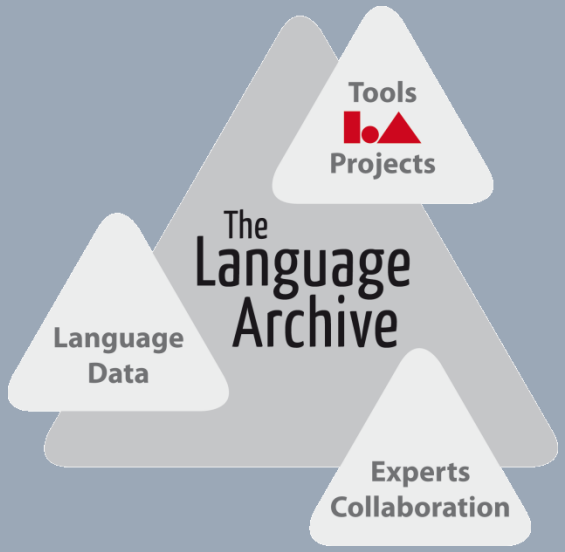


MAX-PLANCK-GESELLSCHAFT

Big Data from Crowdsourcing in the Humanities: Potentials and Challenges

Sebastian Drude
2013-09-24

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Big Data from Crowdsourcing

1. Big Data, Crowdsourcing, Apps
2. Potentials
3. Challenges
4. Conclusion

Big Data from Crowdsourcing

1. **Big Data, Crowdsourcing, Apps**
2. Potentials
3. Challenges
4. Conclusion

1. Big Data, Crowdsourcing, Apps

- Revolution: era of digital and interconnected research
  - New methods, questions
  - New collaborations
- Humanities: traditionally focussed on qualitative research and understanding than quantitative
- Now statistical and data-driven methods are increasing in the Humanities → "Digital Humanities"
- Big data: socio-economic data, the WWW, social media
- Much local, specific knowledge → Several repositories
- Overarching infrastructures are being built (CLARIN, DARIAH, DASISH, EUDAT, Research Data Alliance, ...)

1. Big Data, Crowdsourcing, Apps

- Crowd-sourcing: large numbers of individuals contribute remotely (via the internet) to a research project
- So far by personal computers running specialized computer programs or by interactive websites
- A new revolution: ubiquitous mobile devices
- Exponentially powerful digital technology for everybody
  - high-quality audio and video recordings
  - sensors of location, direction, movements, etc.
  - Soon: collecting human-physiological facts
- Democratizing effect — inclusion of new user groups
- Now: Mobile-device-apps for new-level crowd-sourcing

Big Data from Crowdsourcing

MAX-PLANCK-GESELLSCHAFT

1. Big Data, Crowdsourcing, Apps
2. **Potentials**
3. Challenges
4. Conclusion

2. Potentials

Data creation, elicitation and collection

- It is traditionally often difficult to get balanced samples in experiments or questionnaires
- Apps can reach many more participants
- The researcher doesn't have to be personally present
- Language research: dialect & varieties samples, data from lesser studied languages
- Multimodal research with audio & video recordings
- Complementary data points can be gathered automatically (geographical position, movement patterns, interconnections in social media)

2. Potentials

Data Storing, Pre-processing and Management

- Crowd-sourcing requires a proper infrastructure and analytical capabilities for the mass of incoming data
- New technology will emerge in response
- Connect many mobile devices to a repository that can:
  - provide data packages (stimuli & curation sets,...)
  - manage and store incoming data systematically, with automatically generated metadata
  - pre-process the data (quality checks, conversions)
- →new management, curation and access workflows

2. Potentials

Data “processing” (compilation, enrichment, annotation)

- Currently much time is needed for manual annotation
- Also for typing, OCR correction, mark-up, etc.
- Efficient pre-processing and semi-automatic methods will emerge to deal with the huge amount of data
- Already now crowd-sourcing is successfully used in processing existing and new data
- This is also enhanced by apps, many more participate
- Small tasks of detection of events and structures

2. Potentials

Dissemination

- Current paradigm: paper & text based publications
- Only few results get into larger or even mass media
- Data sets themselves are a scientific result
- In future dissemination by interactive, multi-media applications with advanced visualizations
- Much-needed bridge for the communication of scientific results to scholars and the broader public
- Engaging and allowing the citizens themselves to apply advanced methods and manipulate data sets

Big Data from Crowdsourcing

1. Big Data, Crowdsourcing, Apps
2. Potentials
3. **Challenges**
4. Conclusion

3. Challenges

How to engage the users

- A tool or app has to reach its users
- The challenges are twofold:
  - How to make the potential users or target group aware of the app
  - How to make them actually use it and make the desired kind of contributions
- Needed: Appealing user interfaces, rewards, payment, game-ification, functionality that is useful or fun, etc.

3. Challenges

How to receive, manage, pre-process, and store the data sustainably

- How to set up an appropriate back-end infrastructure
- When, how often, and how to transfer the data to the central back-end infrastructure (offline use and sync)
- How to automatically add metadata to the incoming data points and sets, and store both in a systematic way
- How to automatically pre-process the data, configure general policies of pre-selection or -annotation
- How to sustainably store large amounts of data safely
- How to (re-)distribute stimuli or other datasets (back)

3. Challenges

Provenance information and quality assessment

- Critical part of data or meta-data:
  - Who was the user (name, socio-cultural properties)
  - When and where were the data generated
  - On the basis of which stimulus or in which context
  - Etc. (some information may be sensitive)
- Quality assessment:
  - Spam or less useful data have to be identified and flagged or removed, as far as possible automatically
  - Identifying users with a bad track record
  - Evaluate quality also for potentially useful data

3. Challenges

How to curate the data

- Incoming data has to be further processed or curated
  - Text may have to be orthographically corrected
  - Oral data may have to be transcribed / translated
  - The accurateness and appropriateness of data may have to be checked and possibly improved
- Some activities can be replaced by automatic methods
- “Crowd-curation” may be useful or necessary
- A strategy for data curation has to be included in the plans for an app infrastructure and workflow

3. Challenges

Privacy, intellectual property, authorship, access restrictions

- Protect the privacy of contributors or curators
- Give the appropriate credits for their contribution
- Intricate ethical and even legal questions
- It may be necessary to control and restrict access to certain data sets, types or properties
- Informed consent explicitly ... or by using the app
- Any issues around the authorship of both data collections and scientific results have to be considered and settled beforehand

3. Challenges

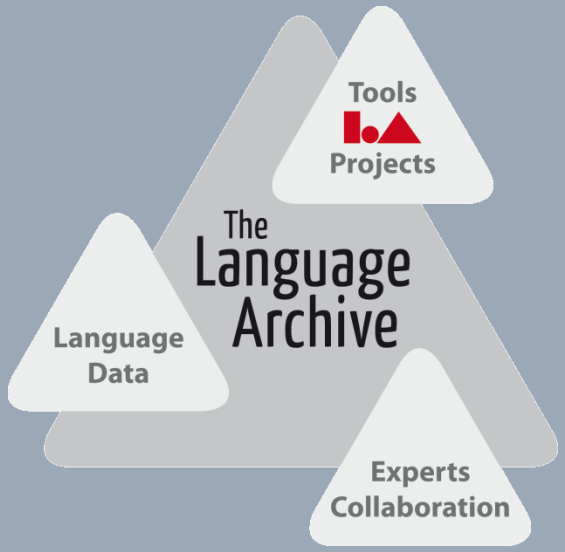
Life-cycle: policy-based handling and de-commitment

- How to deal with sets of data (by collection, selected subsets, data type, etc.) in a systematic manner?
- Some sets will have to be preserved for a certain time to serve as empirical evidence (accountability)
- Others may be popular among users
- Others will have to be transferred to or mirrored at other sites or in other applications, etc.
- This requires policy-based automated treatment
- End of life-cycle: de-commitment of the data

1. Big Data, Crowdsourcing, Apps
2. Potentials
3. Challenges
4. **Conclusion**

4. Conclusion

- The age of mobile devices will revolutionize science
- New levels of crowd-sourcing in the Humanities
- Potentials on various levels & stages in the process
- Challenges of technical, logistical and societal kind
- Successful employment of crowd-sourcing depends on much more than an appealing app for mobile devices
- The back-bone infrastructure needs careful planning and installation
- This is a task for data centres experienced with complex and large sets of digital data



MAX-PLANCK-GESELLSCHAFT

Big Data from Crowdsourcing in the Humanities: Potentials and Challenges

Sebastian Drude
2013-09-24

The Language Archive – Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands