Comparison of US wheat yield with middle-aged marriages in Salzburg

A Data Management Plan created using DMPonline

Creator: Helmuth Breitenfellner

Affiliation: Other

Template: Digital Curation Centre

ORCID iD: 0000-0003-4523-0945

Project abstract:

This project analyses the correlation between yield of wheat in the United States on one side, and the number of marriages by middle-aged (35-44) men and women in Salzburg. The source of the project can be found on GitHub: https://github.com/helmuthb/dmp-exercise1. All software and the report is licensed under MIT license. The data is licensed as by the data providers. • CC0: US Wheat Production Timetable (USDA 2019) • CC BY 4.0: Age of partners at marriage (Salzburg 2019) This data experiment is the result of exercise 1 of the lecture “Data Stewardship”.

Last modified: 22-04-2019

Comparison of US wheat yield with middle-aged marriages in Salzburg

Data Collection

What data will you collect or create?

The data collected is provided by the state of Salzburg ("Land Salzburg"), and by the US Department of Agriculture. The data from Salzburg is the number of husbands and wives, per age group, within a given year. To be counted in the data the husband has to be registered in Salzburg. The data from USDA is the acres used, yearly production, and yield (production per acre) for wheat per year in the United States. From this data, a combined dataset is created, showing both husbands and wives and yield per year.

The source data is already aggregated data, with the sum per year in both datasets.

The total volume of source data is about 1.5 megabytes. The combined data is about 6 KB.

The format used is CSV (for the data from Salzburg) and Excel (the only format provided by the USDA for this data).

The usage of CSV (for the Salzburg and the combined data) is explicitly chosen to allow long-term usage. CSV is an established format which has support in a variety of tools.

The usage of Excel for the USDA data is rather unfortunate. Excel is not an open format and long-term usability is not guaranteed. Since this is the only source format available for this data it had however to be taken.

Source data is provided by Salzburg using CC BY 3.0, and by USDA using CC0. Combined data is licensed as CC BY 3.0 (requiring citation of the Salzburg data source).

How will the data be collected or created?

The data will be collected via download of the current version (as of April 2019) from the providers websites.

Version control (GIT) will be used throughout the project.

Documentation and Metadata

What documentation and metadata will accompany the data?

Dublin Core metadata will be provided together with the data.

A README file is provided for reproducing the results, and a comprehensive report will detail the activities and results.

Ethics and Legal Compliance

How will you manage any ethical issues?

The source data is already aggregated and anonymized, no ethical considerations are necessary.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

As one of the source data (Salzburg Marriage data) is CC-BY-4.0 licensed, this license shall be used for the combined data as a derived work.

Storage and Backup

How will the data be stored and backed up during the research?

Data will be stored and backed up using GitHub.

How will you manage access and security?

Data will be secured via usual Operating System mechanisms (login via username/password), and transport will be secured using HTTPS and/or SSH (git).

No confidential data is part of the data sets and therefore this basic level of security is sufficient.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

The combined data source, the scripts for the processing and the originally used data are to be preserved and shared.

Foreseeable research includes repetition of the experience with more up-to-date data, and comparisong with other yield (e.g. rye) and/or other countries.

What is the long-term preservation plan for the dataset?

The data will be preserved via Zenodo. To recreate the experiment, a Dockerfile is provided, together with the source of the non-standard R-packages used in the experiment.

Data Sharing

How will you share the data?

Data will be shared using GitHub repository.

It will be available for the general public, based on CC0 (for the USDA data) and CC-BY-4.0 (for the Salzburg and the combined data) licenses.

Are any restrictions on data sharing required?

No restrictions on data sharing are required.

Responsibilities and Resources

Who will be responsible for data management?

Helmuth Breitenfellner will be responsible for implementing, reviewing and revising the DMP.

He will be responsible for all data management activities.

What resources will you require to deliver your plan?

Cost-free repository on Zenodo is the main requirement for the implementation of the Data Management Plan.