Sophie's Case Study

Background: Sophie conducts her final-year research in tourism, and to complete it remotely, she gathers diverse datasets from public institutions, utilities, transport providers, and private companies.

1. The challenge lies in determining whether the collected data qualifies as Open Data. By definition, Open Data [1]must be freely accessible, reusable without legal or technical barriers, and redistributable without restrictions. However, Sophie's sources vary significantly in their level of openness. Data locked behind private logins like electricity bills, or gathered via web scraping or Python Beautiful Soup scripts from semi-restricted sites, may not qualify as open in star-based classification[2]. In the case provided, data can be described by specific metrics like publicity, source, open data rules, and legality.

Classification of Sophie's Data Sources

Data Source	Public / Private	Is it Open Data?	Reason	Legality & possible consequences
City council	Public	Yes,1- 2 stars	Published on municipal portals (maybe CKAN/DataHub), PDF are not fully machine-readable but still open)	Legal and safe use. Licensed for public consultation and research. No legal risks.
Municipal water company	Public utility	Yes, 2-3 stars	Aggregated consumption statistics, structured and reusable.	Legal, safe use. Falls under transparency regulations, no risks if anonymised.
Municipal electric company	Public utility	Yes, 2-3 stars	Publicly shared data, (maybe / often via Socrata or DataHub).	Legal, safe use. Maybe part of the Open Government Data initiatives.
Private electricity companies	Private	No	Access requires individual client logins, restricted.	Illegal if bypassed. Accessing without consent same as breach of contract and potential criminal liability (GDPR[3] / Computer Misuse laws).

[1] https://opendatacharter.org/principles/

[2] https://medium.com/transparent-data-eng/5-levels-of-data-openness-a-long-road-from-pdf-to-lod-e10cf36b41c9

[3] https://gdpr.eu/what-is-gdpr/#:~:text=The%20General%20Data%20Protection%20Regulation,to%20people%20in%20the%20EU.

Residents' utility bills	Private individuals	No	Contain personal, sensitive data (name, address etc.)	High legal risk. Use without anonymization breaches GDPR, could lead to fines.
AMTRAK API	Private company but open API	Yes, 3 stars	API provides structured, non-proprietary data.	Legal if used within API license. Breach only if data is resold or used commercially without permission.
AirEuropa	Private company but open API endpoints	Yes, 3 stars	Publicly accessible structured data.	Legal if API terms respected. Risk of contract violations if scraped beyond terms.
Rental car companies	Private	No or semi-open	Only accessible by request, not universally published.	Legally usable if company consent is explicit. Risk if republished without authorisation.
Vacation rental platform	Private company with open portal	Yes, 4 stars	Structured, linked, often exposed via APIs.	Legal for research if under open license. Risk if personal client data is exposed or reused commercially without permission.

Key insights:
- Government and municipal data are fully legal, encouraged by open-data laws.
- Private company APIs are legal only if used under published terms of service.
- Residents' bills and private company logins are illegal to use directly, unless anonymised or with explicit informed consent!
- If data is used unfairly and incorrectly in terms of service, big consequences like GDPR fines or criminal liability under IT/data protection law can be applied!

Conclusions: No, not all the data provided in the case can be considered as Open Data, and in the table and key insights above is shown all necessary parameters for the data to be considered as Open Data.
In this way, Sophie must maintain differentiation between fully open, semi-open, and private data to ensure her study respects both the theoretical framework and privacy regulations.

2. As in the case described, the data provided by the city council and the city planning department - mainly belongs to the one start category according to Tim Berners-Lee's [4] classification (which can be seen in the table above). Just because most of the information is presented in PDF documents and graphic files, which are readable by humans but not by machines, makes it difficult to reuse it in automated analysis. If these same datasets were also published in structured formats such as Excel or ODS, they could receive two or more stars. Therefore, if they were provided in formats such as CSV or JSON, they could receive 3-4 or more stars. To go beyond this, the data would need to be linked to other datasets using URIs (like linking census data to national or regional demographic repositories), which would raise the rating to 4 or 5 stars. With the proper structuring and linking, the data provided by the city council could rise to 3 - 5 stars.

Tim Berners-Lee's 5 star Open Data model

*	Data is made available on the web under an open license. Format can be anything (PDF, image, docs etc.) Pros: accessible to anyone. Cons: difficult for machines to read or process.
**	Data is structured and still under an open license Format can be XLD / ODS. Pros: more usable that PDFs, can be opened in software. Cons: relies on proprietary tools of software.
***	Data is structured and in non-proprietary format. Format can be: CSV, JSON, XML. Pros: machine-readable, easily processed with open tools. Cons: datasets are still isolated, no automatic links to other data.
****	Data is structured, non-proprietary, and uses URIs to link elements. Format can be CSV, JSON, XML etc. + URI. Pros: enables interoperability, which means that different datasets can "communicate to each other".
*****	Data is fully linked to other open datasets on the web. Format can be CSV, JSON, XML etc. + URI + Linked Open Data Pros: creates a "web of data" (linked data) that allows richer analysis and new isights.

[4] https://www.researchgate.net/publication/307845029_Tim_Berners-Lee%27s_Semantic_Web

3. The data provided by AMTRAK and the airline companies is available via API and published in CSV files. This means that the data is structured, non-proprietary, and, therefore, machine-readable, which allows it to be classified as 3 stars according to Tim Berners-Lee's classification. As we mentioned before, the data classified as 3 star is accessible, reusable, and easily processed by using modern tools like Pythom libraries, which include Pandas, Beautiful Soup etc., making it very valuable for Sophie's research.

Classification of AMTRAK and Airline data

Data source	Current format	Stars	Reason	Action to uprafe
AMTRAK API	CSV via API	3	Machine readable, non-proprietary, isolated.	To add URIs for stations/routes. Linking to other datasets like population, tourism etc.
AirEuropa American Airlines	CSV via API	3	Accessible, reusable, structured, interlinked.	Publishing with linked identifiers for airports, regions, flights etc. Connecting with open datasets like tourism, demand emissions etc.

Because AMTRAK and airline data are structured and non-proprietary, they still hold 3 stars. For an upgrade to 4 stars, they need URIs and linkages to external datasets.

4. The vacation service company's dataset received four stars because the data is available in structured, non-proprietary formats (CSV, JSON) and is linked to other internal datasets via API. This special connection between different resources via URI raises the data rating from 3 to 4 stars. To upgrade the data up to 5 stars, a company must ensure not only internal links between datasets, but also external Open Data repositories. For example, accommodation or season demand can be linked to official EU Parliament Tourism databases or the EU / USA Airport Association. By creating these cross-network connections, the data becomes part of a broader Linked Open Data ecosystem, enabling deeper analysis of the data across different platforms.

5. In the table "Classification of Sophie's Data Sources", we have mentioned that utility bills from residents contain personal and sensitive information and can lead to a great fine or even to criminal consequences. This data is usually covered by legal authorities in terms of data privacy and protection regulations. For example, the Author has mentioned in the table "Classification of Sophie's Data Sources" the GDPR (General Data Protection Regulations in Europe), which sets strict rules for handling personal information. To comply with legal and ethical standards, Sophie should implement security and privacy measures:

- Collect only what she needs (excluding names or account)
- Secure storage (encrypted format, restricted access, password-protected folders)
- Anonymise the data (masking, hashing, aggregation)
- Ask for consent (explicit, written consent from residents before using)
- Delete data after use (retention policy)
- Follow GDPR and local laws (data misuse, avoiding fines or academic sanctions)

Sophie must treat residents' bills with the highest privacy safeguards, focusing on anonymisation, consent, and secure handling. This will allow Sophie to transform private information into safe, research-based and usable data.

6. It's always good to dive into the world of modern and niche data. Sophie can look for additional datasets or even enhance her research with predictive techniques. For example, she could use live mobility data to map visitor flows in and out of the popular locations. Nowadays, the use of social media analytics is also popular, which can provide visitor sentiment insights, showing where tourist locations are rising before official or government statistics catch up. With the help of predictive models using open datasets, she can simulate future scenarios and test "what if" cases. This will make her project not only modern with the use of past statistics, but Sophie will also be forecasting future tourism metrics, such as tourist pressures or overall tourism image.

7. More than less, the consultancy's work would be classified as a data management decision, but at a higher and more strategic level. Just because they not only collect datasets, but also create synergy between different studies, ensuring connectivity, management, and decision support. This increases the role of data from raw statistics to a knowledge that directly influences urban development. Once the consultancy aggregates these data, it's no longer Open Data. The information becomes a proprietary asset with added value. After that, the sampled dataset can be enhanced with forecasting, like AI-based analytics, predictive algorithms, or even geospatial modelling. All of this transforms the consultancy's data into a decision support system that can be monetised as a data product or market service.

Bibliography:

- https://gdpr.eu/what-is-gdpr/#:~:text=The%20General%20Data%20Protection%20Regulation,to%20people%20in%20the%20EU.
- https://medium.com/transparent-data-eng/5-levels-of-data-openness-a-long-road-from-pdf-to-lod-e10cf36b41c9
- https://opendatacharter.org/principles/
- https://www.consilium.europa.eu/en/policies/data-protection-regulation/
- https://www.dlapiperdataprotection.com/?c=US
- https://www.pwc.com/m1/en/publications/2025/docs/data-monetisation-and-beyond-redefine-the-economics-of-data.pdf
- https://www.researchgate.net/publication/307845029_Tim_Berners-Lee%27s_Semantic_Web