Observations from a newcomer*

(*with some input from Steven Krauwer)

Franciska de Jong

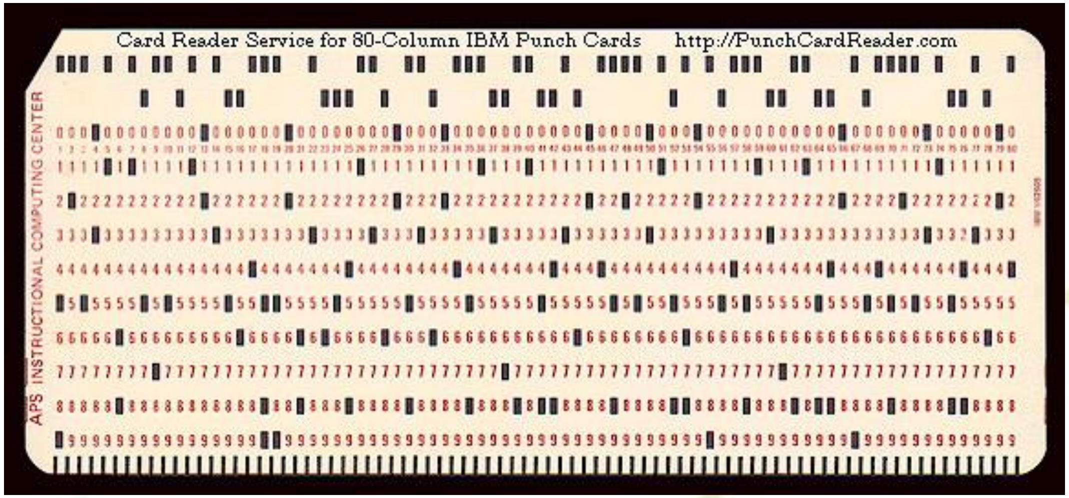
f.m.g.dejong@uu.nl

CLARIN-PLUS workshop

Vienna

2-3 June 2016

Where I come from


Card Reader Service for 80-Column IBM Punch Cards
http://PunchCardReader.com
APS INSTRUCTIONAL COMPUTING CENTER

Characteristics of RIs in SSH

- Typically federations of distributed, pre-existing repositories of digital data, tools and expertise, in multiple countries
- ESFRI examples:
  - Social and statistical data (CESSDA, SHARE, ESS)
  - Humanities, heritage domain (CLARIN, DARIAH)
- No significant investments needed for construction of central nodes; relatively cheap to build and operate
- Main motivation is not to share the financial burden but to get more out of what is available
- Research agendas aim to use the potential for typical EU dimensions and underline perspectives for: comparative research, crossing the language barriers, inclusive societies
- Political agendas: open science, single European market, collaboration across countries

Characteristics of RIs in SSH (ctd)

- Gradual transition from construction to operation; no planned termination point; countries can join in at later stages
- Normally no major capacity constraints, i.e. no access restrictions required for most services
- Most of the action (and the spending) is in the countries; only a modest central coordination point

Sustainability issues for RIs

- Size of investments (effort) call for adequate attention for sustainability
- Various aspects
  - adoption and acceptance by researchers
  - results that count
  - inspiration for the renewal of agendas
  - knowledge sharing mechanisms and practices
  - technical maintenance and updates
  - perspectives for financial support

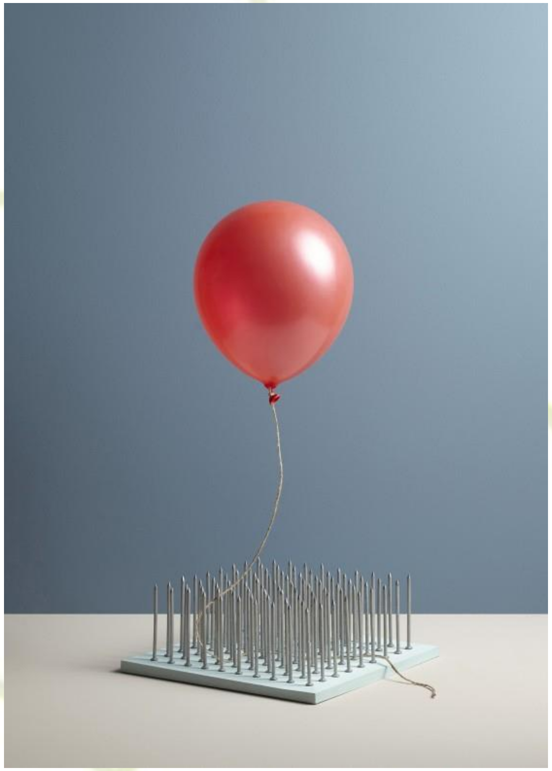


image by Kyle Bean

Sustainability issues for CLARIN

- Size of investments (effort) call for adequate attention for sustainability
- Various aspects
  - **adoption and acceptance by researchers**
  - results that count
  - **inspiration for the renewal of agendas**
  - **knowledge sharing mechanisms and practices**
  - technical maintenance and updates
  - perspectives for financial support

Adoption/acceptance issues

- dominance of data-driven research
- stimulus for number crunching excercises *versus* understanding of the phenomena underlying the data
- lack of attention for theory
- disruptive effects on existing traditions and practises
- technology push due to lack of understanding humanities workflows
- lack of attention for required skills levels
- lack of convincing showcases

How to turn our tools and services into must-haves for scholars?

image by Kyle Bean

Needs of users of oral history data: determined by task

Data curation

- Transcription
- Time alignment
- Metadata creation

Exploration of the data space

- Finding interviews on specific topic, with a specific interviewee
- Finding fragments with a specific name, phrase, etc.

Analysis

- Annotation
- Text / transcript mining
- Link generation

Presentation

- Citation of fragments
- Visualization

Sustainability issues for CLARIN

- Size of investments (effort) call for adequate attention for sustainability
- Various aspects
  - adoption and acceptance by researchers
  - results that count
  - **inspiration for the renewal of agendas**
  - knowledge sharing mechanisms and practices
  - technical maintenance and updates
  - perspectives for financial support

CLARIN and data science agendas

- Analytics for text and speech data as a pillar for data science
- Contribution to the development of new methodological frameworks for the integrated processing of multiple datatypes and multidisciplinary research agendas.
- Europe’s mulitlinguality as a basis for comparative research of societal phenomena, and in particular those that are reflected in language use:
  - Migration patterns
  - Intellectual history
  - Language variation
  - ….
- Text and speech as **social** and **cultural** data

Sustainability issues for CLARIN

- Size of investments (effort) call for adequate attention for sustainability
- Various aspects
  - adoption and acceptance by researchers
  - results that count
  - inspiration for the renewal of agendas
  - **knowledge sharing mechanisms and practices**
  - technical maintenance and updates
  - perspectives for financial support

The + in CLARIN-PLUS

- Look for balance between
  - Tailor-made services to finetune the pipelines to the requirements of the long tail of scholars

  &

  - Generic services
  - Software sustainability
  - Benchmark corpora
- Attract the scholarly brains that fuel the agendas with new questions
- Exchange best practices, on all relevant dimensions