PhenoMapping: A participatory visual tool for curation and verification of historical phenological data

Alois Wieshuber [a], Yu Feng [b], Lina Hörl [a], Peter Valena [a]
[a] Bavarian State Archives (GDA)
[b] Chair of Cartography and Visual Analytics, Technical University of Munich (TUM)

Date of submission: 30.10.2024
Required funding: 6 months

Abstract

Phenological data, especially long-term data, is crucial for understanding the impact of climate change. However, due to the lack of digitization, comprehensive phenological data from before the 20th century is highly incomplete, hindering long-term climate and ecological research and leaving this area underrepresented in Earth System Science (ESS). The innovation of this project lies in developing an interactive tool called *PhenoMapping*, which helps digitize and geocode historical records and matches the priorities of scientists' demands and historians' capacity through an online platform. It encourages volunteers' contributions and also allows public users to explore phenological trends from decades and centuries ago. Using an archive collection from 1856 with 7,000 phenological observations as an example, we can demonstrate the value of this tool. Expected outputs include web-based data visualization and transcription tools, along with the acknowledgement of data contributions to existing phenological databases. The expertise of both teams, GDA in handling historical documents and TUM in geospatial data visualization, will support the project's goal of bridging historical and modern phenological data for ESS research.

I. *Introduction*

Phenology studies the timing of natural events, particularly in the annual life cycles of plants, like leaf budding, flowering, and fruit ripening. Phenological data, especially at long-term scales, has proven to be a valuable indicator for current research in Earth System Science, e.g., the effects of global warming on vegetation and growing seasons (Menzel, A. et al., 2006). Current data are generated from a variety of sources and using standardized methods (e.g., DWD Phenological Data). However, scientific data for the period before the 20th century are scarce. These earlier records can only be extracted from analog historical documents, often handwritten with less detailed location information.

The DWD (German Meteorological Service) primarily provides Historical Phenological Database (HPDB) from 1880 to 1941. However, the observations are very incomplete. There are other repositories for historical phenological data as well, such as PPODB or PEP725. The PPODB (Plant Phenological Online Database) contains records collected by the DWD from 1951 to 2009 supplemented by the HPDB, that holds records of 1880-1941. Those were further supplemented by digitizing observations in printed forms, to fill the gaps from 1941 and 1951 and before 1951. Still, data from before 1880 is missing there. In the PEP725 (Pan European Phenology DB), there

are records for Austria as early as 1775, for the Netherlands from 1868, and for Sweden from 1870 onwards, while the earliest records for Germany are from 1951 (PEP725). They both mentioned that it is laborious to harmonize data from historical sources with current standards. Historical records archived in the Bavarian State Archives (GDA) show that forestry office records from the 19th century alone contain numerous phenological observations. These records are in tabular form on paper, providing structured data that are easier to interpret than pure text paragraphs. Such historical observations are highly valuable, as they can provide a baseline for climate research and be used to estimate long-term trends.

However, interpreting and managing these data requires considerable effort. First, the records must be digitized and transcribed. Even the most advanced AI and transcription software struggle to fully recognize the data accurately, making the involvement of experts essential. To obtain results comparable with current data, the observations in historical records must also be mapped to current observation standards. Second, the accuracy of observation locations needs to be addressed, often requiring the use of historical maps to determine precise geographical coordinates and external data for their elevations. Finally, a major challenge in collecting historical phenological data is that it is not easy to identify which locations and time periods have missing data, even though many records may still exist in archives. Because it is difficult to determine priorities, constructing a complete time series becomes challenging to even start.

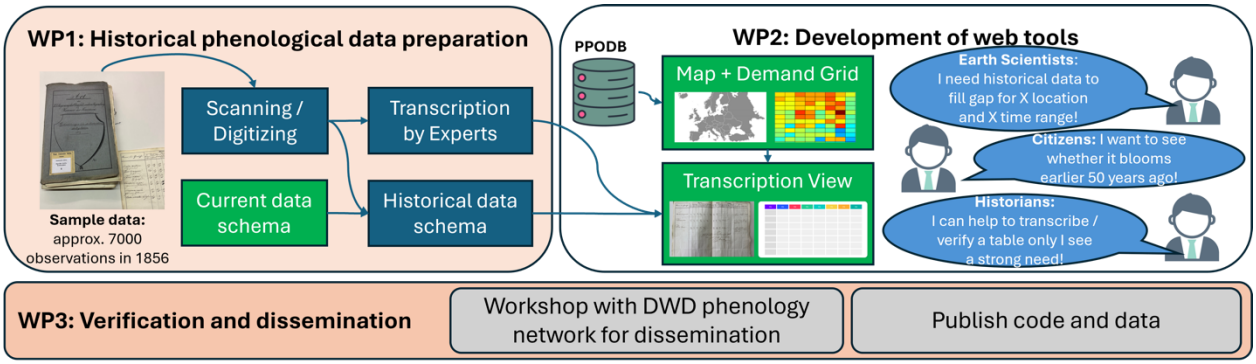
Therefore, we aim to develop a tool called ***PhenoMapping*** that facilitates the interpretation and management of phenological data from historical sources (textual sources or herbaria). The tool can be used by scientists from different backgrounds, e.g., historians who are able to decipher historical writings, or Earth system or life scientists to mark the locations and time ranges with high priority for data completion, to produce valuable datasets for phenology. Furthermore, this project should also ensure that the data and software are accessible and comply with FAIR principles. The GDA team, composed of biologists and historians, provides domain-specific expertise. The TUM team brings experience in visualization and visual analytics, with a recent prior project targeting phenological data visualization. This collaboration ensures comprehensive support, combining field-specific knowledge with technical and visualization expertise.

II. *Incubator Project description*

With the motivation above, the following sample data will be used to demonstrate the potential of the system, which can provide much earlier and well-structured observations compared to existing databases. The **sample source** to be used consists of a file provided by the Amberg State Archive, Government of the Upper Palatinate, Chamber of Finance, District Forestry Office in 1856. The file contains approx. 65 questionnaires for each forest district in the Upper Palatinate, covering the dates for 13 separate developmental stages (i.e., table columns) from a list of 65 different plant species. Each questionnaire is filled with observations of approx. 3-30 species, which amounts to an estimated 7000 single observations in total. With this data as a tester, we propose to implement this project with the following 3 work packages (WP), as illustrated below:

WP1: Historical phenological data preparation (Month 1 – 3, GDA)

- Task 1.1: digitize the sample data as imagery.
- Task 1.2: analyze data schema between historical records and current standards, then create a mapping between them.



- Task 1.3: manual data transcription by experts at GDA, including the metadata, e.g., source, location name, used standard, etc.

WP2: Development of web tools (Month 1 – 4, TUM)

- Task 2.1: development of an interface for the visualization of data points, initially using PPODB as the basis and our own project data as complementary.
- Task 2.2: development of a spatial-temporal grid for researchers marking their interested location and time ranges as demands.
- Task 2.3: development of user transcription and verification interface, where pictural scans of tables with a corresponding table showing experts' transcription for verification or volunteers' transcription from an empty table.

WP3: Verification and dissemination (Month 5 – 6, GDA & TUM)

- Task 3.1: hold a workshop (online) with relevant stakeholders, such as the DWD phenology network and other experts in the field, aiming to introduce this tool and receive feedback.
- Task 3.2: disseminate the code of the tool and publish the data of the example dataset (e.g., contribute to the PPODB).

III. *Relevance for the NFDI4Earth*

The tool allows Earth system and life scientists to identify high-priority locations and time ranges for data completion. Historians or volunteers with the ability to decipher historical writings can contribute by transcribing records based on these priorities, ensuring that the data collection process aligns with both scientific demand and resource availability. Phenological data from the DWD's Climate Data Center (CDC) in the list, i.e., part of the PPODB, will be used in this project.

IV. *Deliverables*

- Digital copies of the sample archived historical file and their transcription
- Code repository of the tool with tutorials, documents, and deployment instructions
- Discussion and feedback collection during the workshop

V. Finance plan

With the work packages scheduled within this project, we apply for funding of EUR 35,800, split between one research assistant (TV-L E13, 6 months 50%, 17,900) at GDA and one research assistant (TV-L E13, 6 months 50%, 17,900) at TUM.

Reference:

Menzel, A., Sparks, T. H., Estrella, N., Koch, E., Aasa, A., Ahas, R., ... & Zust, A. N. A. (2006). European phenological response to climate change matches the warming pattern. Global change biology, 12(10), 1969-1976.