ORD Proposal: opnts

: An Open Framework for Scientific Grade Time Series Datasets

Dr. Matthias Bannert and Minna Heim

1/18/24

Table of contents

# 1. Introduction

|  |  |
| --- | --- |
| Full title | opnts - An Open Framework to Create Scientific Grade Time Series Datasets from Public Data |
|  |  |
| Keywords | Open Data, Official Statistics, Time Series Analysis, Data Literacy, Automation, Machine Readability, Community |
|  |  |

## 1.1 Problem Statement

Opnts bridges the gap between official statistics and academic research. Even though public availability of official statistics is decent in Switzerland, data publications of regional and federal public data providers often lack machine readability. Therefore, these data publications are not convenient for academic researchers and other interested groups.[MAYBE YOU WANT TO ADD SOME EXAMPLES].

In the field of economics, for example, intertemporal comparisons play an important role. This implies regular and timely ingestion of data. Traditional data publications are often unsuitable for the scientific consumer, as modern economic (forecasting) models make use of hundreds or thousands of variables stemming from a large range of datasets. Particularly, researchers who source a large amount of time series data from different providers suffer from the substantial difference in the pace of the digital transformation across providers or even different departments of the same providers.

Though application programming interfaces (APIs) from the official providers have led to great improvements in the machine readability of public data, machine readability is not homogenous across datasets. Opnts mitigates this dynamic problem by not only publishing time series data created from a fully transparent process, but also by providing an open technical processing framework. The opnts framework eases the data sourcing and quality control burden and, at the same time, encourages researchers to contribute time series datasets of their own. In addition, opnts establishes a steady and hands-on exchange between regional and federal data providers and the community of academic users of public data.

## 1.2 Background and Motivation

Though we aim to extend the approach suggested in this proposal to other fields in the future, this *Explore* proposal focuses on our ‘home court’ disciplines of public statistics and empirical economics. We are confident that our strong network in both of these fields will leverage our initial push enough to give the impact of the *opents* project a great chance to spill over to other communities. The following sections introduce our background and explain how the proposed project helps the economics, forecasting and open data communities.

### 1.2.1 Time Series Analysis, Forecasting and Data Science in Economics

Time Series analysis continues to play an important role in many strands of modern empirical economics: Economic forecasting and its evaluation, studies of shock absorption, and many other econometric modeling exercises ask for data that allow intertemporal comparison. Modern models often use hundreds, or even thousands, of time series.

Even though machine readability of publicly available data has improved in Switzerland thanks to great efforts led by the Swiss Federal Statistical Office (SFSO), researchers face a publication mindset that is focused on the wider public and hence on the most recent wave of data. As a consequence, variable names or features may change over time and minor errors may simply get corrected ad hoc or with the next publication — even in datasets that are already available through an application programming interface. In combination with different degrees of machine readability across different providers of public data, the issues mentioned above can lead to two very heterogeneous situations when it comes to automated data ingestion – particularly for those researchers who need a large amount of data from various providers.

Our motivation is to help the time series community *immediately,* while larger organizations face different challenges in their own digital transformation. We intend to do so by not only open-sourcing a framework to create scientific grade time series from official statistics, but also publishing our framework in the most inclusive/transparent way possible: [NOT SURE IF INCLUSIVE IS THE RIGHT WORD HERE] i) We focus on thorough documentation that involves the community; ii) We provide comprehensive application examples, making many thousand proven time series available; iii) We have an infrastructure-as-code approach and share precise information on how to set up an environment that is capable of running our framework.

### 1.2.2 Sustainability through Cooperation: Official Statistics and Public Data

At the same time, our background in regular publication of data-driven indicators and our identification with the official statistics community ensure the sustainability of the *opnts* effort. In addition to the strong empirical focus of its academic research, the KOF Swiss Economic Institute at ETH Zurich has a national mandate to monitor the Swiss economy. In the process, KOF produces forecasts and indicators, many of which rely on publicly available data provided by federal and regional administrations. KOF is also involved (together with the ifo Institute in Munich) in the Joint Economic Forecast in Germany, which serves the German government as a benchmark for future economic development. At the time of composing this ORD proposal, the KOF time series made more than 31’000 time series publicly available in human browsable and machine-readable fashion alike[[1]](#footnote-22).

Because of KOF’s national mandate and the regular production of indicators and forecasts over decades, we have maintained steady exchange with key providers and consumers of economic data, such as the Swiss National Bank (SNB), the Swiss Federal Statistical Office (SFSO), the Secretariat of Economic Affairs (SECO), the State Secretariat of Education, Research, and Innovation (SERI) and the cantonal statistical offices. Through our exchange, we are aware that other institutions have often implemented data sourcing procedures to process publicly available data that are similar to our own processes. Hence, we do not only plan to reduce redundant processing across institutions but also to use source code and example data pipelines as a communication channel to dynamically communicate the needs of the open research community and share a collaboratively developed blueprint working process with public data providers. This type of horizontal and upstream feedback is a core strategy of the opnts project that we validated in talks with these data providers in advance of this proposal. Concrete steps could be the integration of FSO experts into teaching or the use of Swiss data in academic teaching and open-source software packages.

# 2. ORD Project Plan

The following sections explain our approach in greater detail. Though working packages can be seen as stages in general, working packages are not strictly sequential as some activities are continuous efforts. The subsequent chapter gives an overview that shows responsibilities within the team and intensitvity of activities throughout the project.

## 2.1 WP0: Coordination and Planning

**During the starting stage of our project, we coordinate with major public data providers and involve data sources in our planning early on. We do so to establish and maintain active, lively communication between public data providers and the scientific community.**

Publicly available economic data provided by the Federal Statistical Office (FSO), the Secretariat of Economic Affairs (SECO), the Swiss National Bank (SNB), and regional Statistical Offices are important sources for economic research and monitoring of the Swiss economy. To develop an ORD spirit and a community of a data providers and scientific users of the data, it is crucial to involve these public institutions early on. Early dialogue helps to avoid redundancies with existing data community initiatives and to develop a common understanding of *open data*, machine readability and datasets of priority.

One of the core ideas is to leverage the investment into open data by integrating public data into academic teaching to strengthen the connection between public administration and academia. To explore how to effectively use the joint expertise of public data providers, domain experts, data stewards of the ETH domain, and data engineers, we launch a series of monthly meetups early in the process. The meetups are designed to foster inclusive, expert discussions through publicly pre-circulated content. We are working to integrate our meetup series into the academic curriculum at ETH, but we keep the series open to interested participants beyond academia.

## 2.2 WP1: Make Data Processing Framework Inclusive

In Work Package 1 (WP1), we lay the technical foundation for turning KOF Swiss Economic Institute’s existing production process of macroeconomic time series into an open-source framework that creates publicly available, machine-readable scientific-grade time series in fully reproducible fashion. Our approach uses license cost free open source software and aims to publish not only the resulting time series data but also the technical framework and infrastructure setup. Doing so ensures reproducibility and helps to comply with FAIR principles. Thanks to our infrastructure-as-code approach we can publish information on an ideal runtime environment to help others operate their own instance of the *opents* framework.

Because an active community is inevitable to keep the project sustainable, i.e., well maintained, we chose the R Project for Statistical Computing as our main programming language for the implementation of the framework. R is known for its large and inclusive community with its annual user conferences (**inclusivecon?**) and vibrant local communities[[2]](#footnote-27) . Alongside its strong regional and international community, the fact that R is open source and free of license costs lowers the barrier to entrance and contribution considerably. In addition, R offers well-established boiler plating ecosystems like *usethis* (Wickham et al. 2023) and documentation frameworks such as *pkgdown* (Wickham, Hesselberth, and Salmon 2022) or Roxygen (Wickham et al. 2022).

WP1 focuses on the homogenization of the data ingestion process across datasets and data providers. In the first implementation step, we identify common parts of the ingestion process. We bundle these common parts in an R package that forms our core library and fosters reuse of source code. We encourage the community to use our core packages and to contribute by adding further dataset-specific packages. To inspire contributions, we supplement the opents core, based on our long-term experience as consumers of public data, with provider-specific ingestion libraries written in R that cover dataset idiosyncrasies.

|  |
| --- |
| Example of a traditional official data publication of the Swiss GDP. Source: SECO, https://www.seco.admin.ch/seco/en/home/wirtschaftslage—wirtschaftspolitik/Wirtschaftslage/bip-quartalsschaetzungen-/daten.html |

The above screenshot shows a traditional data publication of an official GDP statistic. Its use of multiple worksheets, empty lines and multi-line headers hampers machine readability. Like many regular consumers of the information contained in this spreadsheet, KOF ingest this information immediately after its publication.

|  |
| --- |
| Screenshot of a machine-readable data file, processed by an R-based engine. |

To have a homogeneous, machine-readable representation of the data, we split this information into two files: based on the CSV on the web idea by the World Wide Web Consortium (W3C), we use a long format CSV file for the data and a JSON file for the metadata. The metadata file takes advantage of the JSON format’s ability to handle nested data to store comprehensive, multilingual meta information. Our collaborative approach with the Secretariat of Economic Affairs (SECO) led to a pioneering pilot publication of machine-readable data on the official SECO website: Machine-readable data is published in complementary fashion to the traditional GDP publication[[3]](#footnote-34).

|  |
| --- |
| Excerpt of the corresponding metadata file in JSON format. |

At the end of WP1, we plan to iterate over the resulting datasets and data descriptions and enhance meta information where necessary. We do so in reproducible and transparent fashion by adding all changes to the source code that produces the data and metadata files.

## 2.3 WP2: Publication

*The goal of Work Package 2 (WP2) is to find the platforms that suit our community, data and code components best and facilitate our approach to ORD. We will involve our data consumers and collaborators into this decision-making process. Based on our* open by default\* thinking, we will publish not only the resulting time series data and meta information, but all components of our software framework. In the process, we will work together with the ETH legal service to find the most suitable open source licensing solution[[4]](#footnote-39) for all components of the opents project.\*

The *opents* approach, suggested here, splits data and meta information into two separate text files and formats: a simple CSV spreadsheet and a nested JSON file for multilingual data description. Having two simple text files per dataset allows us to disseminate the resulting time series data on free, standard infrastructure such as GitHub, which is well established in the open-source community. The state-of-the-art rendering of CSV spreadsheets of modern git platforms sets the barrier to exploring and consuming our scientific use time series datasets as low as possible. Hence, the publication of data on GitHub, possibly using git’s Large File Storage (LFS) extension, forms our baseline scenario. Yet, in WP2, we explore the feasibility of other, complementary, regional and international publication channels such as opendata.swiss[[5]](#footnote-40), r-universe or Zenodo.

Technically, the *opents* framework consists of a core R package and several data provider-specific ingestion packages. Again, we see the publication of all these components as open-source libraries on GitHub (or another major Git platform with good visibility) as the baseline form of publication. In addition, WP2 explores which of our components are suitable for a publication on the Comprehensive R Archive Network (CRAN). Publication on CRAN comes with requirements and quality control but opens up our work to a larger audience as endusers face the lowest possible hurdle to install our packages.

|  |
| --- |
| Download of R packages from CRAN is very convenient in all major R environments. The screenshot shows an example of RStudio’s graphical user interface for searching and downloading a package. |

We will also explore more recent and experimental approaches, such as r-universe which helps to improve our reach in the data science community. In WP2, we aim to expose our work to an rOpenSci peer review. rOpenSci shares our values of open and reproducible research through reuse and helps us reach ORD excellence.

## 2.4 WP3: Facilitate Usage and Applications

*The goal of work package 3 (WP3) is to promote usage of the framework and data as well as to encourage others to contribute. In particular we are looking for community contributions such as maintenance of existing datasets, addition of public datasets that have not yet been processed and community activities. At this stage, we will also intensify activities to explore the expansion of* opnts\* into different academic fields.\*

We plan to approach the above goals in three ways: i) integration of opnts data and software into academic teaching aimed at scientists from different disciplines. ii) presentation of opnts at useR! 2024 in Salzburg, Austria[[6]](#footnote-45). iii) participation in local community events iv) hosting of an own event at KOF.

Starting in fall 2024, opnts data and software will be integrated into the Hacking for Science course[[7]](#footnote-46). Originally created for ETHZ D-MTEC PhD students, Hacking for Science is a highly interactive online course whose big picture guidance and hands-on software development approach continues to draw interest from students from a variety of fields. The course last iteration welcomed students from more than 10 different ETH departments as well as guests from outside ETH. We aim to communicate the value of open data to this broad audience and learn which challenges ORD faces in different fields. Currently, we also explore the possibility of running Hacking for Research Assistants – a satellite event for research assistants. In addition our teaching is complemented by the RSEED seminar series started in early 2024 which gives us an additional channel to promote an ORD mindset.

Centered around the submission of opents to useR! 2024, we plan further activity at conferences and community events. At ETH, the Scientific Computing group has recently founded a Research Software Engineering (RSE) group[[8]](#footnote-47) to connect software engineers in science at ETH. Regular RSE events at ETH as well as the vibrant, local open source and opendata community[[9]](#footnote-48) will gives us a chance to present the opnts project, e.g., at the RSE group or the local R user group. KOF hosted useR! in 2021 and is well connected to the R community. This history and connections gives a great chance for a successful application to host an existing special interest conference such as *R in Official Statistics* at KOF. Hosting a well-established, smaller special interest conference allows us to give the entire event an ORD focus push ORD excellence in Official Statistics.

# 3. Impact

It is particularly important to us to not only solve a critical problem in our community, but also to do so in sustainable fashion. We are convinced that aligning opents with the FAIR principles (Wilkinson et al. 2016) is instrumental in establishing our approach in our own community and reaching other research communities from there.

## 3.1 Sustainability in the ETH Domain

Regular and timely data sourcing is mandatory for KOF’s academic and national tasks. Hence, we are fully convinced that pushing an ORD approach at KOF, will help advances ORD practices in empirical economics and official statistics in the long term. The use of the opnts framework in teaching and the economic community helps to leverage digital transformation efforts in official statistics, as future researchers and public servants alike become familiar with Swiss official statistics data and ORD principles. In addition, our inclusive approach helps to keep the community involved, while the idea of modular software packages with their maintainers allows for the passing on of responsibilities in case of changes in personnel. The Comprehensive R Archive Network (CRAN) and the R Project for Statistical Computing have a long tradition at ETH. The ETH domain and its staff contributed substantially to the language and its extension packages, including to CRAN itself[[10]](#footnote-51).

## 3.2 Solving a Critical Problem

Although there has been a certain sense of urgency for digital transformation in the public sector since the pandemic at the latest, change management schedules do not always align with the timeline of individual academic careers or ongoing daily business. While we greatly appreciate advances in the above efforts, we provide an immediate solution to data sourcing to those that need to ingest public data today. In addition, we help identify redundant efforts across institutions and broker usage of common sources, using public money more efficiently. Finally, we help bridge the gap and strengthen the link between public administration and academic research.

## 3.3 Collaboroative Approach

The proposed opnts project is highly collaborative in multiple ways. The technical approach is inclusive and modular. Through open-sourcing and focusing on the core package, we hold ourselves accountable through the community and at the same encourage contribution to the technical core. In our data provider-specific modules, we foster collaboration at the dataset level. Through our own dataset modules, we provide examples that collaborators from other fields can use - along with the documentation - as blueprints to build their own extension packages to process data from their respective fields. Finally, we involved data providers from the very initial phase of the project. Our goal is to use our traditional ties with federal and regional public administration to establish a channel for hands-on two-way feedback between data providers and the research community. In the process, we plan to take part in existing academic and non-academic community events and also explore ways and formats, particularly in a way of teaching that involves students in our collaborative approach.

## 3.4 FAIR Principles

The opnts project works to make its output **findable**. We combine our own experience with the expertise and reach of our partners to find the right outlets to help us reach multiple communities. A high degree of automation allows us to consider and maintain multiple platforms from GitHub to CRAN for our software and from Dryad[[11]](#footnote-55) to opendata.swiss and Zenodo[[12]](#footnote-56) for our datasets.  
In addition, we intend to subject our work to a scientific software review (Ram et al. 2019) by the renowned rOpenSci (Boettiger et al. 2015) organization.

rOpenSci fosters a culture that values open and reproducible research using shared data and reusable software – https://ropensci.org/about/

We make our work **accessible**. With the support of ETH Legal Services, we open-source our software and find appropriate open-source and creative common licenses for our software and data.

We do not compromise when it comes to making our data **interoperable**. We build on the World Wide Web Consortium’s (W3C) idea of CSV on the Web, setting the tone for interoperability. Like CSVW, our data output splits data and metadata into two files: a data .csv and a metadata JSON. While we use standard CSV files for time series to not only allow virtually any programming language to ingest it, but also standard office software to read our data,. Our JavaScript Object Notation (JSON) metadata files extend the basic idea of the W3C. We use a key to refer to data and allow storing comprehensive, multilingual meta-information.

A **reusable** and extendable framework and data pool of macroeconomic time series is a core contribution of the opents project. We do not only provide the data and software as is; we also provide community support on our GitHub page and RSEED chat space. In addition, we use the project as a teaching example and ambassador for the FAIR principles among ETH students.

# 4. Schedule Overview

The following Gantt Chart shows shows the four work packages (WP0: Coordination and Planning, WP1: Homogenization of Data Processing Across Data Sources, WP2: Publication and WP3: Community Activation) and the timeframe in which their activities will be carried out. See Sections 2.1 - 2.4 above for details on the goals and implementation of the WPs. The chart also includes the project activities gives an overview which activity is lead by which collaborators, where PI = Principal Investigator, RSE = Research Software Engineer, and SA = Scientific Assisstant, ZD = Developer of KOF General IT (‘Zentrale Dienste’).

Scroll to the left to see the rest of the table.

# 5. Resources (including costs)

# References

Bannert, Matthias. 2024 forthcoming. *Research Software Engineering - a Guide to the Open Source Ecosystem*. CRC Data Science. CRC/Chapman & Hall.

Boettiger, Carl, Scott Chamberlain, Edmund Hart, and Karthik Ram. 2015. “Building Software, Building Community: Lessons from the Ropensci Project.” *Journal of Open Research Software* 3. https://doi.org/<https://doi.org/10.5334/jors.bu>.

Ram, Karthik, Carl Boettiger, Scott Chamberlain, Noam Ross, Maëlle Salmon, and Stefanie Butland. 2019. “A Community of Practice Around Peer Review for Long-Term Research Software Sustainability.” *Computing in Science & Engineering* 21 (2): 59–65. <https://doi.org/10.1109/MCSE.2018.2882753>.

Wickham, Hadley, Jennifer Bryan, Malcolm Barrett, and Andy Teucher. 2023. *Usethis: Automate Package and Project Setup*. <https://CRAN.R-project.org/package=usethis>.

Wickham, Hadley, Peter Danenberg, Gábor Csárdi, and Manuel Eugster. 2022. *Roxygen2: In-Line Documentation for r*. <https://CRAN.R-project.org/package=roxygen2>.

Wickham, Hadley, Jay Hesselberth, and Maëlle Salmon. 2022. *Pkgdown: Make Static HTML Documentation for a Package*. <https://CRAN.R-project.org/package=pkgdown>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.

1. see also https://ts-explorer.kof.ethz.ch [↑](#footnote-ref-22)
2. Bannert (2024 forthcoming) discusses the importance of communities for the learner, including experience from hosting the useR! Conference 2021 which KOF co-chaired. [↑](#footnote-ref-27)
3. See also: https://www.seco.admin.ch/seco/en/home/wirtschaftslage—wirtschaftspolitik/Wirtschaftslage/bip-quartalsschaetzungen-/daten.html [↑](#footnote-ref-34)
4. see https://choosealicense.com/ [↑](#footnote-ref-39)
5. https://opendata.swiss [↑](#footnote-ref-40)
6. useR! is the annual conference of R users. useR! conference usually draw more than 1’000 international users from a large variety of fields. See also https://events.linuxfoundation.org/user/ [↑](#footnote-ref-45)
7. See also: https://ethz.ch/en/the-eth-zurich/education/innovation/kite-award/kite-award-2022/nominierte-projekte-kite-award-2022/hacking-for-social-sciences.html [↑](#footnote-ref-46)
8. The RSE group is already well connected to the global RSE movement, partiuclarly to the Society of Research Software Engineering in the UK. See also https://rse.ethz.ch/ [↑](#footnote-ref-47)
9. see events such as opendata beer, lovedata week [↑](#footnote-ref-48)
10. see also https://stat.ethz.ch/CRAN/ [↑](#footnote-ref-51)
11. https://datadryad.org/stash [↑](#footnote-ref-55)
12. https://zenodo.org/ [↑](#footnote-ref-56)