Case Study – Social Sciences/Humanities

PARSE.insight workshop

Darmstadt, 21. September 2009

Sabine Schrimpf

Background

Why Social Sciences and Humanities?

– We assumed that their data usage and e-infrastructure needs differ from those of the hard science communities

Why Psycholinguistics and Book Studies?

– Together, they represent the spectrum of Humanities and Social Sciences

Psycholinguistics
- – Concerned with the relationship between the human mind and language as it examines the processes that occur in the brain while producing and perceiving both written and spoken discourse

"Book Studies"
- – Developed in the 19th century as an ancillary science from the study of literature
- – Media-Science, which takes part on the exploration of a cultural history of knowledge
- – Traditional objects of research are ancient books, the invention of printing by Johannes Gutenberg as well as actual tendencies on the book market

Background

Both communities

- work with digital objects
- are well connected
- already have a rudimentary e-infrastructure for scientific data

Difference between the data used

- Psycholinguistics: data is created during experiments and/or observations
- Book studies: work with digitized objects provided by a third party
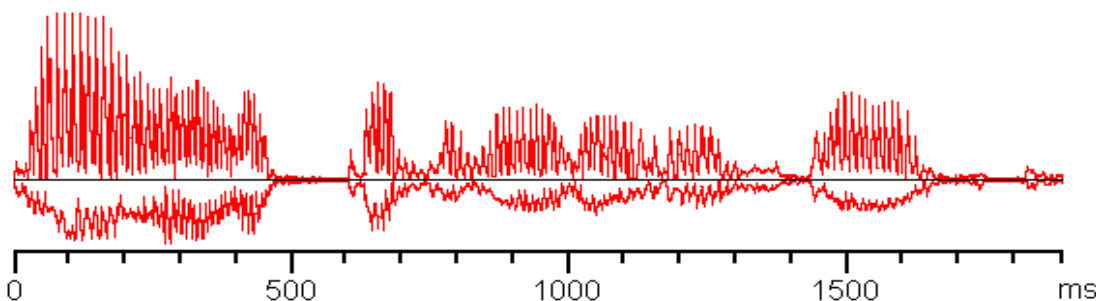
Research workflow in the Psycholinguistics
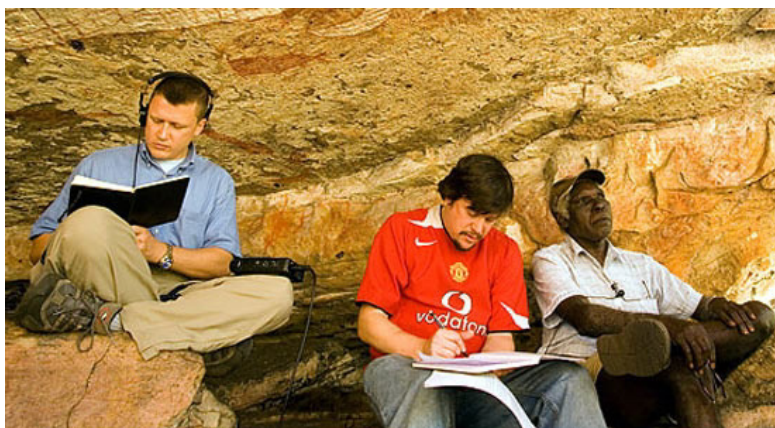
Record spoken language
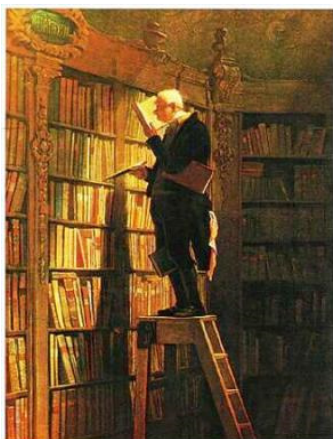

Language production research

Analyse speech patterns

Research workflow in the Book Studies

Consult sources in an archive/library

Consult sources in databases
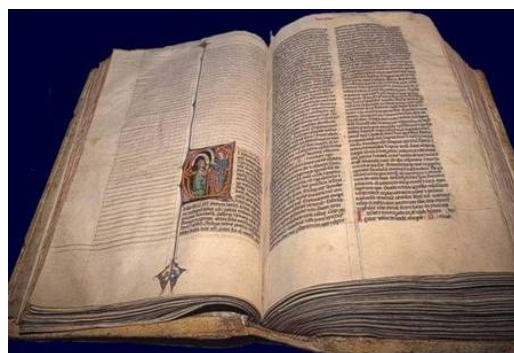
Hermeneutical study of sources
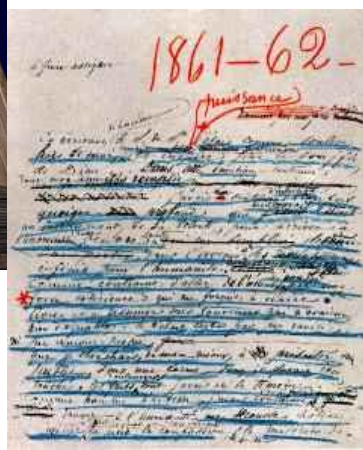
Publications
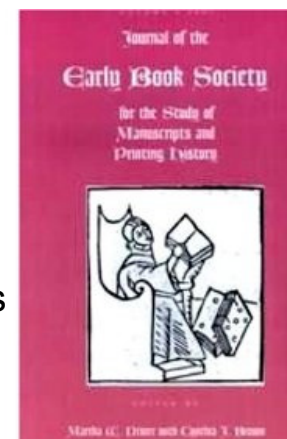
The Survey

Both communities were surveyed in winter/spring 2009

Psycholinguistics (PL)

– 3100 researchers

– 364 respondents

– 205 (56.3%) completed the survey

Book Studies (BS)

– 975 researchers

– 211 respondents

– 124 (58.8%) completed the survey

Background of Respondents

- The majority of respondents came from Europe (68% in both PL and BS)

- The largest groups were individual researchers, followed by research group leaders/managers, employees/participants of research groups, and PhD students

- PL: also data managers

Background of Respondents

- Large consensus that preservation of digital resources is important

- High awareness of threats to long term preservation

- BS: Concept of „digital preservation" rather unclear to some respondents

How important do you consider the following threats over the next 10 years?

Threat	(very) important	(less) important	don't know
The ones we trust to look after the digital holdings may let us down.	99	42	10
The current custodian of the data may cease to exist at some point in the future.	134	16	
Loss of ability to identify the location of data.	112	31	4
Access and use restrictions may not be respected in the future.	100	44	3
Evidence may be lost because the origin and authenticity of the data may be uncertain.	121	28	2
Lack of sustainable hardware, software or support of computer environment may make the information inaccessible.	131	20	1
Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved.	103	45	3

■ (very) important ■ (less) important □ don't know

How important do you consider the following threats over the next 10 years?

Threat	(very) important	(less) important	don't know
The ones we trust to look after the digital holdings may let us down	110	70	16
The current custodian of the data may cease to exist at some point in the future	152	30	14
Loss of ability to identify the location of data	146	40	9
Access and use restrictions may not be respected in the future	106	79	10
Evidence may be lost because the origin and authenticity of the data may be uncertain	153	39	4
Lack of sustainable hardware, software or support of computer environment may make the information inaccessible	155	37	4
Users may be unable to understand or use the data e.g. the semantics, format or algorithms involved	158	35	4

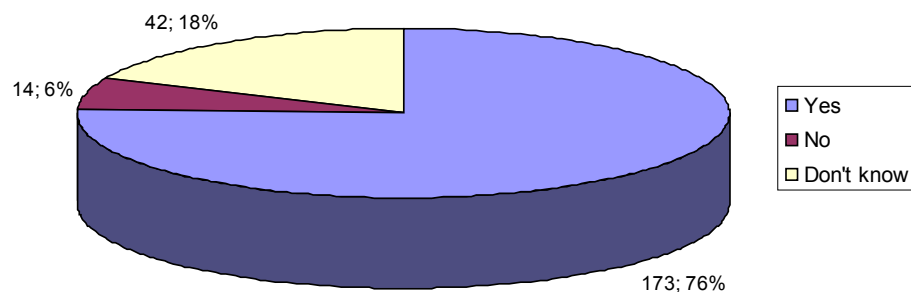Legend: ☐ (very) important ■ (less) important ☐ don't know

• Most respondents think that an international e-infrastructure would help guard against some of these threats
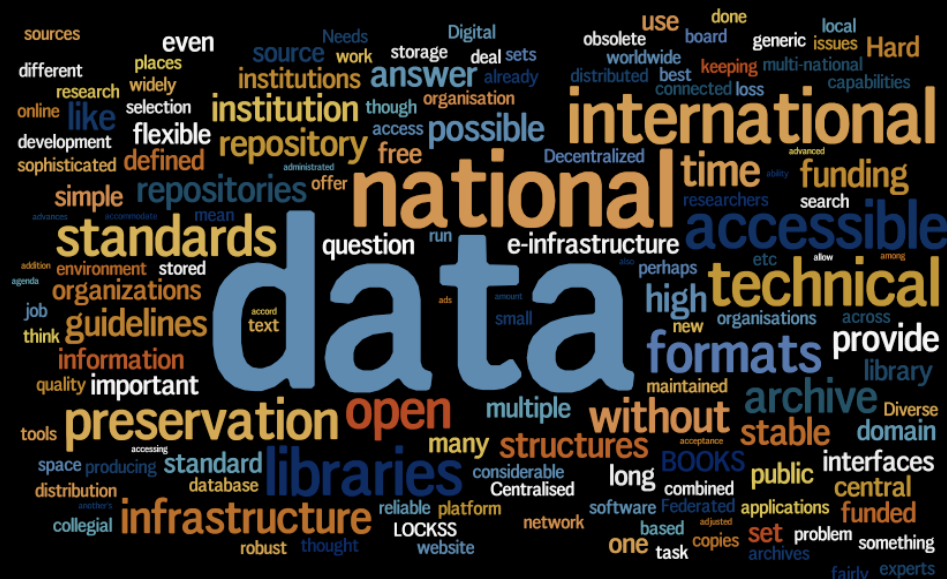
Book Studies
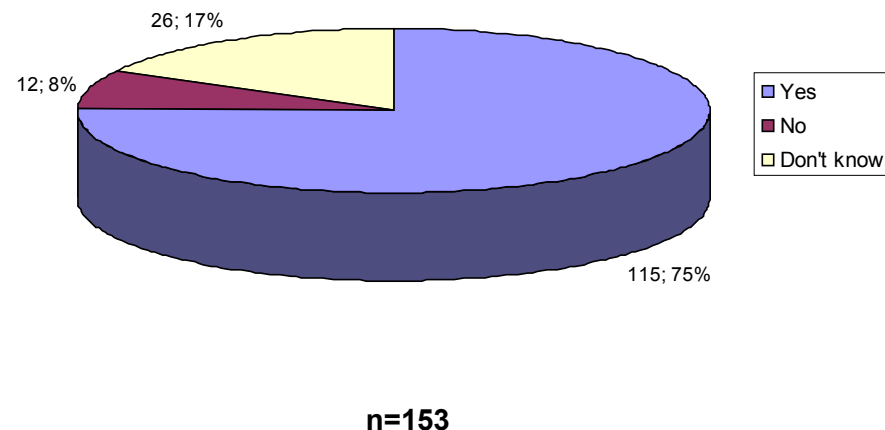
Psycholinguistics

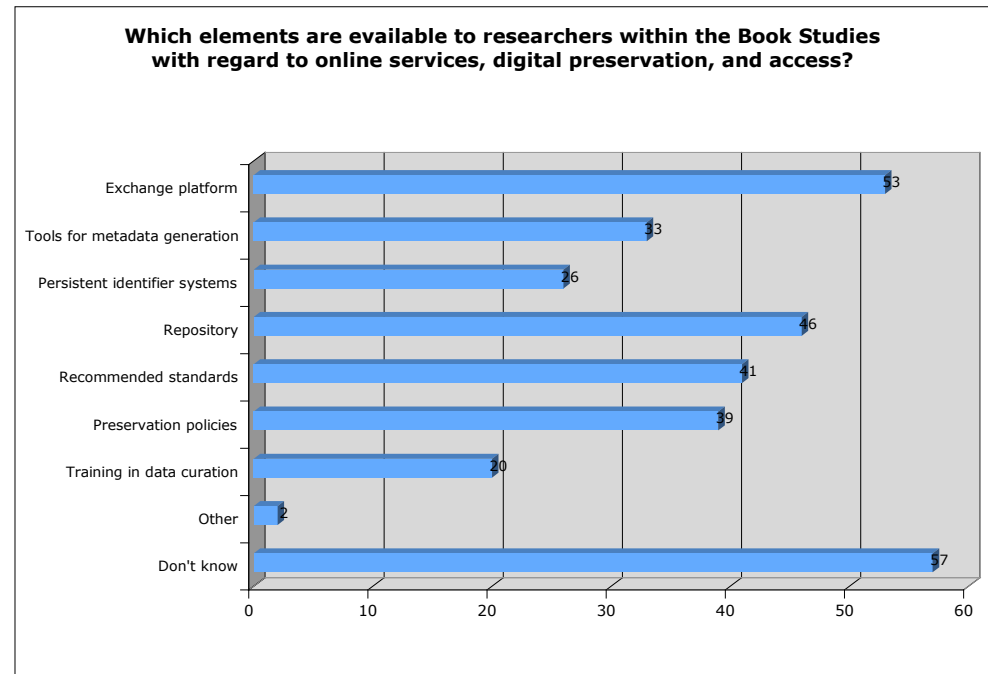- Ideas range from centralized EU agency to completely decentralized e-infrastructure

Psycolinguistics

Book Studies

What is already there in terms of general e-Infrastructure?

- PL: language archives, specialised databases (e.g. audio files, lexical corpora)

- BS: library catalogues, specialised databases (e.g. of manuscripts, early printed books)

- When asked for data sharing and digital preservation specific infrastructure elements, 43% did not know what was available

- Another 59% did not know if the existing tools and elements fit their needs

Which elements are available to researchers within the Book Studies with regard to online services, digital preservation, and access?

Element	Value
Exchange platform	53
Tools for metadata generation	33
Persistent identifier systems	26
Repository	46
Recommended standards	41
Preservation policies	39
Training in data curation	20
Other	2
Don't know	57

n=317 (multiple answers possible)

E-Infrastructure PL

- The Psycholinguists have a better understanding of their discipline specific digital services. They do, however, judge digital preservation related tools (i.e. for metadata generation and persistent identifier systems) little important

- Only 25% are satisfied with the existing components, 39% are not

Which elements are available to researchers within the Psycholinguistics with regard to online services, digital preservation, and access?

Element	Value
Tools for automated data generation	47
Research Tools for language resources	119
Tools for automated data exchange and conversion	69
Tools for Metadata generation	40
Persistent identifier systems	34
Storage	114
Recommended standards	65
Preservation Policies	50
Training in data curation	24

n=562 (multiple answers possible)

Data Storage BS

- Most researchers store their data on their PC at work (77%), or at home (70%), additionally on portable storage carriers (75%)

- Only few researchers submit them to a digital archive (10%) or to an external web service (11%)

- 74% were willing to use a digital archive (of their organisation), if there was one and 69% would make use of a discipline specific repository

Data Storage PL

- PC at work and at home most important data stores (95% each)

- More measures are taken to ensure that data is not lost:
  - 89% submit their data to an external web service
  - 87% store their data additionally on a server of their organization
  - 80% make use of an internal data archive, and 78% of an external data archive

Conclusion BS

- Greater acceptance and open-mindedness towards digital publications than we assumed

- General awareness of importance of digital preservation, but no connection to own work (yet?)

- In the interviews, importance of cross-discipline infrastructures was emphazised (need to make use of materials from other disciplines)

- Uncertainty of responsibility for and funding of e-infrastructure

Conclusion PL

- Use of digital services as a matter of course in the research process

- Still, 54% think existing components and services need to be improved

- Especially, training in data curation is missed

- A couple of e-infrastructure projects address the felt deficits: DOBES, CLARIN, DARIAH, CESSDA, …

Thank you very much for your attention!