gesis

Leibniz Institute
for the Social Sciences

Access to social science research data by an open API

IASSIST 2017 in Lawrence, Kansas*:*
Data in the Middle: The common language of research
May 23-26, 2017
Wolfgang Zenk-Möltgen, Reiner Mauer

Member of the Leibniz Association

Outline

- Social science research data
- Current services at GESIS
- Format conversions
- Provide access
- Build an easy to use API
  - RESTful
  - Documented by OpenAPI
- Future use

Social science research data

- At the GESIS Data Archive …
  - the social science research data have a focus on sociology and political science
  - the data are mainly quantitative rectangular by structure
  - predominant data collection method is survey
- Current expansions are …
  - social media and internet data
  - experimental data
  - georeferenced survey data

gesis Leibniz Institute for the Social Sciences

Social science research data

- A variety of nations

gesis Leibniz Institute for the Social Sciences

Social science research data

- A variety of topics

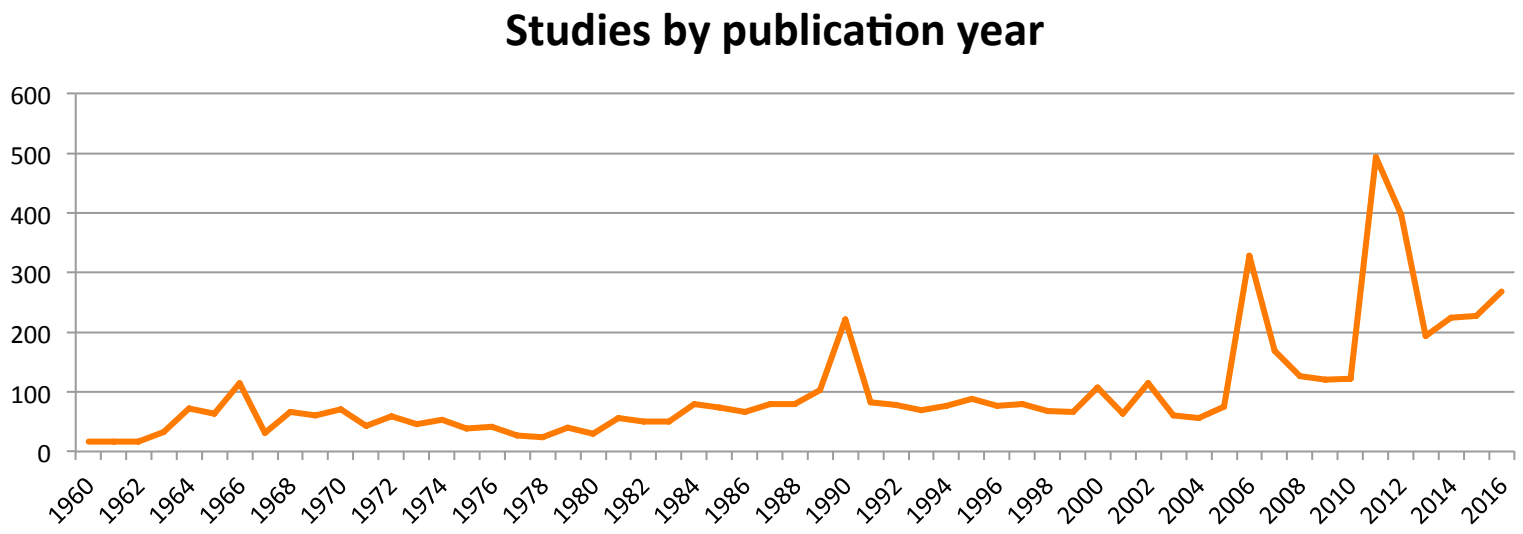
gesis Leibniz Institute for the Social Sciences

Social science research data

- A variety of institutions

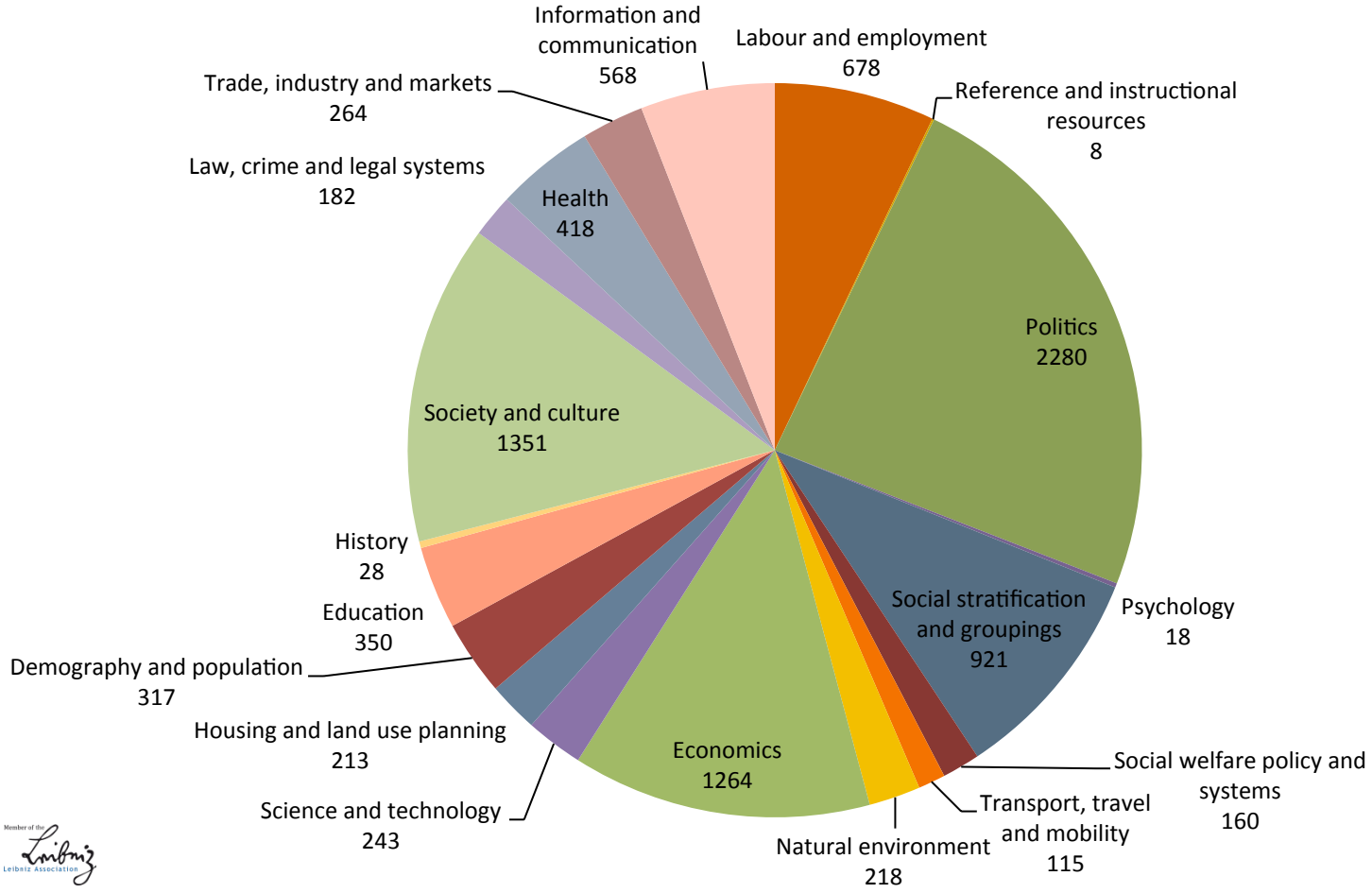
Social science research data

- Increase of published studies per year



Social science research data

- 5767 studies by different topic areas

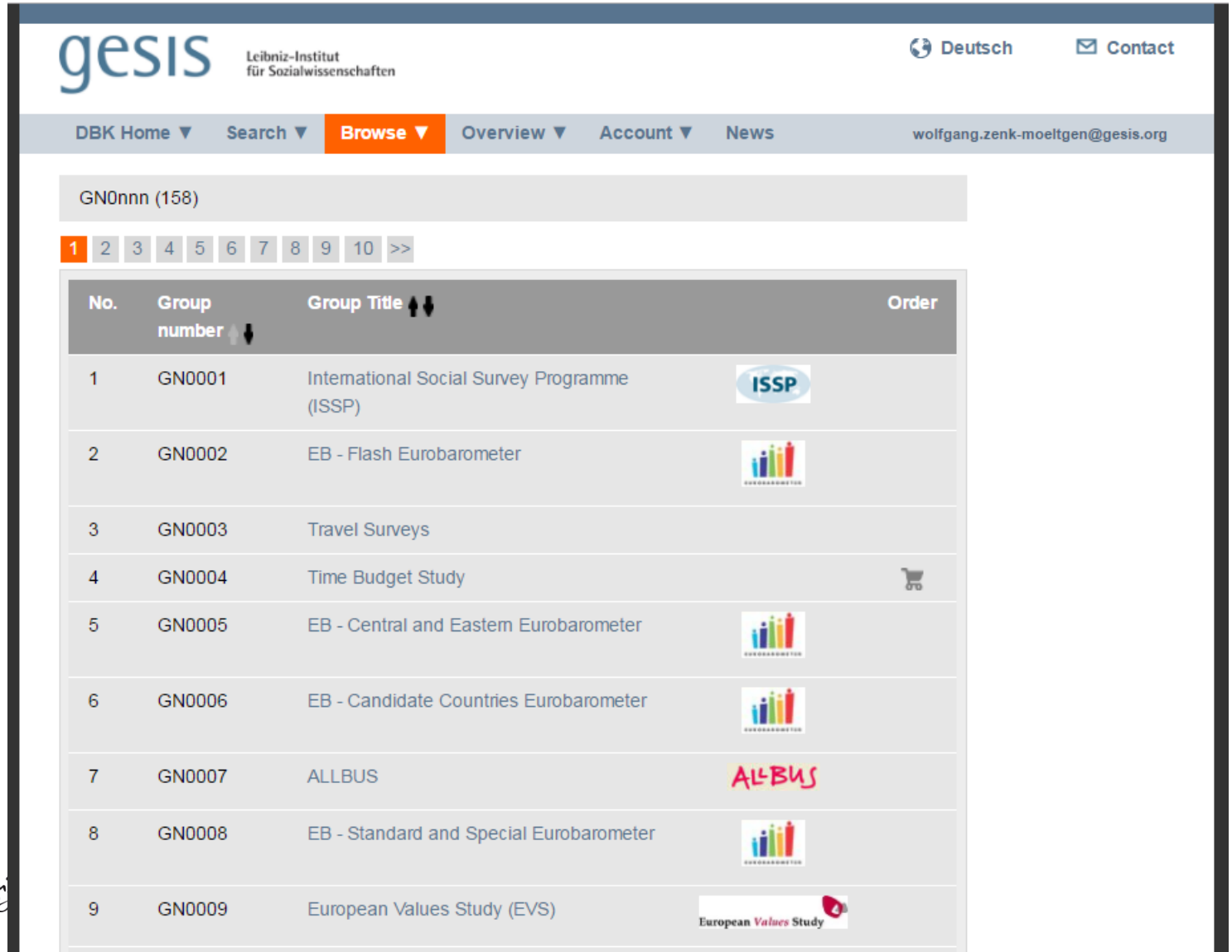


Current services at GESIS

- How do we provide access to data?
- The Data Catalogue DBK
  - Download data sets
  - Order data sets
- datorium
  - Download data sets
- HISTAT
  - Download data

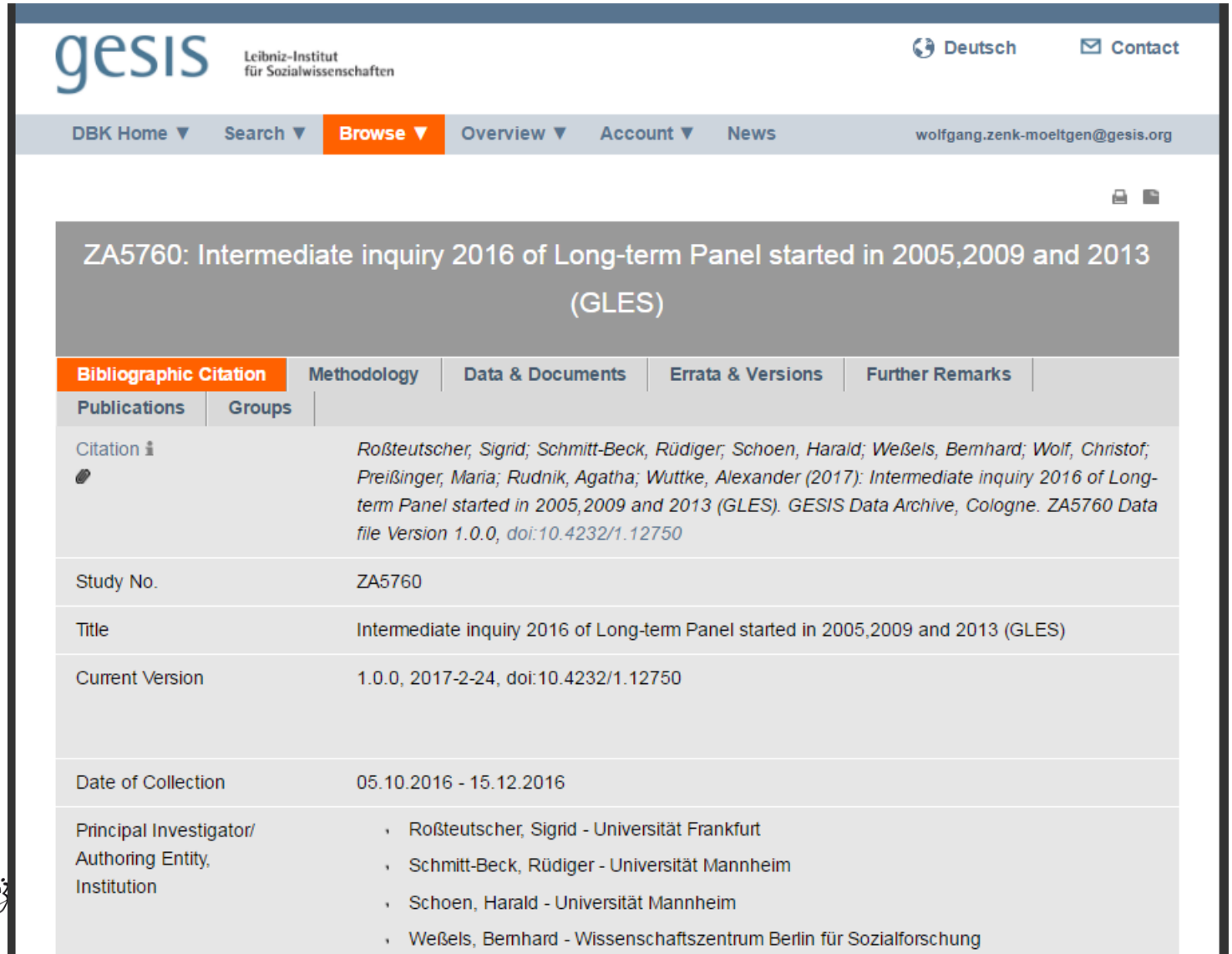
gesis Leibniz Institute for the Social Sciences

Current services at GESIS: DBK



https://dbk.gesis.org/dbksearch/

Current services at GESIS: DBK



gesis Leibniz-Institut für Sozialwissenschaften

Deutsch Contact

DBK Home ▼ Search ▼ Browse ▼ Overview ▼ Account ▼ News wolfgang.zenk-moeltgen@gesis.org

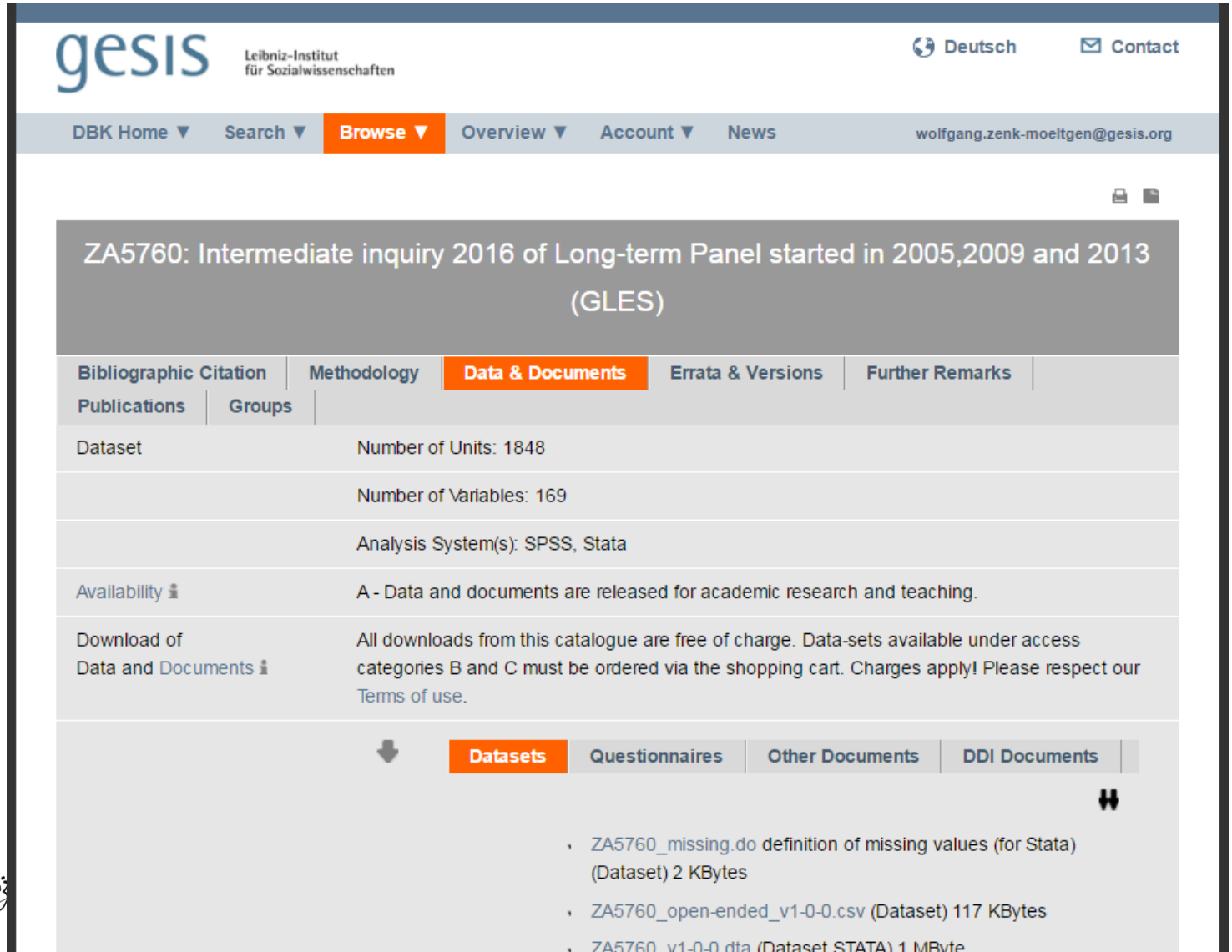
ZA5760: Intermediate inquiry 2016 of Long-term Panel started in 2005,2009 and 2013 (GLES)

Bibliographic Citation | Methodology | Data & Documents | Errata & Versions | Further Remarks | Publications | Groups

Citation	*Roßteutscher, Sigrid; Schmitt-Beck, Rüdiger; Schoen, Harald; Weßels, Bernhard; Wolf, Christof; Preißinger, Maria; Rudnik, Agatha; Wuttke, Alexander (2017): Intermediate inquiry 2016 of Long-term Panel started in 2005,2009 and 2013 (GLES). GESIS Data Archive, Cologne. ZA5760 Data file Version 1.0.0, doi:10.4232/1.12750*
Study No.	ZA5760
Title	Intermediate inquiry 2016 of Long-term Panel started in 2005,2009 and 2013 (GLES)
Current Version	1.0.0, 2017-2-24, doi:10.4232/1.12750
Date of Collection	05.10.2016 - 15.12.2016
Principal Investigator/ Authoring Entity, Institution	• Roßteutscher, Sigrid - Universität Frankfurt • Schmitt-Beck, Rüdiger - Universität Mannheim • Schoen, Harald - Universität Mannheim • Weßels, Bernhard - Wissenschaftszentrum Berlin für Sozialforschung

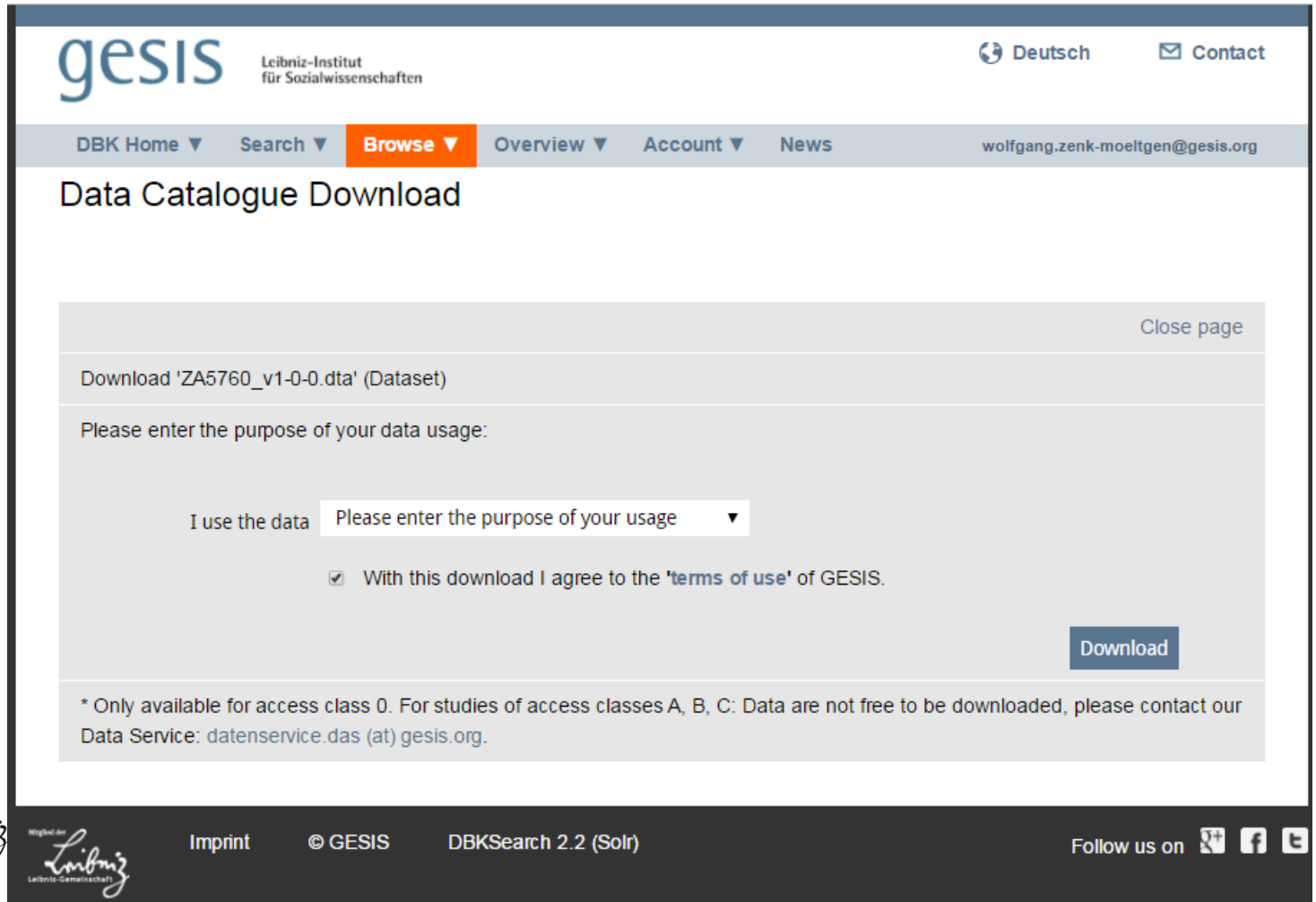
gesis Leibniz Institute for the Social Sciences

Current services at GESIS: DBK



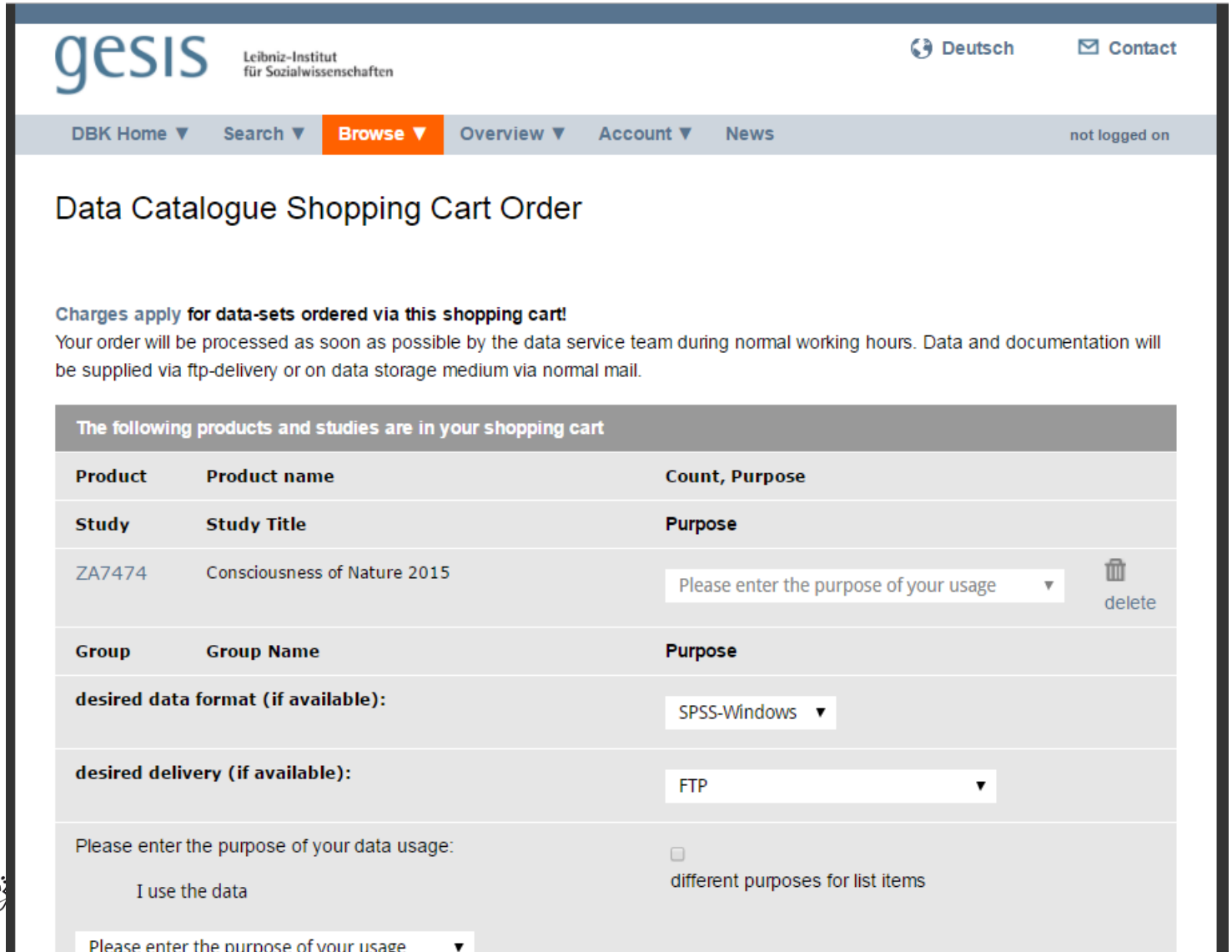
Current services at GESIS: DBK

- DBK data file download and orders



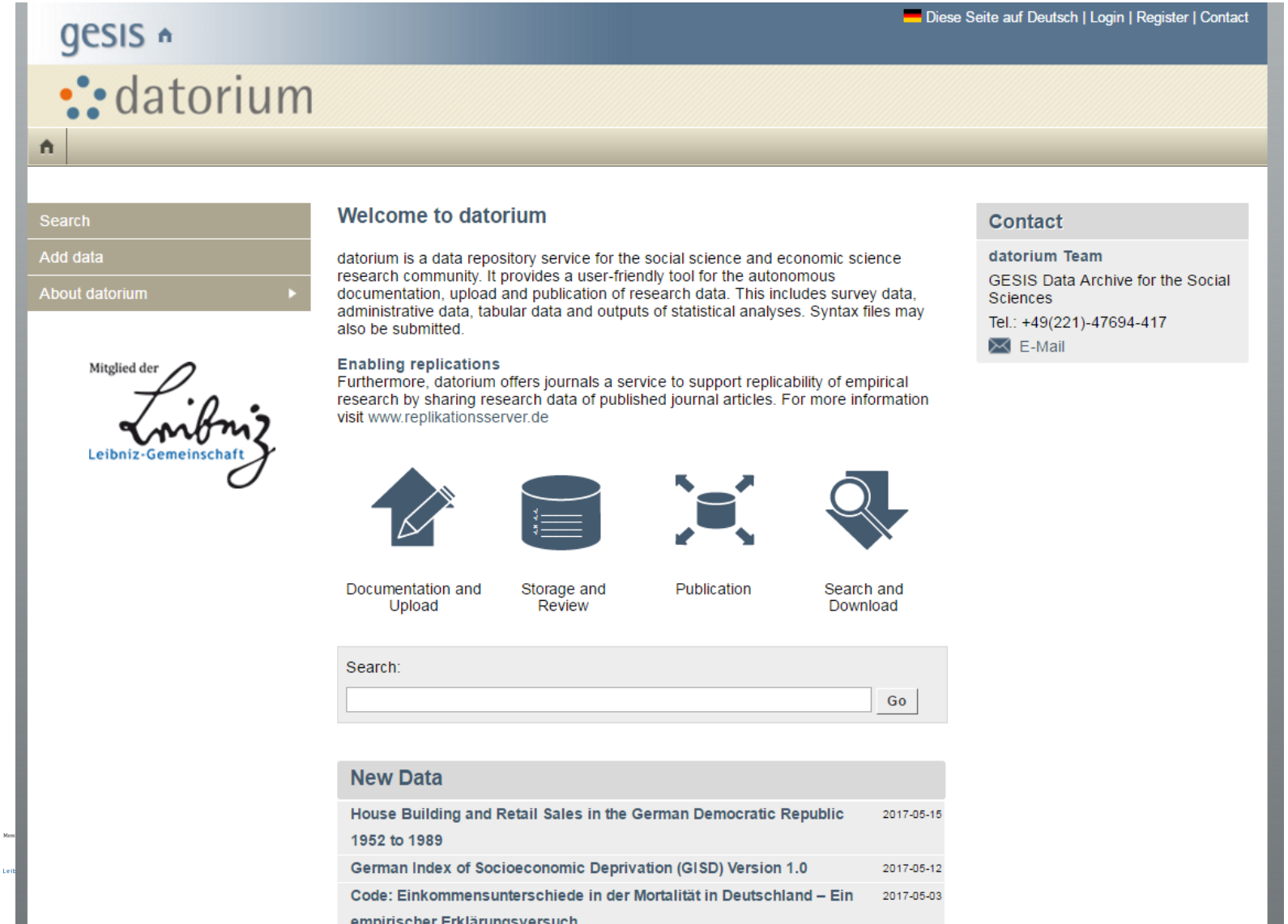
gesis Leibniz Institute for the Social Sciences

Current services at GESIS: DBK



gesis Leibniz Institute for the Social Sciences

Current services at GESIS: datorium



https://datorium.gesis.org/

gesis Leibniz Institute for the Social Sciences

Current services at GESIS: datorium

gesis Leibniz Institute for the Social Sciences

Current services at GESIS: datorium

gesis Leibniz Institute for the Social Sciences

Current services at GESIS: HISTAT

gesis

histat: Historical statistics 7.640.310 Values 367.234 Time series

Home | Data | About | Login | Searchtext

Historical Time Series

Welcome to the online-database histat. The database has been offered via GESIS since 2004. A substantial revision has been made in 2012. The number of database calls has been more than doubled within the last two years. This year (in 2012) we will attain more than 3.000 registered users. In the meantime, the database encloses a quarter of a million time series, 359 studies, and more than five million values.

Citation

Kindly cite the sources of the used data and the respective publication of the downloaded studies in your own publication according to the scientific conventions. Recommendations regarding the citation of historical data can be found under the topic "Bibliographic citation of research data and study related documents".

In cooperation with

DESTATIS
Statistisches Bundesamt

What others say about histat

"A really good and important project."

Prof. Dr. Martin Schulze Wessel
Chairman des Verbandes der Historiker und Historikerinnen Deutschlands

"An indispensable data-repository for German economic and social history."

Prof. Dr. Jörg Baten
General Secretary der International Economic History Association

"It is most certainly not possible to make it more comfortable."

Dr. Oliver Volckart
London School of Economics

"It's great that GESIS collects this data and provides it to researchers."

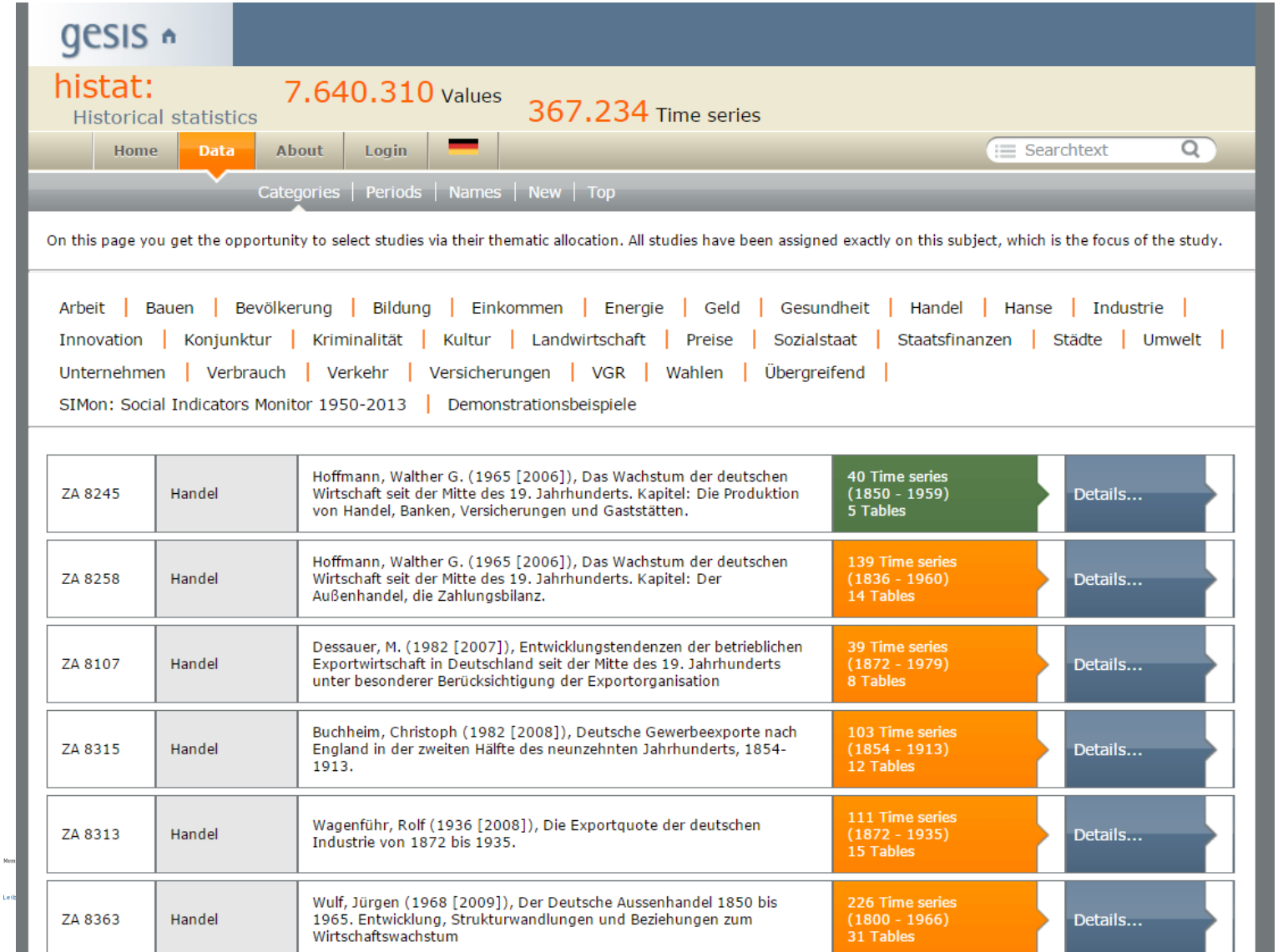
Robin Winkler
DPhil candidate in Economic History, Oxford University

Further cooperating Partners

https://histat.gesis.org

gesis Leibniz Institute for the Social Sciences

Current services at GESIS: HISTAT



gesis Leibniz Institute for the Social Sciences

Current services at GESIS: HISTAT

- HISTAT time series data in tables per study

gesis

histat: Historical statistics 7.640.310 Values 367.234 Time series

Home | Data | About | Login | Searchtext

Categories | Periods | Names | New | Top

Tableview

ZA 8245	Handel	Hoffmann, Walther G. (1965 [2006]), Das Wachstum der deutschen Wirtschaft seit der Mitte des 19. Jahrhunderts. Kapitel: Die Produktion von Handel, Banken, Versicherungen und Gaststätten.	40 Time series (1850 - 1959) 5 Tables	Details...

01. Die Produktion des Handels (1850-1959) (found timeseries: 17) Download

Produkte *	Nahrungsmittel	Brennstoffe	Holz	Baumaterial	Metallwaren	Papier	Chemikalien	Seife	Bier
Tables	Tab. 93	Tab. 93	Tab. 93	Tab. 93	Tab. 93	Tab. 93	Tab. 93	Tab. 93	Tab. 93
Sources	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...	Hoffmann, W. G. (196...
Notes	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....	1913=100 bzw. 10000....
Graph									
1850	887	8	43	14	20				44
1851	864	9	43	15	21				44
1852	894	10	47	15	21				41
1853	913	11	47	17	23				42
1854	940	12	49	10	25				38
1855	962	14	48	12	27				38
1856	981	15	57	11	31				40
1857	1026	16	60	13	35				47
1858	1041	18	59	14	39				50
1859	1025	17	53	13	37				53
1860	1082	18	57	16	36				51
1861	1129	20	63	18	34				52
1862	1150	22	70	19	36				57
1863	1240	24	74	25	45				63
1864	1287	28	72	26	50				66

Format conversions

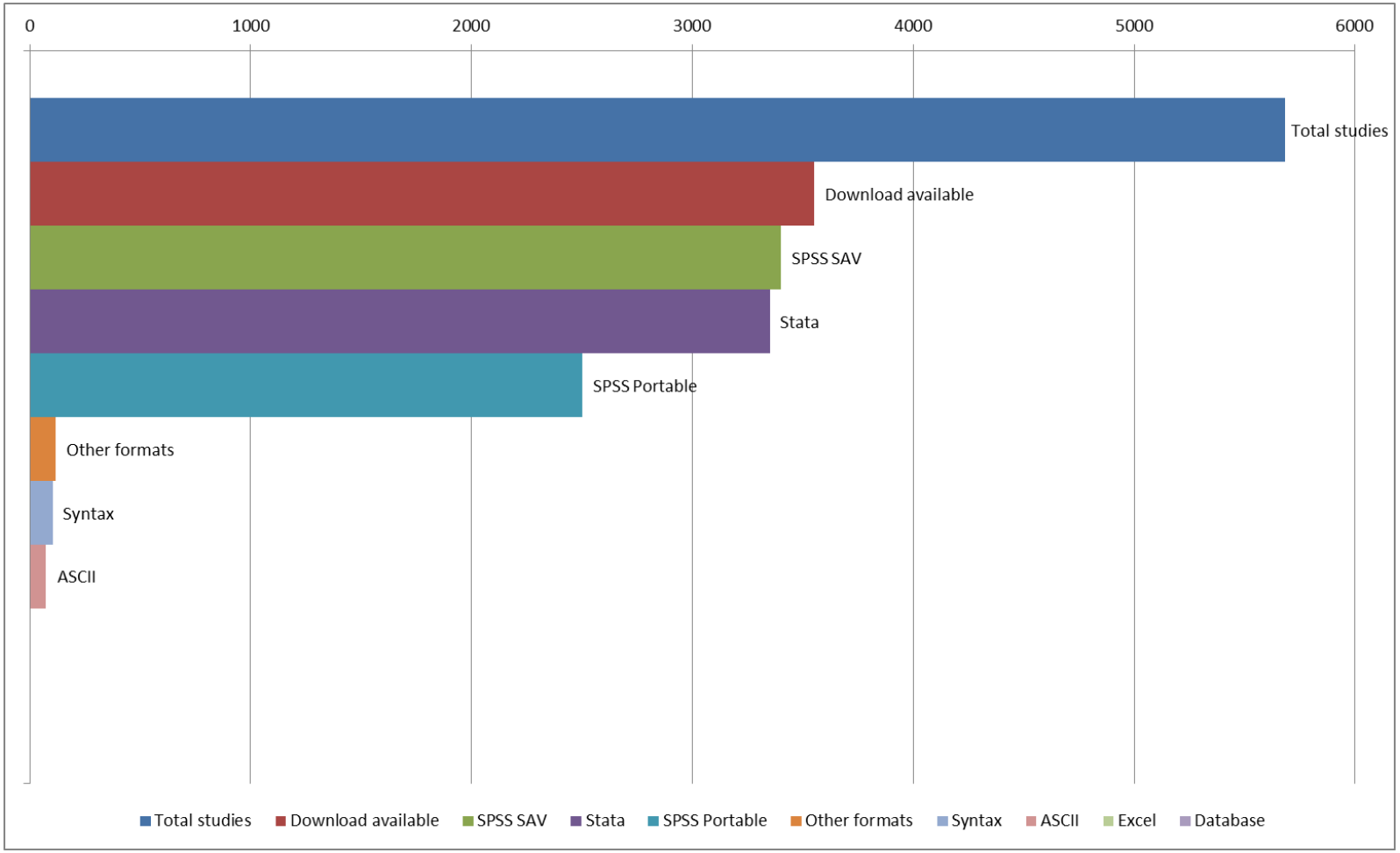
- Current file storage holds a variety of formats
- The Data Catalogue DBK
  - Predominantly SPSS (Portable, Windows)
  - STATA
  - Excel (CSV, XLS, XLSX)
  - ZIP-compressed files
  - Others
- datorium
  - What the researchers provide
- HISTAT
  - Excel (CSV, XLS, XLSX)

Format conversions

- DBK & HISTAT
  - internal long-term archiving repository
  - the originally submitted files
  - the processed files to be archived
  - the service files to be distributed
    - SPSS, STATA
    - Future: vendor independent, unicode based format
- Datorium
  - bistream preservation

gesis Leibniz Institute for the Social Sciences

Format conversions



Member of the Leibniz Association

Format conversions

- Tools to convert
  - PSPP
  - StatTransfer
- Issues with StatTransfer
  - ISO-8859-1 coded source files
  - Variable names in capital letters (sometimes)
  - Variable name replacement errors
  - Umlaut problems
  - Embedded newlines
  - Numbers shortened/rounded
  - minor updates already break compatibility

Format conversions

- Shell script with conversions and sanity checks
- Output: UTF-8 encoded, UNIX file format
  - <Study>[_<version>].csv (the data)
  - <Study>[_<version>].txt (metadata)
  - <Study>[_<version>].sps (SPSS syntax script, DOS)
  - <Study>[_<version>]-DDI-2_5.xml (metadata)
  - <Study>[_<version>].sql (SQL script)

- Special Thanks to *Andre Müller*

Format conversions

- Handle data

```
## Logging
echo "PSPP meta log:"  > $logdir/${basename}.log_pspp
head -n`expr $startline - 1` pspp_meta.proc2 | grep -v "^$" >> $logdir/${basename}.log_pspp
head -n`expr $startline - 1` pspp_meta.proc2 | grep -v "^$"
echo -e "\nPSPP data log:" >> $logdir/${basename}.log_pspp
cat pspp_data.log >> $logdir/${basename}.log_pspp
    # begin data file paranoia
rows=`grep "cases were transferred$" st_out-csv.stlog | cut -d\  -f1`
cols=`grep "^Input file has" st_out-csv.stlog | cut -d\  -f4`
h1=`awk -F"\t" '{print NF}' pspp_out.csv | uniq -c`
echo -n " Table is $cols by $rows "
# [ `echo $h1 | wc -w` = "2" ] || { echo "Error: $basename: Number of colums is not constant." >> $logdir/${basename}.log; echo "Number columns is not constant. Will exit";
if [ `echo $h1 | wc -w` != "2" ]; then
        mv pspp_out.csv tmp.pspp_out
        echo "" > pspp_out.csv
        while [ "`diff -q pspp_out.csv tmp.pspp_out`" ]; do

            echo `diff -q pspp_out.csv tmp.pspp_out` " "`wc -l pspp_out.csv`
            sed ':a;N;$!ba;s/
\n/ /g' tmp.pspp_out > pspp_out.csv
            sed ':a;N;$!ba;s/\(\t"\)\n/\1 /g' pspp_out.csv | sed ':a;N;$!ba;s/\(\t"[^\n\t]\+""\)\n/\1 /g' | sed ':a;N;$!ba;s/\(\t"[^"]\)\n/\1 /g' | sed ':a;N;$!ba;s/\(\t"[^
            echo -n " Removed embedded newlines from dataset. "
            echo "Info: $basename: removed embedded newlines from dataset." >> $logdir/${basename}.log;
        done
        h1=`awk -F"\t" '{print NF}' pspp_out.csv | uniq -c`
        if [ `echo $h1 | wc -w` != "2" ]; then
            echo " newlines remove leaves disarray "
            echo "Error: $basename: the newlines removal loop no helpee. Dataset in disarray." >> $logdir/${basename}.log;
            exit
        else    echo -n " successfully removed newlines from pspp cells "
            echo "Info: $basename: removed newlines from table cells." >> $logdir/${basename}.log;
        fi
        tabse=`expr $cols - 2`
        wc -l st_out.csv
        echo strini
        sed ':a;N;$!ba;s/
\n/ /g' st_out.csv > tmp.st_out
        echo "" > st_out.csv
        while [ "`diff -q st_out.csv tmp.st_out`" ]; do
            mv tmp.st_out st_out.csv
            wc -l st_out.csv
            echo stround
            sed ":a;N;\$!ba;s/\(\n[^\t\n]*\(\t[^\t\n]*\)\{1,$tabse\}\)\n/\1 /" st_out.csv | sed ":a;N;\$!ba;s/\(\n[^\t\n]*\(\t[^\t\n]*\)\{1,$tabse\}\)\n/\1 /" | sed ":a;N;\
        done
```

gesis Leibniz Institute for the Social Sciences

Format conversions

- Handle metadata

```
for i in `tail -n+2 ${fbase}-Vars_Lines | cut -d: -f1`; do
    length=`expr $i - $startline - 1`
    tail -n+$startline $indir/$file | head -n$length > cutfile
    v_index=`grep -m1 "^ V_Index:    " cutfile | cut -f2`
    v_format=`grep -m1 "^ V_FormatS:    " cutfile | cut -f2`
    v_measure=`grep -m1 "^ V_Measure:   " cutfile | cut -f2`
    v_sql_datatype=`grep -m1 " V_Datatype:  " cutfile | cut -f2 | sed -e 's/Integer/INT/' -e 's/String/VARCHAR/' -e 's/Decimal/DECIMAL/'`
    v_sql_varlen=`grep -m1 "V_VarLen:   " cutfile | cut -f2 | tr "." ","`
    v_label=`grep -m1 "^ V_Label:    " cutfile | cut -f2`
    v_label_ddi=`echo "$v_label" | sed -e 's/\&/\&amp;/g' -e 's/</\&lt;/g' -e 's/>/\&gt;/g' -e 's/\x27/\&apos;/g' -e 's/"/\&quot;/g'`
    v_label_spss=`echo "$v_label" | sed "s/'/''/g"`
    v_miss_spss=`grep -m1 "^ V_MissingR:    " cutfile | cut -f2`
    v_miss_d=`grep -m1 "^ V_MissingD:   " cutfile | cut -f2`
    v_sha=`grep -m1 "^ V_ValueL_sha:    " cutfile | cut -f2`
    if [[ $v_sha ]]; then
        sed "s/^\(${v_sha}\t.*\)$/\1${v_name} /" tmp.shalist > tmp.shalhelp
        mv tmp.shalhelp tmp.shalist
    fi
    echo -e "  $v_name\t$v_format" >> SPSS_var_form
    echo -e "  $v_name\t'$v_label_spss'" >> SPSS_var_label
    echo -e "label variable $v_name \`\"$v_label\"'" >> Stata_var_label
    echo -n " $v_name" >> tmp.M-$v_measure
    [[ $v_miss_spss ]] && echo -e "  $v_name\t($v_miss_spss)" >> SPSS_missing
    echo -e "\t\t<var ID=\"$v_name\" name=\"$v_name\">\n\t\t\t\t<labl>$v_label_ddi</labl>" >> DDI2_var
    echo -e "$v_index\t$v_name\t$v_label\t$v_format\t$v_measure\t$v_miss_spss\t$v_sha" >> tmp.sql-var_table
    echo -e "\t$v_name $v_sql_datatype($v_sql_varlen)," >> tmp.sql-data_def
    grep "^ " cutfile > tmp.vall_il
    cut -f2 tmp.vall_il > tmp.vall_i
    [ "`grep -m1 "^$v_sha    " tmp.sql-vallab`" ] ||  sed "s/^/$v_sha/" tmp.vall_il >> tmp.sql-vallab
    echo $v_miss_d | tr " " "\n" > tmp.miss_d
    for toto in `cat tmp.miss_d tmp.vall_i | sort -n | uniq`; do
        missus="N"
        if [ "`grep -x -m1 -- $toto tmp.miss_d`" ]; then
            missus="Y"
            echo -e "$v_name\t$toto" >> tmp.sql-missings
        fi
        lab_lab=`grep -m1 "^    $toto    " tmp.vall_il | cut -f3 | sed -e 's/\&/\&amp;/g' -e 's/</\&lt;/g' -e 's/>/\&gt;/g' -e 's/\x27/\&apos;/g' -e 's/"/\&quot;/g'`
        echo -e "\t\t\t<catgry missing=\"$missus\">\n\t\t\t\t<catValu>$toto</catValu>\n\t\t\t\t<labl>$lab_lab</labl>\n\t\t\t</catgry>" >> DDI2_var
    done
    [ -s tmp.vall_il ] && mv tmp.vall_il tmp.vl-$v_sha
    echo -e "\t\t</var>" >> DDI2_var
    startline=$i
    v_name=`grep "^$i:" ${fbase}-Vars_Lines | cut -d: -f3`
done
```

Provide access

- DBK download
  - in addition to the other formats
  - as a replacement for the archival version and downloads
- Enable direct access: API
  - machine-actionable
  - enables data linking
  - provides support for different formats via interface
  - crawling of data and metadata possible
  - makes visible what usage is needed
  - access regulations must be implemented

Build an easy to use API

- API should be RESTful
  - To enable good machine interaction

Example Action	Classic Webservice	RESTful
Get study list	GetStudies()	dbk/study/ HTTP Method GET
Get study description	GetDescription(studyID)	dbk/study/{studyID} HTTP Method GET
Get variables of study	GetVariables(studyID)	dbk/study/{studyID}/variable HTTP Method GET
Get variable description	GetVariable(studyID, var)	dbk/study/{study}/variable/{var} HTTP Method GET
Get variable data	GetData(studyID, var)	dbk/study/{study}/variable/{var}/data HTTP Method GET

Build an easy to use API

- API should be documented by OpenAPI
  - To enable easy programming against it

The OpenAPI Specification (fka The Swagger Specification)

build passing

Looking for the upcoming 3.0 version? See here:

The pre-release OAS 3.0.0 Specification Branch

The goal of The OpenAPI Specification is to define a standard, language-agnostic interface to REST APIs which allows both humans and computers to discover and understand the capabilities of the service without access to source code, documentation, or through network traffic inspection. When properly defined via OpenAPI, a consumer can understand and interact with the remote service with a minimal amount of implementation logic. Similar to what interfaces have done for lower-level programming, OpenAPI removes the guesswork in calling the service.

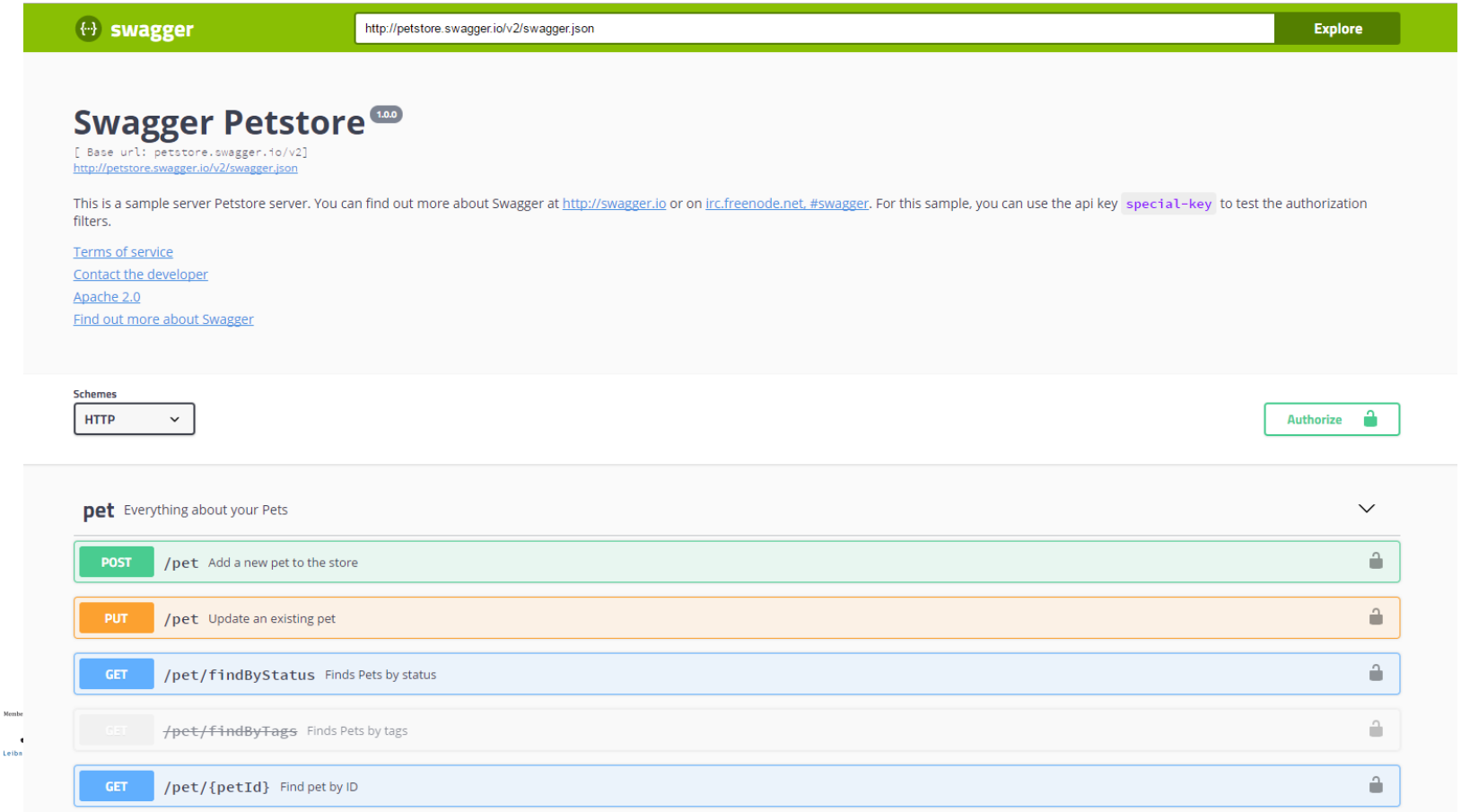
Use cases for machine-readable API interfaces include interactive documentation, code generation for documentation, client, and server, as well as automated test cases. OpenAPI-enabled APIs expose JSON files that correctly adhere to the OpenAPI Specification, documented in this repository. These files can either be produced and served statically, or be generated

https://github.com/OAI/OpenAPI-Specification

gesis Leibniz Institute for the Social Sciences

Build an easy to use API

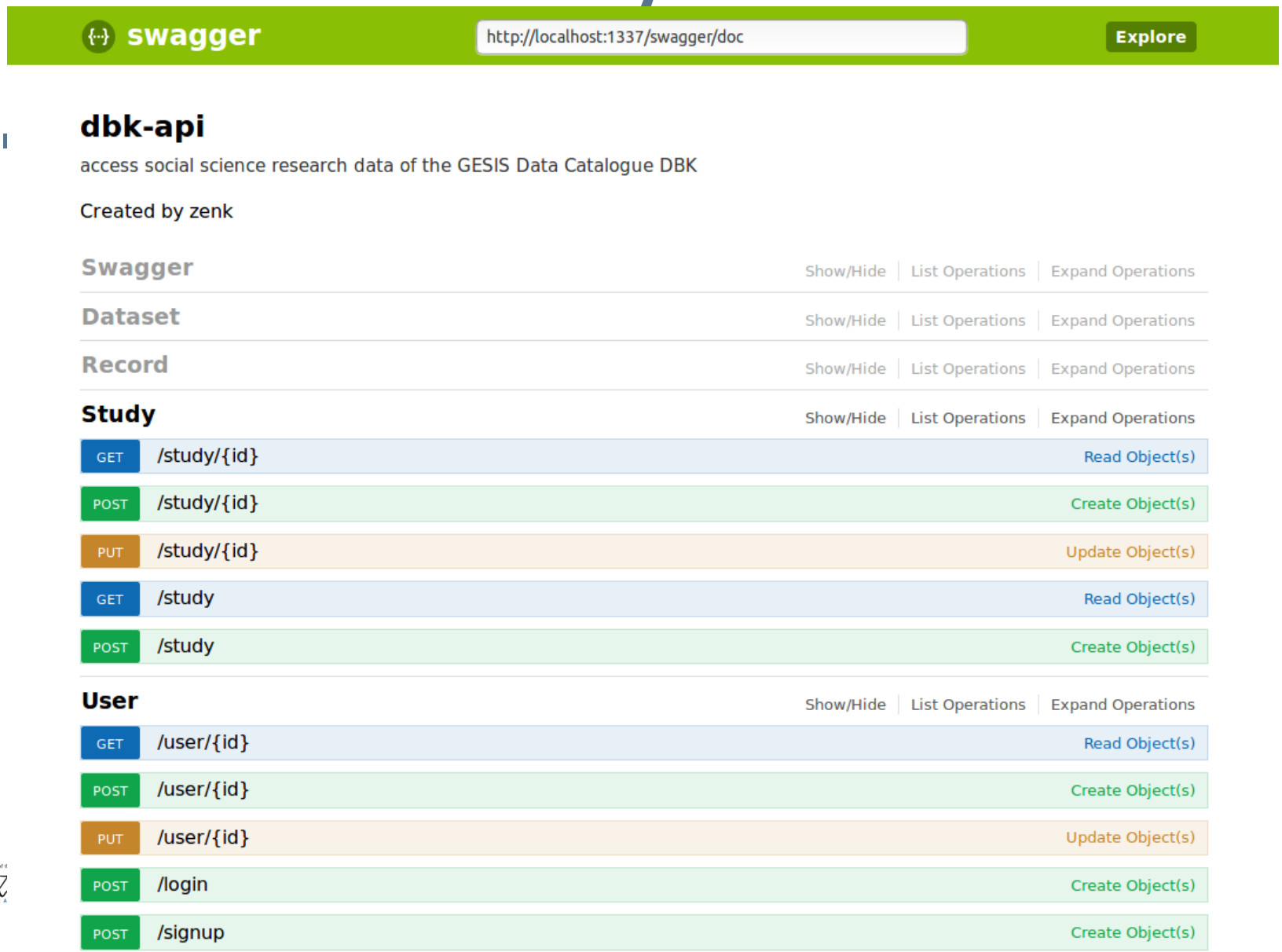
- API should be documented by OpenAPI
  - To enable easy programming against it



http://petstore.swagger.io/

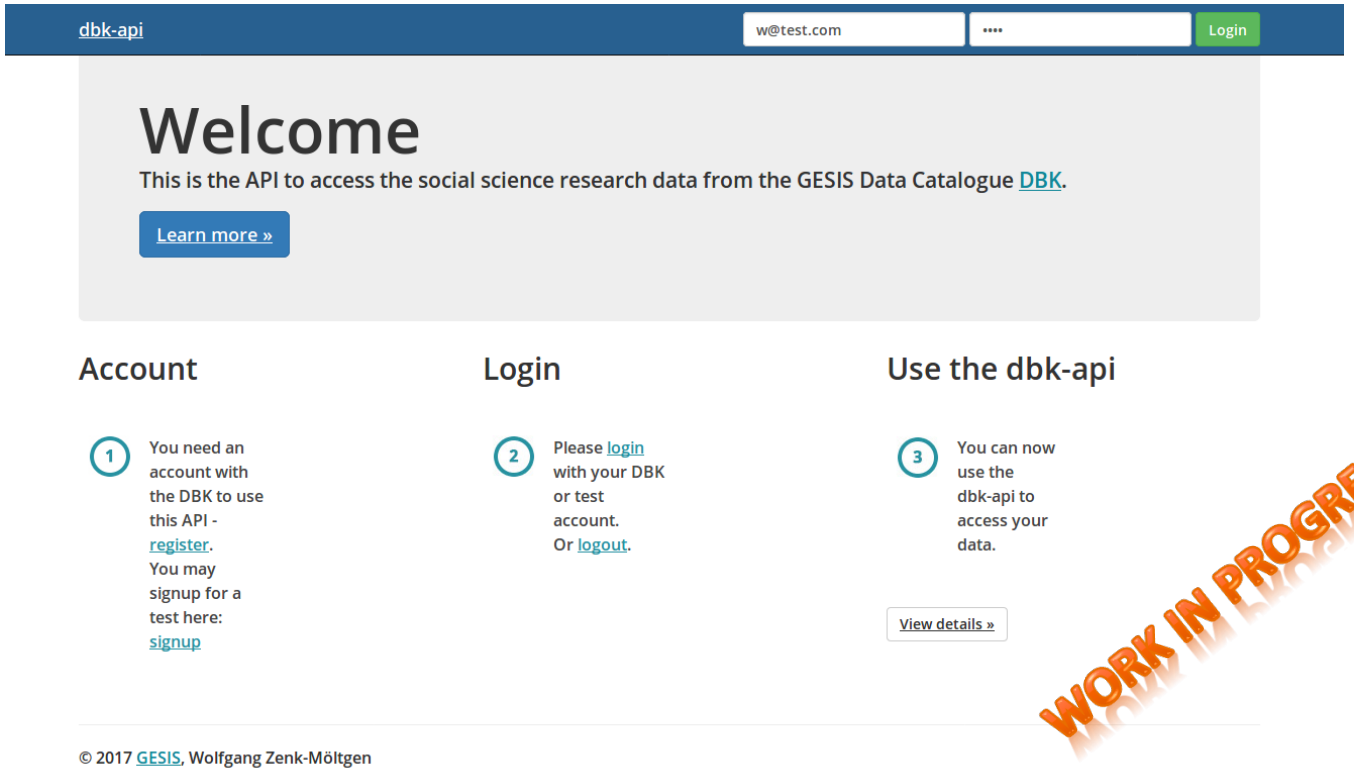
gesis Leibniz Institute for the Social Sciences

Build an easy to use API



Build an easy to use API

- Prototype
  - Using Sails.js, MVC web framework built on Node.js and EJS



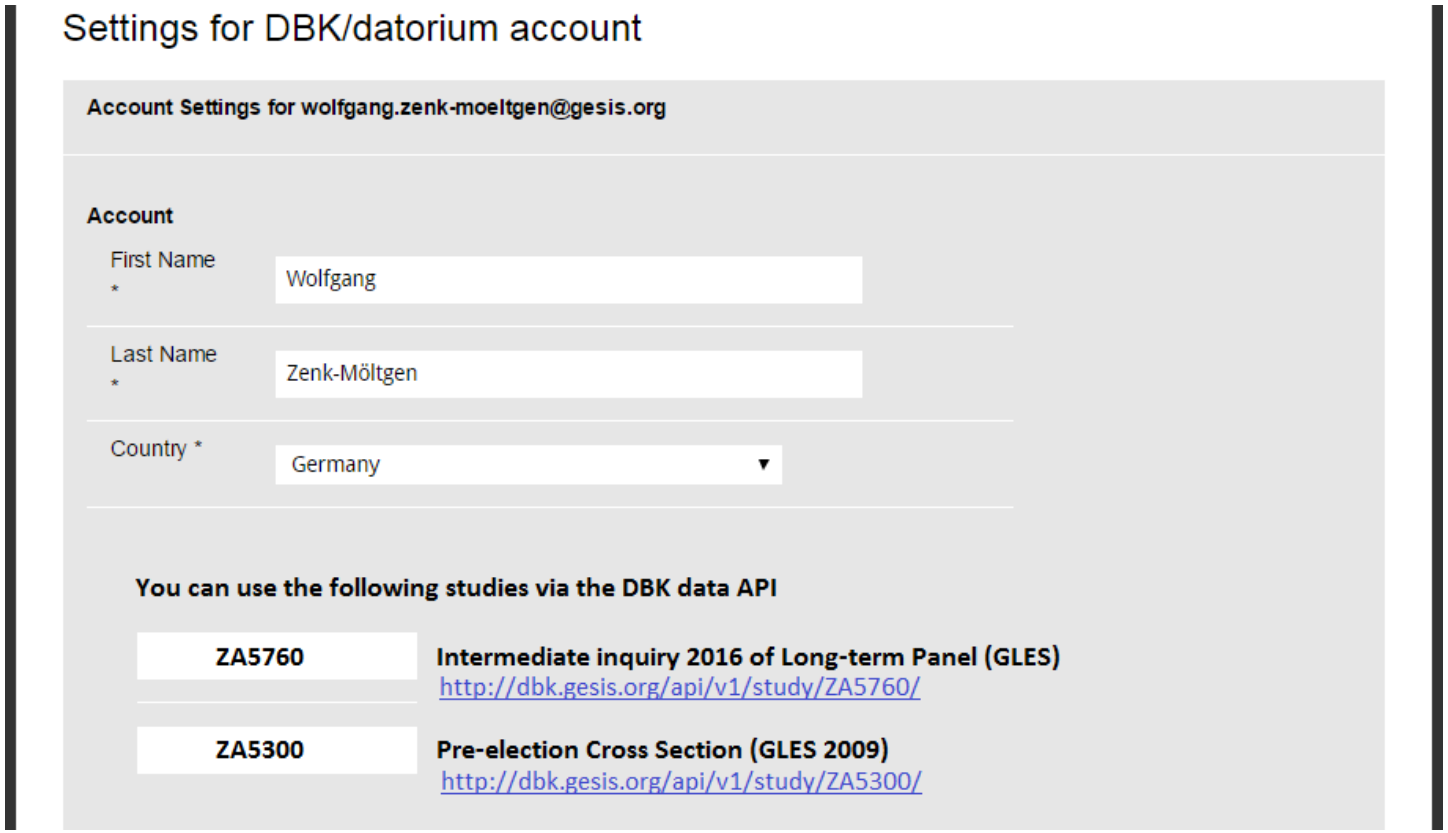
http://eis.co/

https://nodeis.org/

http://sailsis.com/

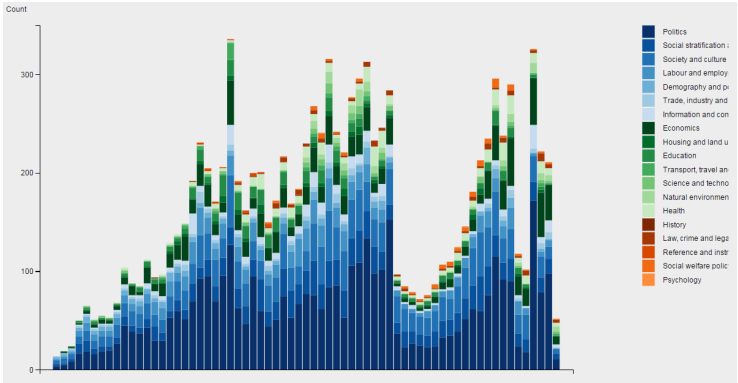
Future use

- Manage dataset access with a DBK user account



Future use

- Use Case
  - Access the API to fetch data for analysis from any programming language
  - E.g. use the R packages httr or RCurl
  - Combine the data with data from other sources
- There are many tutorials available, e.g.
  - https://www.r-bloggers.com/getting-data-from-an-online-source/
  - https://www.r-bloggers.com/accessing-apis-from-r-and-a-little-r-programming/
- In the future, you can use the DBK API to analyze data!



Thank you for your attention

gesis Leibniz Institute for the Social Sciences

Member of the Leibniz Association