Journal Title
XX(X):1–11

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

# A geo-referenced micro-data set of real estate listings for Spain's three largest cities

**Abstract**

This article presents an open data product with large geo-referenced micro-data sets of 2018 real estate listings in Spain. These data were originally published on the idealista.com real estate website. The observations were obtained for the three largest cities in Spain: Madrid (n = 94,815 observations), Barcelona (n = 61,486 observations), and Valencia (n = 33,622 observations). The data sets include the coordinates of properties (latitude and longitude), asking prices of each listed dwelling, and several variables of indoor characteristics. The listings were enriched with official information from the Spanish cadastre (e.g., building material quality) plus other relevant geographical features, such as distance to urban points of interest. Along with the real estate listings, the data product also includes neighborhood boundaries for each city. The data product is offered as a fully documented R package and is available for scientific and educational purposes, particularly for geo-spatial studies

**Keywords**

Housing prices; hedonic price analysis; idealista.com; geo-referenced data; point-level data; open data; Spain

## Introduction

Interest in the characteristics of the housing market and housing prices has been a growing area of research in recent decades, generating a vast amount of theoretical and empirical literature. Including the spatial component to analyze the real estate market and incorporating geographic variables has significantly improved the understanding of this

market: to this end, it is essential to have information/data at the point level. Therefore, it is becoming common for spatial analysis of urban environments to be use micro-data, geo-referenced as points (López et al., 2015). In some cases, researchers have had to resort to webscraping processes to obtain data for research (e.g., Gupta et al., 2022; Arbia and Nardelli, 2020; Li et al., 2019; López et al., 2015). Alas, webscraping is a chancy process prone to download errors, missing data, duplicated records, etc. Furthermore, researchers do not always share their webscrapped datasets, which limits reproducibility of their research. As we witness a growing interest in openness and reproducibility in geographical data science (Arribas-Bel et al., 2021a; Páez, 2021; Brunsdon and Comber, 2021), it becomes increasingly urgent to have open data products to support research (Arribas-Bel et al., 2021b).

Some researchers have already responded to this need for publicly available, geo-referenced datasets to support open, reproducible research. A few such datasets are now available to support the analysis of real estate markets, but they are sometimes geo-referenced at the level of large geographical zones, such as Fuerst and Haddad (2020), an open dataset that includes $n = 4,201$ property prices geocoded to the level of nine regions in England and Wales. Other datasets are geo-coded as points, including Bonifaci and Copiello (2015), which includes $n = 1,042$ observations for Padua, in Italy; Del Giudice et al. (2018), who share a dataset with $n = 576$ observations relating to rental prices in Naples, Italy; and Solano Sánchez et al. (2019) present a dataset with $n = 1,623$ daily rental prices in Seville, Spain.

To contribute to the growing inventory of international micro-data sets of real estate markets, this paper introduces an open micro-data set of geo-referenced dwelling listings. The data have been provided by the Idealista company* and contain information about 189,923 dwellings located in Spain's three largest cities. To date, this data product is one of the largest open geo-referenced micro-data set of the housing market in the world. The most similar in terms of geographical disaggregation and sample size is the dataset of Song et al. (2021) which includes transactions for four cities in South Korea, namely Busan ($n = 61,152$), Daegu ($n = 32,363$), Daejeon ($n = 21,114$), and Gwangju ($n = 25,984$). It is certainly the largest in Spain.

The data set has been supplied directly by Idealista, and therefore is clean and free of download errors. However, in order to be able to share publicly the data we have slightly modified some variables by applying some random noise which will not bias the main results derived from its usage.

This micro-data set can be used to benchmark new analytical methods in a reproducible fashion. Applied and theoretical researchers on real estate mass appraisal and valuation methods might use this dataset to canonically compare the performance of their proposed comparable and hedonic models, among others. The data can also be used to study the segmentation of housing submarkets and related topics such as the impact of suburban areas on house prices. The listings have been enriched with official information from

*Idealista is the major real estate listing website in Spain, and present in other southern european countries as Italy and Portugal

the Spanish cadastre along with other relevant geographical features, such as distance to urban points of interest. In any case, the data might be easily extended by spatially joining other datasets that contain information at administrative, census tract, or postal code levels.

The data set is distributed as an `R` package, named {idealista18}, which can be accessed from a public Github repository[†]. The open data product is made available under the Open Database License. For transparency, we also share the randomisation process applied to the original data in the aforementioned Github repository.

## Data description

The open data product {idealista18} is an `R` package composed of nine objects, three objects for each of the three main Spanish cities: Barcelona, Madrid, and Valencia. For each city, dwelling listings, neighborhood polygons, and a set of points of interest have been included in the package.

The data provider (idealista.com) is a leading real estate portal in Spain, on par with Fotocasa. Smaller portals include Habitaclia and Milanuncios, the latter focusing almost exclusively on individual (i.e., non-professional) advertisers. In September 2021, according to data from Similarweb (a site specialized in sites' traffic volume comparison), there were a total of 103 million page views on real estate portals in Spain. Of this, 94% of traffic was concentrated on four portals, of which idealista was the most important, followed by Fotocasa, Habitaclia, and Pisos.com. Traffic is also highly concentrated, and idealista.com was the the leader by far with 58.6 monthly million visits (57% of the total traffic) compared to its closest competitor with 19.9 million visits (19.3 % of total traffic).

As a result of its share of advertisments, idealista.com covers fairly well all segments of the Spanish market, including both individual and professional advertisers. This dataset includes information about listing prices, and therefore represents the market situation from the perspective of asking prices. This is a necessary compromise in the present case, since actual sale prices are not publicly available, and the information can only be accessed by paying high fees to Colegio de Registradores. However listing prices reflect quite well the (transaction) reality of real estate markets, and correlations between idealista listings and transaction prices can be established see Banco de España. Official transactions and asking prices can be taken as complementary and are of great interest when studying asking-transaction price gaps or the relation between listing site demand variables (i.e., ad contacts or ad views) and price gaps.

To provide some context about the coverage of the {idealista18} dataset Table 1 shows the number of listings with respect to the total residential stock in each city in 2018. As seen in the table, the number of listings ranges between 6.1% of the total number of properties (in Madrid) and 8.1% (in Valencia). Information from Instituto Nacional de Estadistica[‡] shows that the number of listings in the {idealista18} package correspond to

[†] https://github.com/paezha/idealista18
[‡] https://www.ine.es/jaxiT3/Tabla.htm?t=6150&L=1

81.3% of recorded real estate transactions in Barcelona, 80.8% in Madrid, and 91.1% in Valencia.

**Table 1.** Total properties and transactionst three Spanish cities. Year 2018

| City | Total properties (P) | Total transactions (T) | Listings (L) | L/P | T/L |
|---|---|---|---|---|---|
| Barcelona | 789,740 | 56,012 | 61,329 | 7.8% | 81.3% |
| Madrid | 1,545,397 | 76,603 | 94,802 | 6.1% | 80.8% |
| Valencia | 416,004 | 30,615 | 33,593 | 8.1% | 91.1% |
| Total | 2,751,141 | 163,230 | 189,724 | 6.9% | 86.0% |

**Sources**

Total properties (P): Ministerio Español de Hacienda y Función Pública

Total transactions (T): Instituto Nacional de Estadística

The following subsections describe each object. A full description of the data is available in the help section of the package. We have tried to the best of our ability to comply with FAIR principles regarding research data (Wilkinson et al., 2016): upon publication, the dataset has a persistent digital object identifier; publication as a data article makes the data findable; the data and metadata are packaged together and protocols for help files in the R ecosystem mean that documentation is easily searchable; distribution as an R package means that only open software is needed to access the data; and a public repository documents all data processes followed to generate the distributed open data product.

## *Dwelling listings*

The dwelling listing of each city includes a set of characteristics for each dwelling published on the idealista real estate website as an ad. The dwelling listing has been included in the {idealista18} package as a simple features (sf) object (Pebesma, 2018) with point geometry in latitude and longitude. The name of each sf object with the list of dwellings is the name of the city followed by '_Sale' (e.g., Madrid_Sale). Each data object includes a total of 42 variables and the complete set of listings corresponding to the four quarters of 2018. Table 2 shows the number of dwelling listing ads included in the data set for each city and quarter. The record counts for each city are: 94,815 listings for Madrid, 61,486 for Barcelona, and 33,622 for Valencia. Note that the same dwelling may be found in more than one period when a property listed for sale in one quarter was sold in a subsequent quarter. The variable ASSETID, included in the sf objects, is the unique identifier of the dwelling.

| City\Quarter | First | Second | Third | Fourth | Total ads |
|---|---|---|---|---|---|
| Barcelona | 17826 | 7951 | 12375 | 23334 | 61486 |
| Madrid | 21920 | 12652 | 15973 | 44270 | 94815 |
| Valencia | 9305 | 4655 | 5644 | 14018 | 33622 |

**Table 2.** Number of dwelling listing ads for each city and quarter.

Each record of the dwelling listing contains a set of indoor characteristics supplied by the advertisers on the Idealista website (e.g., price, surface area, number of bedrooms, basic features, etc.), including the location of the dwelling. The exact location has been masked, as described in Section Anonymizing the data set). Table 3 lists the main indoor variables with a short description and the mean value of each variable. The dwelling listings were enriched with a number of additional attributes from the Spanish cadastre (Registro Central del Catastro, 2021). The cadastral information is described in Table 3, with the prefix CAD in the variable name. The cadastral features were assigned by applying the features of the nearest parcel to the coordinates. The year the dwelling was built (CONSTRUCTIONYEAR) given by the advertiser was revised since the year of construction is entered on the website by users, and it is therefore subject to errors and incomplete data (40% missing data). To resolve this issue, an alternative variable (CADCONSTRUCTIONYEAR) was included, assigning the cadastral construction year from the nearest cadastral parcel whenever the value was outstanding (date was after publication date or year of construction was before 1500) or when the value supplied by the advertiser was missing.

Additionally, the distance of each dwelling to several points of interest was included in the sf object: distance to the city center, distance to the closest metro station, and distance to a major street (La Diagonal for Barcelona, La Castellana for Madrid, and Blasco Ibañez for Valencia). The last rows of Table 3 show the mean values of these variables.

| Variable | Sort Description | Barcelona | Madrid | Valencia |
|---|---|---|---|---|
| PRICE | Asking price | 395770.58 | 396110.11 | 199678.31 |
| UNITPRICE | Asking price per m^2 (euros) | 4044.86 | 3661.05 | 1714.54 |
| CONSTRUCTEDAREA | Surface (m^2) | 95.46 | 101.40 | 108.95 |
| ROOMNUMBER | Number of bedrooms | 2.86 | 2.58 | 3.07 |
| BATHNUMBER | Number of bathrooms | 1.52 | 1.59 | 1.59 |
| CONSTRUCTIONYEAR | Construction year (advertiser) | 1952.58 | 1964.69 | 1969.43 |
| CADCONSTRUCTIONYEAR | Construction year (cadastre) | 1952.19 | 1965.70 | 1970.55 |
| CADMAXBUILDINGFLOOR | Max build floor | 6.85 | 6.38 | 7.04 |
| CADDWELLINGCOUNT | Dwelling count in the building | 28.56 | 39.19 | 36.83 |
| CADASTRALQUALITYID | Cadastral quality. 0 Best-10 Worst | 4.31 | 4.85 | 5.34 |
| DISTANCE_TO_CITY_CENTER | Distance to city center | 2.80 | 4.49 | 2.09 |
| DISTANCE_TO_METRO | Distance to subway station | 0.27 | 0.48 | 0.64 |
| DISTANCE_TO_(MAINSTREET) | Distance to major street | 1.77 | 2.68 | 2.07 |

**Table 3.** List, sort description, and mean of the main quantitative variables included in the dwelling listing for the three Spanish cities. See the help section in the **idealista18** R package for details and formal definitions. Some variables have been excluded from this table to save space. Check the full list in the **idealista18** package.

In addition to the variables listed in Table 3, the sf object includes a set of dummy variables with information about the basic characteristics of the dwelling. Table 4 shows the more relevant variables included in the sf object.

| Variable | Sort Description | Barcelona | Madrid | Valencia |
|---|---|---|---|---|
| HASTERRACE | =1 if has terrace | 0.33 | 0.36 | 0.25 |
| HASLIFT | =1 if has lift | 0.74 | 0.70 | 0.79 |
| HASAIRCONDITIONING | =1 if has air conditioning | 0.47 | 0.45 | 0.47 |
| HASPARKINGSPACE | =1 if has parking | 0.08 | 0.23 | 0.17 |
| HASNORTHORIENTATION | =1 if has north orientation | 0.13 | 0.11 | 0.13 |
| HASSOUTHORIENTATION | =1 if has south orientation | 0.31 | 0.24 | 0.19 |
| HASEASTORIENTATION | =1 if has east orientation | 0.24 | 0.20 | 0.25 |
| HASWESTORIENTATION | =1 if has west orientation | 0.16 | 0.15 | 0.15 |
| HASBOXROOM | =1 if has boxroom | 0.12 | 0.26 | 0.13 |
| HASWARDROBE | =1 if has wardrobe | 0.30 | 0.57 | 0.53 |
| HASSWIMMINGPOOL | =1 if has swimmingpool | 0.03 | 0.15 | 0.07 |
| HASDOORMAN | =1 if has doorman | 0.08 | 0.25 | 0.05 |
| HASGARDEN | =1 if has garden | 0.04 | 0.18 | 0.06 |
| ISDUPLEX | =1 if is duplex | 0.03 | 0.03 | 0.02 |
| ISSTUDIO | =1 if is studio | 0.02 | 0.03 | 0.01 |
| ISINTOPFLOOR | =1 is on the top floor | 0.02 | 0.02 | 0.01 |
| BUILTTYPEID_1 | =1 if is new contruction | 0.01 | 0.03 | 0.03 |
| BUILTTYPEID_2 | =1 is second hand to be restored | 0.17 | 0.19 | 0.13 |
| BUILTTYPEID_3 | =1 is second hand in good condition | 0.82 | 0.78 | 0.83 |

**Table 4.** List of dummy variables, sort description, and ratios of dwellings with the specific characteristics. See the help section in the **idealista18** R package for details and formal definitions. Some dummy variables have been excluded from this table to save space

## *Neighboorhood polygons*

The second data object included in {idealista18} is the neighborhoods in the three cities as polygons. There is an sf object for each city with the name of the city and the suffix '_Polygons'. Figure 1 shows the quantile maps of the number of dwellings in the listing for the different neighborhoods in the three cities. The neighborhoods are based on the official boundaries but slightly changed by Idealista[§]. In practical terms, we can assume they are the same since the website combines areas when there are few ads for that area. In the case of Madrid, they combined four areas into two.

There are a total of 69 neighborhoods in Barcelona, 135 in Madrid, and 73 in Valencia. The sf object includes a unique identifier (LOCATIONID) and the neighborhood name (LOCATIONNAME).

The total number of dwellings is available from the Spanish cadastre aggregated by district (districts are groups of neighborhoods). Figure 2 shows the percentage of listed dwellings relative to the total number of dwelling by district in each of the three cities. This gives a sense of how active residential real estate markets were in different parts of each city in 2018.

[§]There are two criteria used to make this division. If an area is small enough and similar enough to another, the two areas are merged, and, if the official area is not homogeneous, it is divided into a series of new polygons.

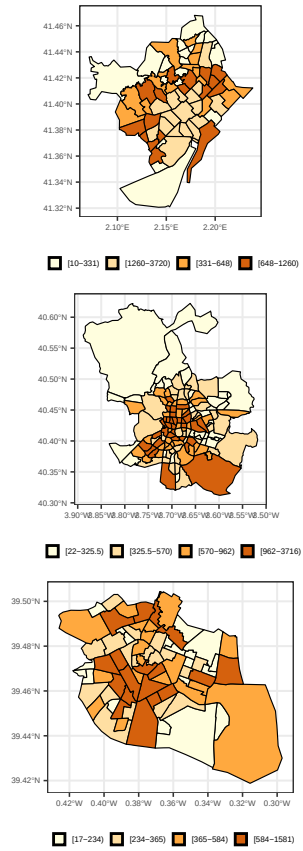


**Figure 1.** Quantile maps of the number of dwellings in each neighborhood. Boundary for Barcelona (Left), Madrid (Center), and Valencia (Right).

### *Points of Interest*

The last data object included in the package is a set of Points of Interest in each city as an object of the class list. The name of the list includes the name of the city with the suffix '_POIS'. These lists include three elements: (i) the coordinates of the city center, the central business district; (ii) a set of coordinates that define the main street of each city; and (iii) the coordinates of the metro stations.

## Anonymizing the data set

To comply with Spanish regulations, two variables were slightly modified to provide anonymity. A masking process was applied to asking prices and location (coordinates).

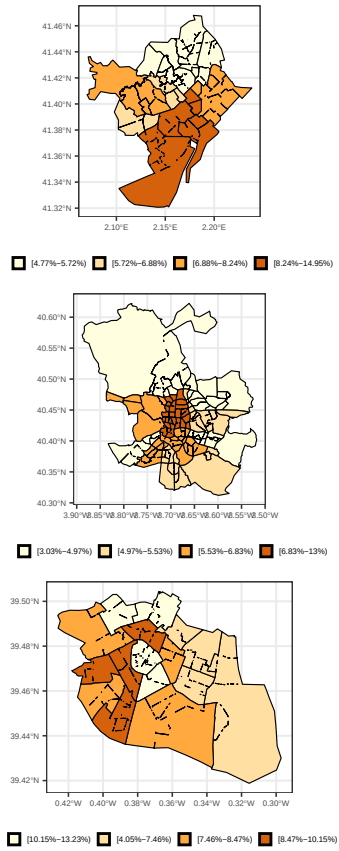


**Figure 2.** Quantile maps of the percentage of listings relative to total number of dwellings by district. Boundary for Barcelona (Left), Madrid (Center), and Valencia (Right).

In terms of the asking prices, the original values were obfuscated with the addition or subtraction of a random percentage of their original values, ranging from -2.5% to +2.5%. Since asking prices are usually multiples of 1,000, after the first price modification, the prices were aligned to multiples of 1,000.

With respect to the location of the dwelling, a spatial masking process was implemented to maintain the spatial properties of the original data set. The coordinates of each listing were displaced using a stochastic procedure. The listings were recorded using coordinates contained in maximum and minimum displacement circles, as shown in Figure 3 (left). To preserve inclusion in a neighborhood, the spatial masking procedure was constrained to ensure that the masked coordinates remained in the original neighborhood of the listing.

**Data:** all idealista listings
**Result:** all idealista listings with masked coordinates
initialization;
**for** *each listing L* **do**
  take geographical location of L as $(X, Y)$ **repeat**
    take a random angle $\alpha$ from 0 to 360 degrees take a distance $R$ as a random value from 30 to 60 meters determine a new point $(X', Y')$ calculated as a point located $R$ with the angle $\alpha$
  **until** *this stop condition*;
  set $(X', Y')$ as the new location for the listing L
**end**

**Algorithm 1:** Coordinate displacement process for anonymisation purposes

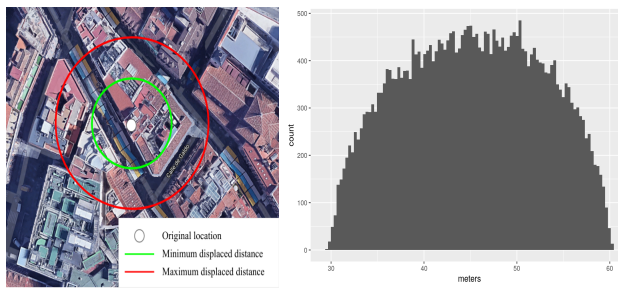


**Figure 3.** (Left) Masking coordinates. Spatial range. (Right) Coordinate displacement in meters (Valencia)

Algorithm 1 iteratively displaces the coordinates of each listing with a minimum distance and a maximum distance with the restriction that the new coordinates do not fall into a different neighborhood. This ensures that neighborhood attributes are preserved.

Figure 3 (right) shows the histogram of the displacements in meters for all the listings in the city of Valencia. The average distance between the original and masked coordinates is 45 meters.

## Conclusion

This paper describes a data product of a geo-referenced micro-data set of Spain's three largest cities. This is an excellent data product to help understand the complex mechanisms related to the housing market and housing prices. Researchers can apply hedonic models with spatial effects, identifying housing submarkets or machine learning techniques (e.g., Rey-Blanco et al., 0). The data product can also be used for educational proposes and teaching activities.

## Declaration of Competing Interest

Author One and author Two are employed by Idealista. They have been granted permission to share the data presented in this article. None of the authors have financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## Acknowledgments

The authors wish to thank Alessandro Galesi for their support in the paper revision and Juan Ramón Selva for collecting and cleaning the spatial data. This work has been partially funded by the Spanish Ministry of Economy and Competitiveness Grants PID2019-107800GB-100

### References

Arbia G and Nardelli V (2020) On spatial lag models estimated using crowdsourcing, web-scraping or other unconventionally collected data. *arXiv preprint arXiv:2010.05287* .

Arribas-Bel D, Alvanides S, Batty M, Crooks A, See L and Wolf L (2021a) Urban data/code: A new EP-b section. *Environment and Planning B: Urban Analytics and City Science* 48(9): 2517–2519. DOI:10.1177/23998083211059670.

Arribas-Bel D, Green M, Rowe F and Singleton A (2021b) Open data products-a framework for creating valuable analysis ready data. *Journal of Geographical Systems* 23(4): 497–514. DOI: 10.1007/s10109-021-00363-5.

Bonifaci P and Copiello S (2015) Real estate market and building energy performance: Data for a mass appraisal approach. *Data in Brief* 5: 1060–1065. DOI:https://doi.org/10.1016/j.dib.2015.11.027.

Brunsdon C and Comber A (2021) Opening practice: supporting reproducibility and critical spatial data science. *Journal of Geographical Systems* 23(4): 477–496. DOI:10.1007/s10109-020-00334-2. URL https://dx.doi.org/10.1007/s10109-020-00334-2.

Del Giudice V, De Paola P and Forte F (2018) Housing rental prices: Data from a central urban area of naples (italy). *Data in Brief* 18: 983–987. DOI:https://doi.org/10.1016/j.dib.2018.03.121.

Fuerst F and Haddad MFC (2020) Real estate data to analyse the relationship between property prices, sustainability levels and socio-economic indicators. *Data in Brief* 33: 106359. DOI: https://doi.org/10.1016/j.dib.2020.106359.

Gupta A, Van Nieuwerburgh S and Kontokosta C (2022) Take the q train: Value capture of public infrastructure projects. *Journal of Urban Economics* : 103422.

Li H, Wei YD, Wu Y and Tian G (2019) Analyzing housing prices in shanghai with open data: Amenity, accessibility and urban structure. *Cities* 91: 165–179. DOI:10.1016/j.cities.2018.11.016.

López FA, Chasco C and Gallo JL (2015) Exploring scan methods to test spatial structure with an application to housing prices in m adrid. *Papers in Regional Science* 94(2): 317–346. DOI: 10.1111/pirs.12063.

Páez A (2021) Open spatial sciences: an introduction. *Journal of Geographical Systems* 23(4): 467–476. DOI:10.1007/s10109-021-00364-4.

Pebesma E (2018) Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10(1): 439–446. DOI:10.32614/RJ-2018-009.

Registro Central del Catastro (2021) https://www.sedecatastro.gob.es/ .

Rey-Blanco D, Arbués P, López FA and Páez A (0) Using machine learning to identify spatial market segments. a reproducible study of major spanish markets. *Environment and Planning B: Urban Analytics and City Science* 0(0): 23998083231166952. DOI:10.1177/23998083231166952.

Solano Sánchez MÁ, Núñez Tabales JM, Caridad y Ocerin JM, Santos JAC and Santos MC (2019) Dataset for holiday rentals' daily rate pricing in a cultural tourism destination. *Data in Brief* 27: 104697. DOI:https://doi.org/10.1016/j.dib.2019.104697.

Song Y, Ahn K, An S and Jang H (2021) Hedonic dataset of the metropolitan housing market – cases in south korea. *Data in Brief* 35: 106877. DOI:https://doi.org/10.1016/j.dib.2021.106877.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 'T Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J and Mons B (2016) The fair guiding principles for scientific data management and stewardship. *Scientific Data* 3(1): 160018. DOI:10.1038/sdata.2016.18.