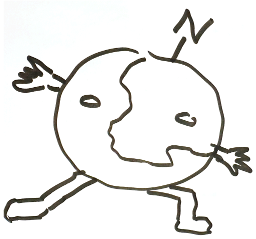
DCN Planetears
Geospatial Data Curation:
an introduction



Module 2: Ethics and GIS Instructor Notes

[This handout accompanies a slidedeck with the file name “2_Ethics_and_GIS_Data”]
2024-02-08 1000

Lesson Plan: Ethics and GIS Data (30 mins)

Slide 2: Objectives:

- Identify 3 examples where there would be ethical concerns with sharing geospatial data
- Identify 3 recommendations for action when there are ethical concerns

Slide notes: see slides (15 mins)

Activities: (15 mins)
Discussion prompt: Have you encountered any situations working with sensitive GIS data? What were your ethical concerns and what actions did you end up taking?

Notes for remaining Slides:

3. **Ethical checks**
   Today we are going to talk about three ethical scenarios that may come up when preparing geospatial research data for sharing. Re-identification, sensitive data (beyond protected health information) and permission to share. For each we will look at a few examples and then discuss recommendations you might have for the researchers.

4. **Re-identification**
   Spatial information (like addresses) are considered personally identifiable pieces of information. There are a lot of excellent specialized trainings and primers about how to handle human subjects data, we are not going to get too deep into this topic. We would just like to emphasize that location is very identifiable. One of the things that makes geospatial research so powerful is that you can look across data layers and analyze them together with location as a common point. But that also means that you can’t think about the dataset in isolation, you have to be thinking about what other sources someone might have access to that could help re-identify subjects. Some examples where you would need to be cautious include if there is any protected health information

for an area that has a smaller population. Another example would be data about traffic accidents. If you have when and where a crash took place, even with minimal demographic information, it could be combined with news articles and police reports to identify people.

5. **Re-identification Prevention: Real Life Example** (Optional more detailed example)
    a. In May 2023 a dataset that recorded traffic crashes, trauma center location, patient demographic metadata, and, alcohol and drug use, was submitted to a repository.
    b. During the CURATE(D) Check step, the curator became concerned that patient demographic metadata and the crash and/or trauma center location data could lead to re-identification when paired with publicly available news reports or police reports of crashes.
    c. The curator read through the final report, and realized that a number of demographic variables that the trauma centers collected were not analyzed by the researchers and were not needed to support research conclusions, and were not needed for reproducibility or replication.
    d. Curator suggested researchers remove the variables **Year**; **Month**; **Race**; **Ethnicity (Hispanic)**; **Comorbidities**; and **Complications** to greatly reduce re-identification risk.
        i. In states such as Iowa, the non-white population is a small percentage of the population, and the **Race** and **Ethnicity** variables could make it easier to re-identify people, some who could become targets for discrimination or deportation.
        ii. Some medical **Comorbidities** or **Complications** are quite rare, and could lead to re-identification, and possible insurance or genetic discrimination.
        iii. The inclusion of **Month** and **Year** of crash could help to match event to news reports.
        iv. Taking all of these variables together, a re-identified driver or crash victim who was using prescription or non-prescription drugs or alcohol, or were or were not wearing a seatbelt or a helmet at the time of the crash could find themselves discriminated against for employment, health coverage, housing, and other social needs.
    e. The curator also suggested that while the **Age** of the people involved in the crashes (as driver, passenger, collision victim, etc.) was binned to reduce re-identification, the age bin 18-20 was too small (3 years) while other bins were 10 to 20 years wide.

6. **Recommendations when re-identification is a concern**
   Again there is much more robust and detailed on this topic that is available, but a couple of key things:
   - You might recommend that the researcher remove any unnecessary attributes that are not part of the analysis.
   - For protected health information, researchers should aggregate data to geographic units that have populations of more than 20,000 people or remove geographic identifiers for any areas that are too small. (The Safe Harbor Method promoted by HIPAA explains how to achieve this.) (Street addresses are removed. − Zip codes are truncated to only show the first three digits. − Zip codes, cities, towns and counties with < 20,000 residents are obfuscated. The zip codes will display as ‘000’. Cities, towns, and counties will display ‘Other/Unknown’.)
   - You also might recommend placing this data in a repository that has restricted access.

7. **Sensitive Data (beyond PHI)**
   Even when datasets do not directly represent human subjects, locations may be sensitive. Some examples: archeological sites or movement of endangered species. Publishing the data in these cases could reveal sensitive locations to vandals, poachers, or tourists. Also resources on lands of indigenous communities should be treated with CARE -- publishing data may infringe on tribal data sovereignty and tribal authority over sacred names/sites.

8. **Recommendations when sensitive locations are concern**
   What are some recommendations you can offer a researcher when sensitive locations are concerned? You might suggest that they obscure or randomize spatial information. There are a few different ways they might do this.

9. **Offset geospatial locations**
   Offset geospatial locations. With this approach, you maintain relationships between points, but disconnect them from real-world locations. If the researcher is studying animal movement with respect to land cover, they could offset the datasets to an unrelated location maintaining only the spatial relationships between data points.
   a. The example on the slide shows the movement of a wolf in Alberta - in an area where hunting wolves is legal. Publishing location specific information could harm the animal being tracked.

10. **Aggregate data to a larger area**
    Another approach is to aggregate data to a larger area. In this example, instead of publishing specific locations of where endangered or threatened species had been found, observations are aggregated together to indicate the broader critical habitat.
    a. Another example would be if water samples are being used to describe the state of a watershed, exact locations may not be important to the results and could be replaced with a watershed identifier.

11. **Introduce noise or randomly distribute points within a larger area**
A third approach that is sometimes used is introducing noise, fabricating data based on the characteristics of the original, or randomly distributing points within the areas of analysis. So if you are doing an analysis at the level of census tracts, you can randomly move points around within this area.

12. **Recommendations when sensitive locations are concern (2)**
There is a lot that you can do to obscure spatial information. But sometimes the specificity of locations is essential to the research.
    a. For example, if you are looking at how spatial environments impact health outcomes - if you generalize things and decrease the resolution of the data when sharing the findings won't be reproducible. [Specific example: Studying the impacts of a superfund site on health - if you aggregate the data to the nearest geographic area with 20,000 people the signal may be completely lost]
    b. Obscuring information also can dramatically reduce the reuse potential. Going back to the wolf example, let's say a future researcher wanted to explore movement with respect to something other than land cover. If you have removed the spatial information, that won't be possible. In cases like this, you'd want to suggest that the data be submitted to a repository with restricted access (such as ICPSR).

    Finally you might request that the researcher contact communities represented or affected by the data. This could include signed permission to share or evidence of a governance mechanism.

13. **Permission to share**
A third ethical topic is around permission to share data. This is not something that is unique to geospatial data, but it is something that comes up all of the time. Because geospatial data offers the ability to look for patterns across multiple phenomena at a location, researchers very often supplement data they gather directly with existing data published by government entities or research partners. A lot of this content (demographics, satellite imagery, land cover) is going to be publicly available through federal agencies with no limitation on use or redistribution. But data can also be coming from project partners or proprietary sources. Examples include:
    a. Multiple data layers from outside sources combined together to identify the best location for conservation actions
    b. Columns appended to a spreadsheet of information collected by another graduate student in a lab

14. **Recommendations when data sharing permissions are a concern**
When data sharing permission are a concern, you will want to recommend:
    1. Checking that the authors have permission to share all of their data
    2. Removing data for which permission has not been granted and replacing it with information about how to acquire it.
    3. Adding additional authors and attribution as needed.

    We are going to talk more about establishing provenance and documenting data sources in the Metadata module. We just wanted to bring it up here because if you can catch data sharing concerns as early as possible, that is optimal.

15. **Resources for later**
This slide has some links with more resources and information about ethical considerations for health, human subjects, and indigenous data. Including:

- De-identification of Protected Health Information in Accordance with HIPAA https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
- Human Subjects Data Essentials https://conservancy.umn.edu/handle/11299/216579
- the CARE Data Principles, for data related to Indigenous Peoples and Interest https://conservancy.umn.edu/handle/11299/256919

16. **Small Group Activity**
    Before we transition to the next topic, we would love to hear about any situations you might have encountered working with sensitive GIS data? What were your ethical concerns and what actions did you end up taking?