Sandbox projects

Global Conference on Big Data for Official Statistics
20-22 Oct 2015, Abu Dhabi, UAE

Fernando Reis, Eurostat
Boro Nikic, SURS
Albrecht Wirthmann, TF Big Data, European Commission (Eurostat)

Sandbox Projects

- Wikipedia
- Enterprise Websites
- Mobile Phones
- Web Scraping
- Prices
- Traffic Loops
- Smart Meters
- Social Media

Each experiment team produced a detailed report on its activity, available in draft format on the UNECE wiki

Wikipedia as a big data source

Insights on world heritage from analysis of Wikipedia use

Project team

- **Eurostat (EU)**
- **CSO (IE)**
- **ISTAT (IT)**

Time table

- **Final report: Mid November 2015**

Wikipedia as a big data source

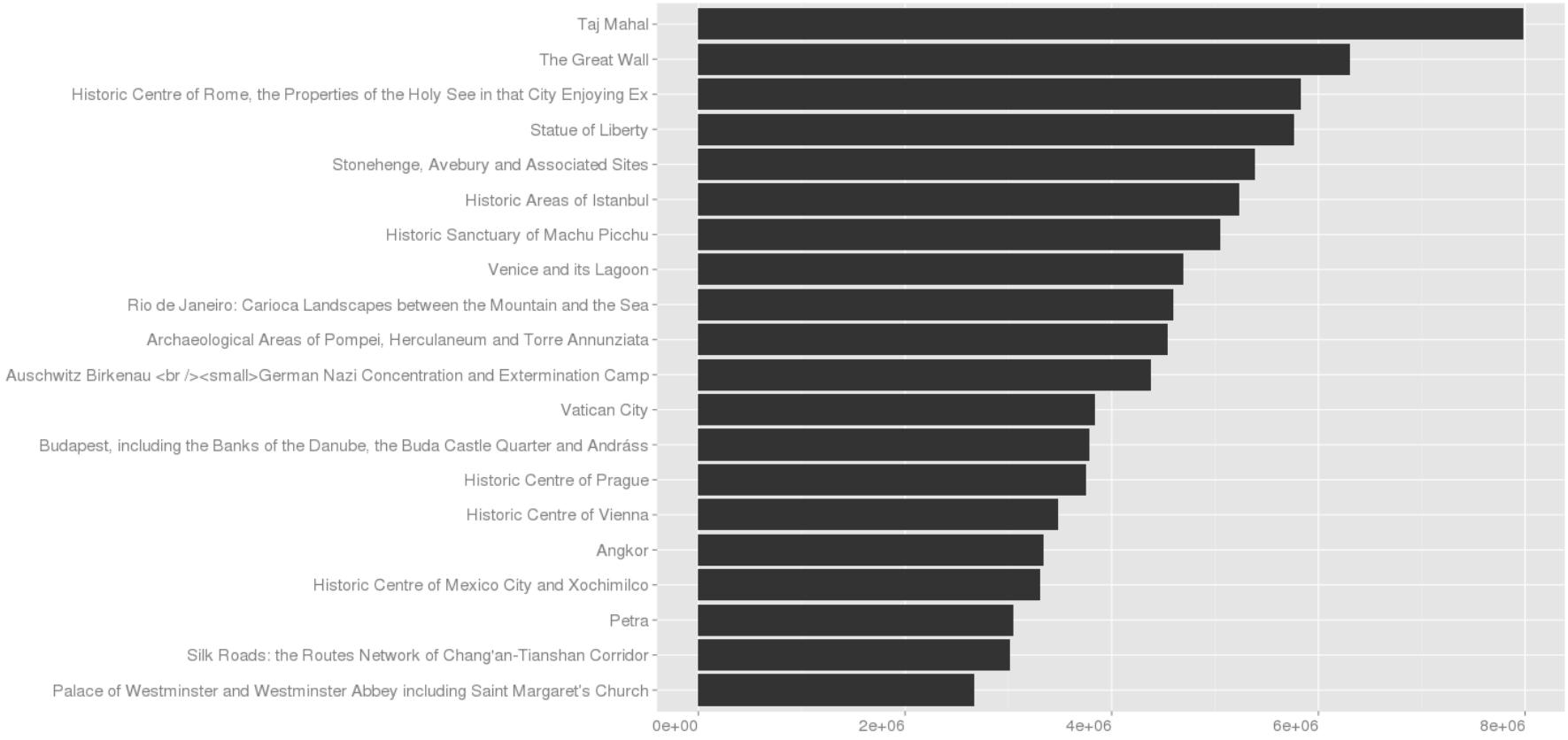
Insights on world heritage from analysis of Wikipedia use

- List of World Heritage Sites (WHS) from UNESCO
  - **Public source**
  - **Official information**
- Wikipedia
  - **Public source**
  - **Digital traces left by people in their activities**
  - **Widely used**
    - In 2013, 44% of individuals 16 to 74 years old living in EU consulted wikis to obtain knowledge (e.g. Wikipedia)
    - This was 69% for individuals between 16 and 24 years old
  - **Content (text and links)**
    - Selection of articles related to World Heritage sites
  - **Page views**
    - Wikistats: hourly number of page views for all articles of all wiki projects of the Wikimedia foundation;
  - **English Wikipedia only was used**

Wikipedia as a big data source

Insights on world heritage from analysis of Wikipedia use

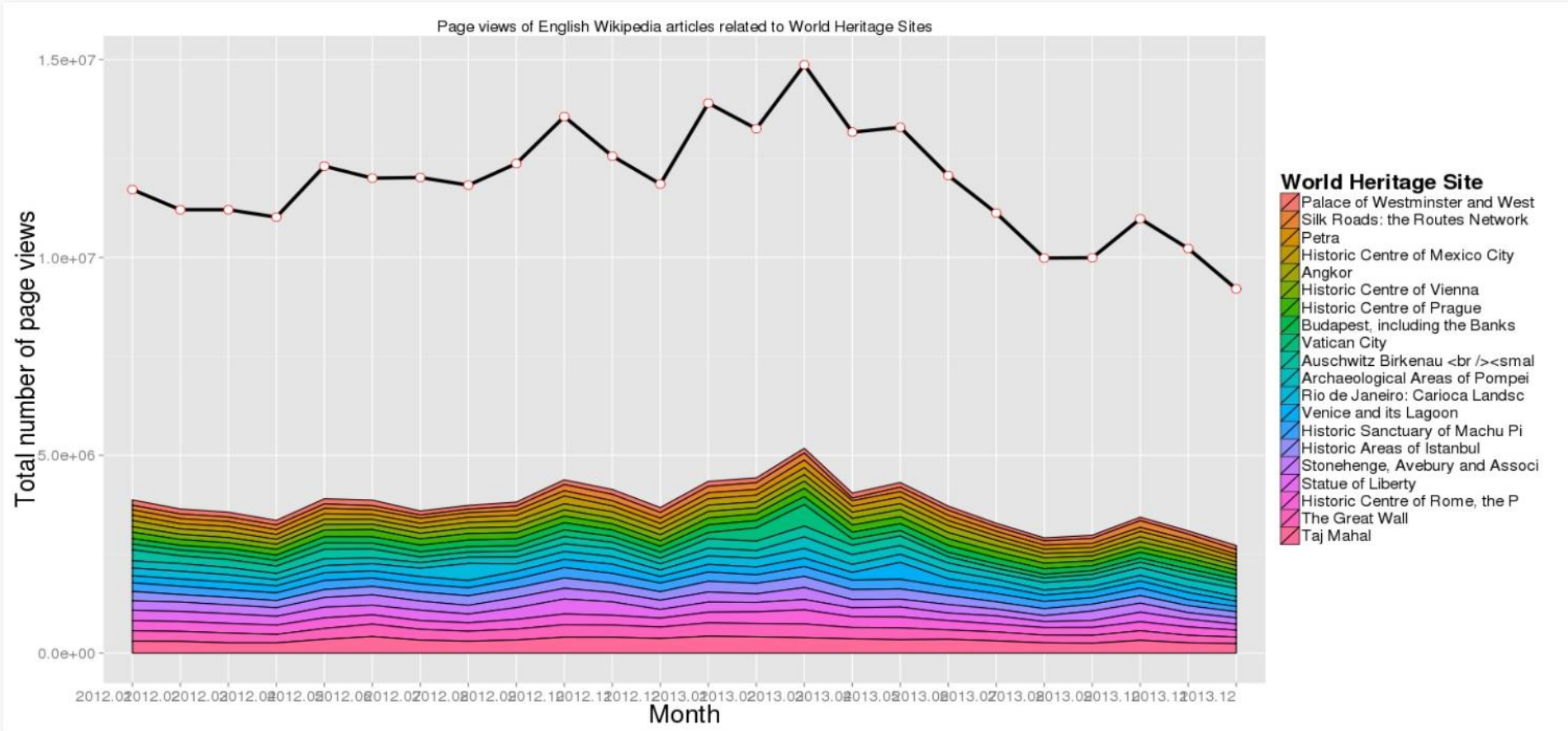
Total number of page views during 2012-2013 for the top WHS with most visits to its articles



European Commission

Wikipedia as a big data source

Insights on world heritage from analysis of Wikipedia use



Wikipedia as a big data source

Insights on world heritage from analysis of Wikipedia use

Further developments

- **Include more Wikipedia language versions**
- **Include page views of the mobile version of Wikipedia**
- **Explore other topics**

  Tourism lead indicators

Sandbox Project: Enterprise Web Sites

Aims:

Analyse and structure the information of websites in order to produce statistics

- **Link enterprises with websites using URLs**
- **Collect information in the websites**
- **Classify sub-pages according to topics, e.g. job-advertisements, analysing context and structure**
- **Structure information on the webpage, e.g. job descriptions, job titles, etc.**

Enterprise Web Sites

Project team

- **Switzerland**
- **Netherlands**
- **Poland**
- **Sweden**
- **Slovenia**

URLs of Enterprises

Two ideas:

- *List of URLs in NSIs (Business Register, Surveys,..)*
- *Names of enterprises from Business Register and web searching engines in order to collect the URLs;*

The Collection Process

Spider: *to find sub-pages relating to employment*

Downloader: *download sub-pages*

Splitter: *split the content of the downloaded sub-pages into different documents.*

Determinator: *detect the JV ads in the documents*

Classifier: *classify the identified JV by occupation, deadline, address, region…*

Improve identification of JV ads

- **Decision Tree**
- **Keywords (white- and blacklists)**

Outputs

- *The team works on methodology how to detect the Job Vacancy (JV) ads*
- *Possible JV statistics: number of JV ads broken down by NACE groups and by country*

Under the consideration:

- *Use of different software tools*
- *Detection of URLS from given the list the names of enterprises*
- *Further classification of JV (by occupation,..)*
- *Integration with survey data*