The necessity to handle multiple, conflicting formats when pulling the data from respective authorities further impedes research and development at a transnational scale.

[2] URL hidden for blind review..

[3] URL hidden for blind review..

[4] `https://earthengine.google.com/`.

[5] `https://eurodatacube.com/marketplace/services/edc_geodb`.

[6] `https://code-de.org/en/`.

[7] `https://eo-lab.org/en/`.

3.4 Scalability

In the cases where individual country-dependent datasets are fetched, harmonised, and combined into an aggregated collection, the collection itself is still encapsulated within the pool of available national datasets. The effort of including new sources then requires restarting this harmonisation process from the beginning, to ensure the content of the additional data complies with the existing collection, while the existing structures have to be altered to accommodate the new dataset. This is not only true when new countries are added to the transnational collection, but also when ministries or authorities of a certain country migrate from legacy data platforms, or update content structure or format. These can happen irregularly over the course of years and therefore require manual monitoring and intervention.

3.5 Distributability

When releasing or updating the complete dataset, there is the problem of ensuring that all relevant stakeholders receive notification and central access to the data. Under current circumstances, this typically involves reaching out to each member state individually. Even though European data repositories exist, these currently lack the functionality to act as a distribution hub, where all information is brought together and subsequently used as an exchange platform for all concerned parties: it is still required to actively engage authorities to ensure they receive the latest updates.

3.6 Maintainability

With the increasing size and number of interactions with a dataset, progressively entangled, moving parts begin to co-exist and need to be continuously maintained. In the case of ANOMDATASET , a number of country-owned variables will undergo constant development, including: changes to crop policies; infrastructure updates, such as to platforms and authority websites; and standards. Each of these provides some underlying structure or metadata to the dataset itself, and so require strict manual verification to ensure a high dataset quality is maintained.

When changes to country-owned variables do occur, these are typically carried out without communication, version-control, and tend to be obscured in the background, with error corrections even understated on the authorities' data platforms. This makes it extremely challenging to pin down and correct outliers in datasets without significant effort.

4 Opportunities

While construction of these large, interwoven datasets presents the array of challenges as discussed above, the opportunities of providing communities with dataset like ANOMDATASET far outweigh the effort involved. In the following section, we discuss three of the key target groups and the complementary benefits that the dataset would deliver.

4.1 For Earth Observation

As many in the Earth observation community are aware, provision to easily-accessible and complete datasets is often limited and difficult to obtain. Initiatives to create open-access datasets covering vast swathes of Europe considerably lowers the barrier to entry, providing members of the community with large-scale, manipulable datasets on which they can begin to operate. These datasets are important, as large-scale Earth observation analysis might require substantially more reference ground data for training machine learning models than one country may be able to provide alone.

Furthermore, Earth observation reference data is often complicated to both understand and work with, being frequently obscured by multiple layers of abstraction and more difficult to find when compared to natural image labels. By overcoming this hurdle, the domain itself sees further benefits through increased accessibility to outside researchers.

4.2 For Europe

A clear benefit of transnational datasets is the weakening of countries' data sovereignty, which has historically led to encapsulated and incompatible granular datasets and models.

By giving the citizens of the European Union the ability to access and use all available data in a consolidated location and format, research can take a transnational direction, allowing for broader, pan-European problems to be tackled and solutions developed. This in turn enables far greater potential for both bi- and multilateral research projects and data analysis across borders.

4.3 For Ministries and Authorities

When considering ministries and authorities, the benefits of actively participating within transnational datasets take another clear step: data can be accessed, controlled, and maintained within a singular ecosystem; changes can be version-controlled and publicly-stated; and metadata can be standardised.

By providing a unified node that holds all necessary information to download and work with the dataset, authorities are provided with a simple means of interacting with the public and those who request access, as well as establishing a forum for discussion and feedback to be held.

Additionally, with the growth of a singular, established, transnational dataset, there comes the appeal for other countries to participate. This provides consequent and progressive reward to participants, as they can rely on an existing baseline against which they can fit their data, instead of developing another from the ground up. Making the entire process far more scalable, both horizontally and vertically, delivers tangible time and effort benefits from the start.

5 Conclusion

In this article, we have presented and discussed both the challenges and opportunities that the development of a large-scale, transnational dataset may deliver.

Through the introduction of ANOMDATASET , we have started to tackle the *discoverability* and *accessibility* barriers by providing one publicly available dataset, which is currently being published on several, well-established and easily discoverable platforms. The development of *HCAT* expands on this, introducing a sense of *homogeneity* through harmonisation of reference data and allowing for *scalability*, both with new data from existing countries, as well as newly participating countries.

Finally, during the project itself, we managed to build up a community of data providers and data users and form stable connections with a number of authorities and paying agencies across several member states of the European Union. This way, we hope to actively address the issue of *maintainability* of the dataset and *distributability* of information to all participants of the process, through our now established lines of communication and continually addressing feedback and concerns.

References

[1] R. Connelly, C. J. Playford, V. Gayle, and C. Dibben. The role of administrative data in the big data revolution in social science research. 59:1–12. doi: 10.1016/j.ssresearch.2016.04.015.

[2] R. d'Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eiselt, A. Palmieri, P. Dominici, J. Gallego, H. I. Reuter, C. Joebges, G. Lemoine, et al. Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union. *Scientific data*, 7(1):1–15, 2020.

[3] European Commission. Common agricultural policy. *agriculture.ec.europa.eu* `https://agriculture.ec.europa.eu/common-agricultural-policy_en`.

[4] C. Marini and V. Nicolardi. Big data and economic analysis: The challenge of a harmonized database. In *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 235–246. Springer. doi: 10.1007/978-3-030-51222-4_18.

[5] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner. Breizhcrops: A time series dataset for crop type mapping. *arXiv preprint arXiv:1905.11893*, 2019.