# HIGH VALUE DATA SETS IN GERMANY – EXECUTIVE SUMMARY

STUDY COMMISSIONED BY THE GERMAN FEDERAL MINISTRY OF ECONOMIC AFFAIRS AND ENERGY

## Initial situation and objective of the study

The amended EU Directive on Open Data and the re-use of Public Sector Information (PSI) obliges Member States to identify and publish so-called High Value Datasets (HVD) in six thematic areas. This study was commissioned by the German Federal Ministry of Economic Affairs and Energy to inform the German position in subsequent negotiations between EU Member States and the EU Commission to determine these datasets. The study identifies *potential* German HVD, analyses their technical, legal, and economic status quo and gives recommendations on required adjustments should the relevant data be classified as HVD.

## Selection of HVD

The analysis began with a list of already existing HVD proposals, particularly those mentioned in the PSI Directive and in a simultaneously conducted, currently (i.e. in December 2020) unpublished study on behalf of the European Commission. To gather insights into the value-added potentials of different groups of prospective HVD, an online survey was conducted. Subsequently, concrete data sets from Germany were identified, corresponding as closely as possible to the aforementioned HVD proposal lists. In a third step, the selection of data sets, their potential for added value, and relevant challenges were evaluated separately by the project's expert advisory board and in six stakeholder workshops. In the process, the views of both data providers and users were taken into account. Eventually, a set of policy and legal recommendations has been developed based on these inputs.

## Results

Following the rationale of the revised PSI Directive, a dataset has high value, if its potential for added value is high, particularly if the data were to be provided openly and in a good technical order. Classifying such datasets as HVD can result in new services, products or business models, more efficient production processes and increased competition in general. There can also be positive socio-economic effects, such as more transparency, better access to knowledge for citizens, and the preservation of natural resources. However, a classification as HVD can also negatively impact existing business models, make it more complex to prepare data for publication, or lead to the loss of revenues for data providers.

Given the prominence of these essentially economic considerations, this research did not only assess the legal applicability of the PSI Directive for each dataset, but also evaluated the positive effects and assumed costs of classifying data as HVD.

Starting with the legal assessment, almost all analysed datasets fall into the scope of the PSI Directive. Notably, this also includes companies' register data as the courts providing such data are public authorities in the sense of the PSI Directive. However, the examined companies' register also contains personally identifiable data. In this respect, further clarifications should be sought from EU and national data protection authorities. Potential privacy problems can certainly be addressed via technical means, particularly through suitable anonymisation procedures. Nevertheless, there may be trade-offs in the shape of additional implementation costs and a reduced utility of relevant data.

Clear and non-restrictive, ideally truly open licensing conditions are an essential success factor to increase the reusability of data. Here, the legal analysis showed that, on the one hand, many of the examined German data sets can already be used subject only to attribution requirements. On the other hand, some datasets are subject to more restrictive licensing and reuse conditions, making reuse more difficult and complicated as past experiences have shown. Missing information on the type of licence or copyright, as well as excessively narrow or unclear rules for the combination of data sets with different licences are frequent factors that limit the reusability of datasets. On the EU level, the implementing act on HVD could address these issues by referring to uniform or standard licensing conditions and model licences.

For the economic assessment of any potential added value, the supply cost and licensing conditions of potential HVD are an essential starting point. The study shows that the examined potential HVD fall broadly into two groups:

For data sets that are already available free of charge, in good technical order and with no restrictions on reuse, HVD classification offers little to no added value. Since the most essential conditions for easy reusability are already fulfilled, their potential is already largely being realised, e.g. in business models. However, the challenges associated with an HVD classification are also often low for such datasets.

For data sets that are currently only available for a fee, not in good technical order, or that come with restrictive re-use conditions, an HVD classification offers, in some cases, medium to high potential for added value. In particular, the (improved) availability of standard land values, terrain models or companies' registers could drive new business models. Still, depending on the exact scope of necessary adjustments, an HVD classification of those data could lead to the loss of fees and revenues as well as other, potentially high follow-up costs for affected data providers. Specifically, classifying cadastral data as HVD would raise substantial challenges in Germany, including substantive initial investment costs, mainly in the form of monetary costs for data providers, and indirect effects through distortions of competition. In the case of cadastral data, sub-national regulations would have to be harmonised between German federal states. Such harmonisations would have to align regulations on chargeable costs, deal with revenue losses for data providers, and achieve the application of uniform technical standards. The associated costs of such changes may, however, exceed any expectable positive effects.

From a technical point of view, a large proportion of the examined datasets meet the HVD requirements already, i.e. data provisioning in machine-readable formats and via programming interfaces (API). This is particularly the case for meteorological and statistical data. However, the analysis also showed that a significant number of the examined data do not yet meet the HVD requirements. Fulfilling precisely these conditions would ensure that relevant data is, technology-wise, more accessible and reusable. Additionally, to fully realise further value-added potentials associated with an HVD-compliant data provisioning, closely related, further requirements must also be fulfilled. These include good documentation of the data *and* interfaces, good usability of the API itself, and a reliable performance of the API. From a technology perspective, data providers should generally strive for such qualitative improvements.

In economic terms, however, the expected benefits from such efforts must be weighed against potentially substantive additional investments. In this context, it is particularly critical that substantive value-added potentials can already be unlocked if data is available free of charge, via download, and subject to open licensing terms. Accordingly, in the conducted workshops, some stakeholders questioned whether a technologically more complex and more costly data provisioning via API is a justifiable, efficient expense - specifically in relation to a bulk download. But according to this study, the conclusion that such costs are necessarily prohibitive cannot be readily drawn only from the existence of requirements that are usually associated with API-based data provisioning, such as comprehensive documentation, good metadata management and the implementation of a scalable database architecture. Today, these requirements should already be met by high-quality data offers via download. Interestingly, the detailed analysis in this study points to considerable implementation deficits among German data providers. Instead, where these conditions are already met by data providers, the additional investment required for API-based data provisioning should also be lower. Eventually, thus, the argument that an API-based, free-of-charge data provisioning according to HVD-criteria is too costly for data providers may, if at all, only carry weight if, for example, the ongoing operation of any indispensable infrastructure itself is very costly. This is particularly the case for use cases that depend on the transfer of *very*, i.e. truly big data, such as some geo and satellite data. Where such data sets are already openly available and no clear, additional potential for added value is discernible, the option of an HVD classification should be examined with particular care. Any such evaluation must, however, also consider the possibility that a provisioning of data via API also enables a more targeted, selective querying of databases, potentially reducing the scope of transferable data for many requests. As a result, data usage barriers would effectively be lowered for many users, leading to the unlocking of higher potential added value.

The implementing act should therefore not only specify clearly which datasets are meant to become HVD, but also detail exactly how HVD should be provided. In doing so, the limits of the PSI Directive must be observed. This implies, for example, that no obligation to archive old data records or to prepare entirely new ones must applied, i.e. not yet existing datasets must follow either directly or indirectly from such specifications. Where this is necessary to support the effective reusability of designated HVD, relevant adjustments may also be required in sectoral regulations at the European or national level.

Lastly, in relation to sectoral regulations on the EU level, particularly the INSPIRE and ITS Directive, the PSI Directive only serves as a minimum standard. Any further specifications of this minimum standard through implementing acts must account for existing sectoral regulations.  Accordingly, PSI implementing acts must reflect the technical and organisational requirements of the existing sectoral regulations in such a way that, where this would provide little benefits, duplicate structures and efforts on behalf of data providers are avoided.  In the area of mobility, this is specifically true for data that fall already into the scope of the ITS Directive. If such data were to be qualified as HVD, the ITS regime would not be changed significantly and the requirements for (certain) private companies to provide certain data would still apply.

### Recommended data sets

To conclude, the following table lists the data sets that were examined in the six thematic areas. All examined datasets are assumed to offer at least minor potential for added value if they were to be classified as HVD. An HVD-classification can thus be recommended, if the PSI Directive is applicable to the relevant data. Depending on the concerned datasets, a classification as HVD may however necessitate different types of interventions. The second column in the table below lists datasets whose HVD-classification would not require further interventions. These data sets that are indisputably covered by the PSI Directive, their classification as HVD adds value (because it would offer economic benefits at no significant cost), and they are already machine-readable and provided via API. The third column lists data records that offer significant potential but whose classification as HVD would also necessitate further interventions. However, this does not imply that these data are generally unsuitable as HVD. Rather, it is specifically important in these cases to deploy economic and legal policy instruments in a targeted manner, ensuring an optimal interplay of any technical, legal and economic interventions. The last column lists data sets that were deemed to fall outside the PSI Directive’s scope during the analysis.

| Thematic Area | Classification as HVD requires **no** intervention | Classification as HVD requires intervention(s)[1] | Datasets out of scope of PSI Directive |
|---|---|---|---|
| Geospatial | Geo_05 rivers; Geo_06 national and local data | Geo_01 cadastral data (E); Geo_02 postcode regions_alt (2 digits only) (T); Geo_03 standard land values (T,E); Geo_04 standard ground value (T); Geo_07 orthoimagery (E); Geo_08 digital surface models (L,E); Geo_09 digital terrain models (E) | Geo_02 Postcode routing data (5 digits) (R) |
| Earth observation and environment | Erd_02 air quality data; Erd_03 noise mapping; Erd_09 CORINE Landcover Germany; Erd_14 earthquakes; Erd_18 soil condition; Erd_19 satellite imagery | Erd_01 air emissions (T); Erd_04 water quality (T); Erd_05 groundwater quality (T); Erd_06 public water supply (T); Erd_07 energy consumption (T); Erd_08 waste balance (T); Erd_10 land use (T); Erd_11 forest condition (T); Earth_12 logging (T); Earth_10 soil areas (T); Earth_11 forest condition (T); Erd_12 logging (T); Erd_13 flooding (T,L); Erd_15 nature conservation areas (T); Erd_16 biodiversity (T,L); Erd_17 fishing quotas, imports & exports (T) | - |
| Meteorological | Met_01 numerical weather forecast for Germany and Europe; Met_02 weather warnings and advance information on municipal level; Met_04 Grid of quarterly means of air temperature for Germany; Met_05 regionalised climate projections; Met_06 2m Temperature at RBSN Stations; Met_07 radar composite RV | - | - |
| Statistics | Sta_01 population statistics; Sta_04 gross domestic product; Sta_06 health statistics; Sta_07 school statistics; Sta_08 income statistics; health statistics; Sta_07 school statistics (pupil enrolment); Sta_08 private household income statistics | Sta_02 federal budget data (T); Sta_03 rate of unemployment (T,L) | Sta_05 Ifo-Business climate Index (R) |
| Companies and company ownership | - | Unt_01 companies’ register (free access) (T,L); Unt_02 companies’ register (chargeable access) (T,L,E); Unt_04 names of shareholders (T,L) | Unt_03 transparency register (R) |
| Mobility | Mob_09 shipping infrastructure and facilities | Mob_01 traffic signs (L,E); Mob_02 road network (L,E); Mob_03 electric vehicle charging stations map (T,L); Mob_04 cycling infrastructure (L,E); Mob_05 public transport data for busses and trains (L,E); Mob_06 public transport timetable / schedule (L); | - |

[1] T = Data not provided in machine readable format and/or via API.
L = Legal applicability or exception from PSI Directive uncertain or significant legal follow-up adjustments required.
E = Classification as HVD entails costs equal to or greater than the expected benefits.

| | | | |
|---|---|---|---|
| | | Mob_07 train station data (T,L); Mob_08 quantities carried and transport performance by mode of transport (T); Mob_10 wave measurements in the North Sea and Baltic Sea (T); Mob_11 broadband penetration (T) | |