**FAIR Data Assessment Homework - *DUE Friday Oct 3, 2025***

Throughout our foundations module we have discussed the FAIR data principles to various degrees.  Understanding what these principles are is important to understanding if other data sets you are using are FAIR, and for you to be aware of how to make your data FAIR throughout your research career.  In this assignment you will be given a dataset and asked to assess its FAIRness in each category: Findable, Accessible, Interoperable, and Reusable.

**Instructions:**

* For each of the 10 questions in the FAIR-Aware Assessment you previously filled out ([https://fairaware.dans.knaw.nl/Links to an external site.](https://fairaware.dans.knaw.nl/)), ***describe how*** the GEO dataset with accession GSE276609 meets or does not meet the criteria for being FAIR and provide an example if possible.  **Create a MS Word document for this assignment.**  Click on the Info icons next to each question in the FAIR-Aware Assessment to get more details about what each item means.
  + URL: [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276609Links to an external site.](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276609)

1. Dataset assigned a globally unique persistent and resolvable identifier when deposited with a data repository

Yes, the data repository is Gene Expression Omnibus (GEO) which provides an Accession number (unique stable identifier) for a series or project as well as for each sample. The GEO website states that investigators can cite GEO submissions using the accession number GSExxxx. The file names in supplementary are also stable (.tar and .gz)

1. Provide discovery metadata to make datasets easier for others to use

Yes, samples have additional information about their origin (eg patient derived xenograft), disease type (eg basal like triple negative breast cancer), organism for platform sequencing (homo sapiens). There is also descriptive information on the dataset like the date, summary, and keywords. Additionally, the supplementary files includes information about the samples. The Design section provides additional details on the data making it easier for humans to process the data. There are also explicit links to other research outputs that were used.

1. Data repository for data should make metadata available in format readable by machines and humans

The data is readable to machines and humans like the .soft and .txt files. Some supplementary materials require special tools and programs like the Loupe Browser. For the .cloupe files, and the .tar files may require extraction or some pre-processing before it can be machine read. GEO has NCBI E-utilities (API) but does not have the REST API or OAI-PMH systems that are listed through FAIR.

1. Access to data may need to be controlled. Metadata has license information so data can be reused

The deposited data is associated with a publication. It is recommended to add a data availability statement which was included in the publication with the GSE number and github link. The deposited data in GEO is open access and anyone on the internet can access it.

1. Metadata is available over time even if data is no longer accessible

GEO policy is that only GEO staff can remove records from the database if the investigators contact them via email. GEO also allows for updates and edits after submission. These considerations make it easier to maintain metadata availability over time.

1. Metadata describing data should use controlled vocabularies

Yes, the GEO submission follows NCBI taxonomy language such as the use of Homo sapiens. Other nomenclature for example for diseases or tissue type were not as explicitly noted. There are not explicit classes given in the examples like UBERON but the vocabulary used is descriptive.

1. Provenance information about collection and generation of data should be included in metadata

Provenance information is included, for example for the sources of data generate and collection, the specific equipment (Illumina HiSeq 4000 and NextSeq 2000) was listed. The data is also disaggregated by the contributors involved in the author contributions section of the publication. There is a data versioning process as some of the data was reanalyzed from a previous publication and project, and this process was clear based on the information the authors provided.

1. Metadata describing data should follow specifications of community endorsed standard

It is unclear whether the data deposited meets the specifications of community endorsed standard.

1. Data should be deposited in a file format that is open and supported by data repository for long term preservation

The files the data is stored in are .tar, .csv, and .txt which are all open and supported for long term preservation.

1. Keeping data FAIR over time requires professional data curation and digital preservation

The dataset is deposited in Gene Expression Omnibus, which is a government led open source data repository that is regularly maintained by National Library of Medicine staff to ensure the data is curated and preserved.