The Collective Action of Data Collection: Innovations in Gathering Political Data

Holger Döring

University of Konstanz

Holger.Doering@uni-konstanz.de

Structure talk

- Data collection in political science
  - traditional approaches
- Recent innovations in data collection
  - political science data
  - more structure online sources
- ParlGov database
  - a new approach towards data collection
- Summary

Part I

Data collection in
political science

Studying government formation

- Research questions
  - – Who gets into government?
  - – Which coalitions form?
- Studying parliamentary democracies
- Evolution within comparative politics
  - – much work in the 1960s and 1970s
  - – revival of over the last decade

Evolution: than and now

- Theories
  - new game theoretical models
- Data collection
  - no progress
- Statistical techniques
  - new estimation techniques
  - more important: new software

Replication possible?

- Theories
  - proof is in the paper
- Data
  - often data available online
  - data as sausage: better don't look into it
  - researchers use different data
- Estimation techniques
  - norm to provide replication scripts

Data on political institutions

- Which data?
    - election results
    - government composition
    - party positions
- Which indicators?
    - – positional information
        - median, veto players etc.
    - – institutional parameters
        - effective number of parties, polarization etc.

What do we have?

- Good books
  - – data handbooks as part of political science
  - – eg. Mackie/Rose, Müller/Strom, EJPR
- Digital data
  - – OECD
  - – survey data with national data archives
  - – replication data provided online

Other (sub)disciplines

- Political economy
  - OECD and Worldbank data
  - provided by international institutions
- Political behavior
  - survey data
  - (national) institutionalized infrastructure
- International relations
  - COW (correlates of war)
  - Penn World Table
  - systematically (updated) sources

Best data we have

- A good data set
  - digitally provided
  - with codebook and good documentation
  - still a valid approach for 'static' data (eg. survey data)
- Shortcomings
  - no regular updating
  - combining primary and calculated observations
  - hard to combine with other sources (esp. party data)
  - no or in-transparent feedback loops
  - spreadsheet thinking

What is missing?

- Data on political institutions
  - Don't care about data!
  - institutional provision of data
  - same data basis
- Current reality
  - hours of coding
  - everyone has it 80-90% correct
  - little systematic knowledge from country experts

Part II

Recent innovations in data collection

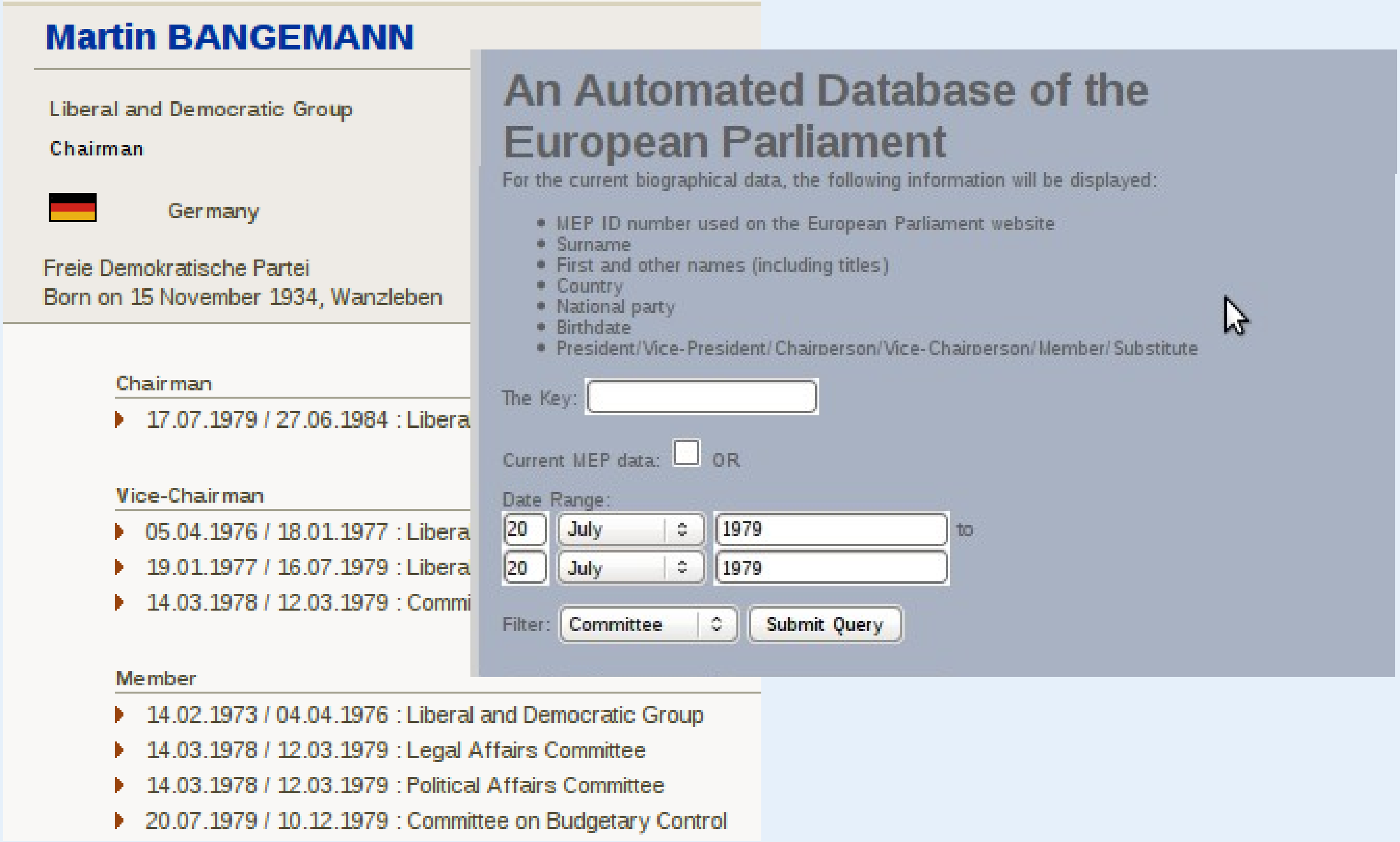
- political science data
- more general trends
- software tool

Data in political science

- Does good data matter?
  - researchers realize shortcomings
  - data sections in journals
  - funding for data collection easily available
- Increasingly more online data
  - mostly not explicitly for political science
  - digitalizing official documents
    - eg. information on MPs
    - official election results
    - Wikipedia

- Innovative data projects
  - The Dataverse Network Project
    - http://thedata.org/
    - million dollar NSF project
      - professional computer programmers
    - digital data archive for the 21th century
      - a library for data
      - central archives for replication data
  - Quality of Government (QoG) data
    - http://www.qog.pol.gu.se/
    - combining many, many existing data sets
    - three or six hundred variables

Hoyland ea. 2009



Martin BANGEMANN

Liberal and Democratic Group

Chairman

Germany

Freie Demokratische Partei
Born on 15 November 1934, Wanzleben

Chairman

- 17.07.1979 / 27.06.1984 : Libera

Vice-Chairman

- 05.04.1976 / 18.01.1977 : Libera
- 19.01.1977 / 16.07.1979 : Libera
- 14.03.1978 / 12.03.1979 : Commi

Member

- 14.02.1973 / 04.04.1976 : Liberal and Democratic Group
- 14.03.1978 / 12.03.1979 : Legal Affairs Committee
- 14.03.1978 / 12.03.1979 : Political Affairs Committee
- 20.07.1979 / 10.12.1979 : Committee on Budgetary Control

An Automated Database of the European Parliament

For the current biographical data, the following information will be displayed:

- MEP ID number used on the European Parliament website
- Surname
- First and other names (including titles)
- Country
- National party
- Birthdate
- President/Vice-President/Chairperson/Vice-Chairperson/Member/Substitute

The Key:

Current MEP data: □ OR

Date Range:

20 July 1979 to
20 July 1979

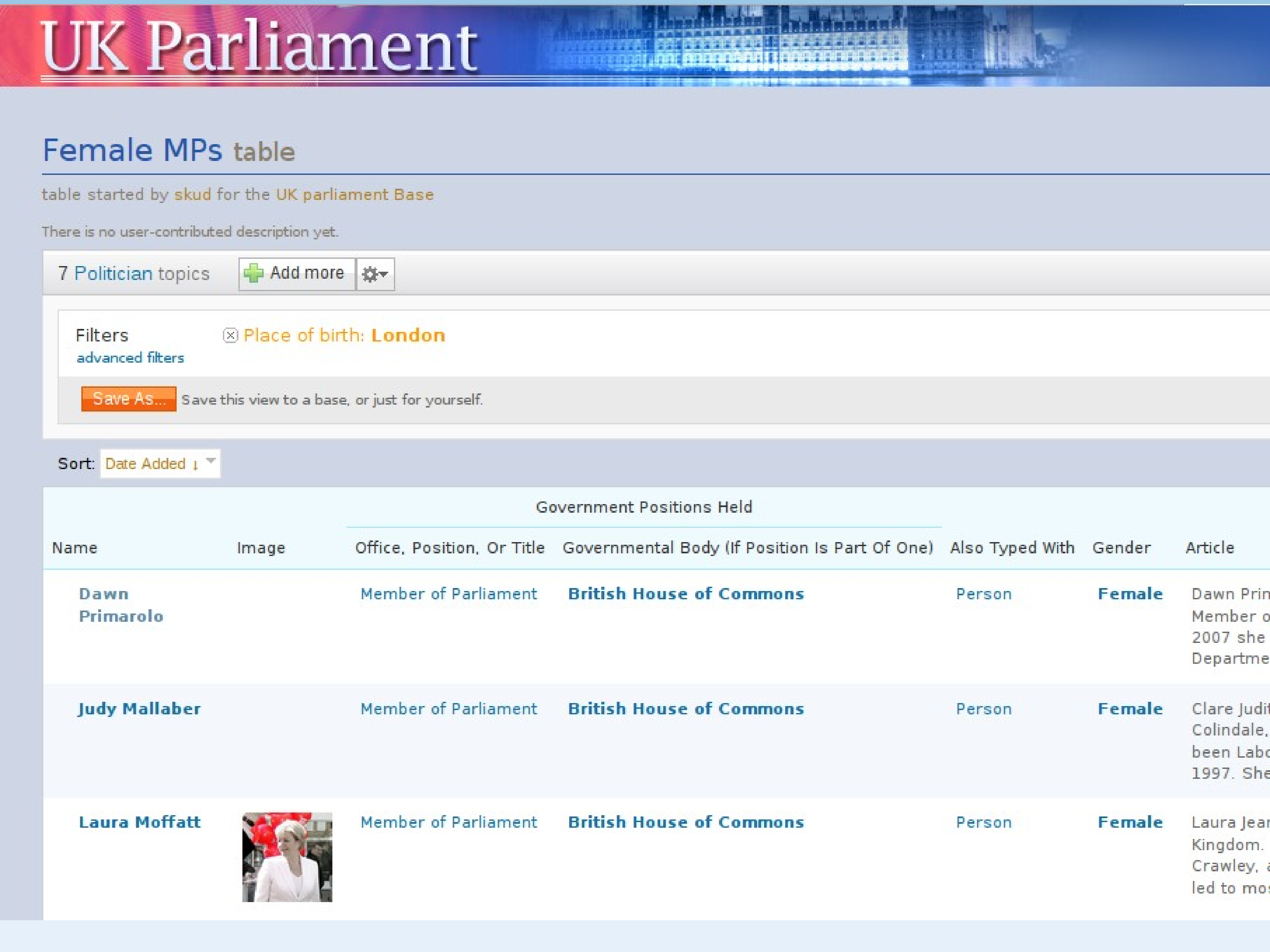
Filter: Committee

Hoyland ea. 2009

- An Automated Database of the European Parliament (Bjorn Hoyland, Indraneel Sircar, and Simon Hix), European Union Politics, 2009, Vol 10 (1), 143 – 152
- online data
  - http://www.europarl.europa.eu/members.do
  - http://folk.uio.no/bjornkho/MEP/
- from online to political science data
  - automatic conversion
    - computer script parses data anew every month
    - no human coding

Online data

- (semi)structured data
  - mixing content and format
    - web scraping required
  - Wikipedia
- structured data
  - semantic data
  - public data interface (API)
  - Freebase
    - eg. http://ukparliament.freebase.com/



UK Parliament

Female MPs table

table started by skud for the UK parliament Base

There is no user-contributed description yet.

Save this view to a base, or just for yourself.

Sort: Date Added ↓

		Government Positions Held				
Name	Image	Office, Position, Or Title	Governmental Body (If Position Is Part Of One)	Also Typed With	Gender	Article
Dawn Primarolo		Member of Parliament	British House of Commons	Person	Female	Dawn Prin Member o 2007 she Departme
Judy Mallaber		Member of Parliament	British House of Commons	Person	Female	Clare Judit Colindale, been Labo 1997. She
Laura Moffatt		Member of Parliament	British House of Commons	Person	Female	Laura Jear Kingdom. Crawley, a led to mos

Computer toolbox

- Skills needed to draw on this data
  - online interface -- browser
  - statistical software
    - data difficult to manage with SPSS or STATA
      - enforces spreadsheet thinking
    - R (http://www.r-project.org/) more flexible
  - databases
    - some knowledge of SQL helpful
  - programming languages
    - Python (or Ruby, PHP, Perl etc.)

Part III

ParlGov database

- a new data set concept
- computer tools

ParlGov introduction

- ParlGov
    - Parliament and government composition database
    - Holger Döring and Philip Manow
  - – Inspired by modern software development
  - – Current status
    - beta version online (password restricted)
    - double-checking and updating of data
    - needs feedback and user demands
  - – Future plans
    - open publicly under a free license
    - regular updated releases once or twice a year

What do we need?

- Better data on political institutions
  - visual interface (online website)
  - separate data and indicators
  - software routines to calculate indicators
  - provide linkage to other data sets
  - feedback mechanisms
  - include country experts
- Web 2.0 approach to data collection
  - – institutions/software to foster collaboration
  - – 'The collective action of data collection'

What do we need?

- Data
  - database
    - separate information
    - primary, secondary and foreign data
  - datasets automatically from database
  - regular release cycles
- Internet interface
  - dev.parlgov.org
    - more intuitive presentation of data
  - User interaction
    - wiki and example scripts
    - ticket system for updating and progress monitoring

ParlGov concept I

- Data types
  - Primary data
    - election results and government compositions
    - politician data (MPs, MEPs, ministers) – not public
  - Secondary data
    - database views
    - calculated indicators – scripts provided
  - Foreign data (linked)
    - party positions
    - electoral turnout (IDEA)
    - Eurobarometer (planned)

ParlGov concept II

- Online administration
  - editing of data user based
  - feedback mechanisms
    - ticket system to file bugs
    - wiki section (example scripts)
- Country experts
  - quality control
- Release cycle
  - regular releases (once or twice a year)
  - releases in data archive
- Open license

Summary

How to create better data

- Care about data!
  - help others to build upon your work
  - learn new techniques
    - some of you should learn programming
      - or get funding for programmers
    - online interfaces help significantly
      - for coding of data
      - for presenting content of data set
    - draw on new online sources

Hands-on

- Afternoon hands-on session
  - ParlGov presentation
    - structure internet presentation
    - database tables
    - working with ParlGov and R
      - replication data EP second-order election
  - more advanced techniques
    - web scraping
      - Italien election results
    - record linkage
      - fuzzy matching of data sets

Finally

Hopefully, there is no need for this talk five to ten years from now.

Thank you for your attention!