Linguistics and EUDAT

Pavel Straňák

Charles University in Prague

stranak@ufal.mff.cuni.cz

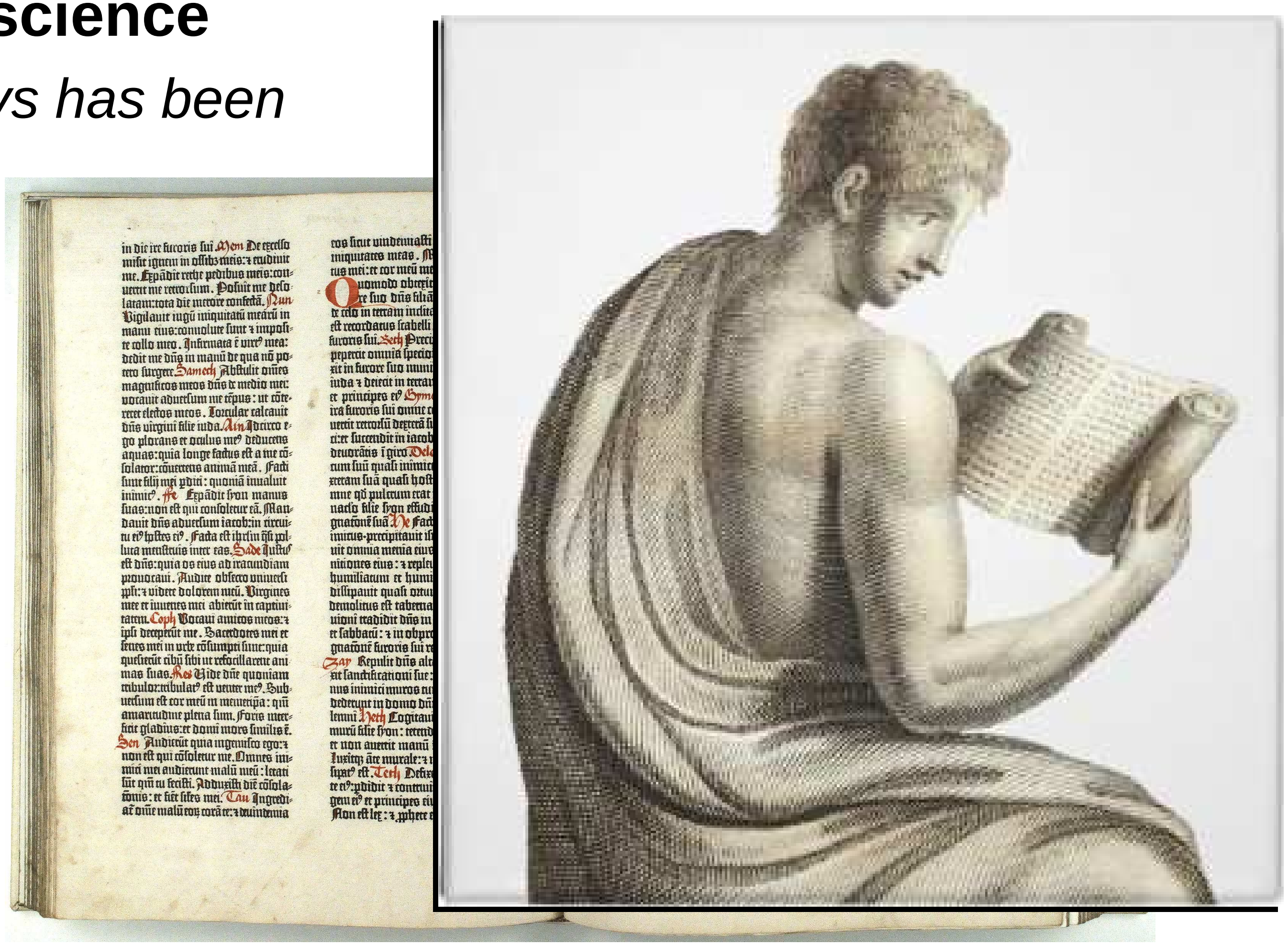
1st EUDAT User Forum
Barcelona
7. March 2012

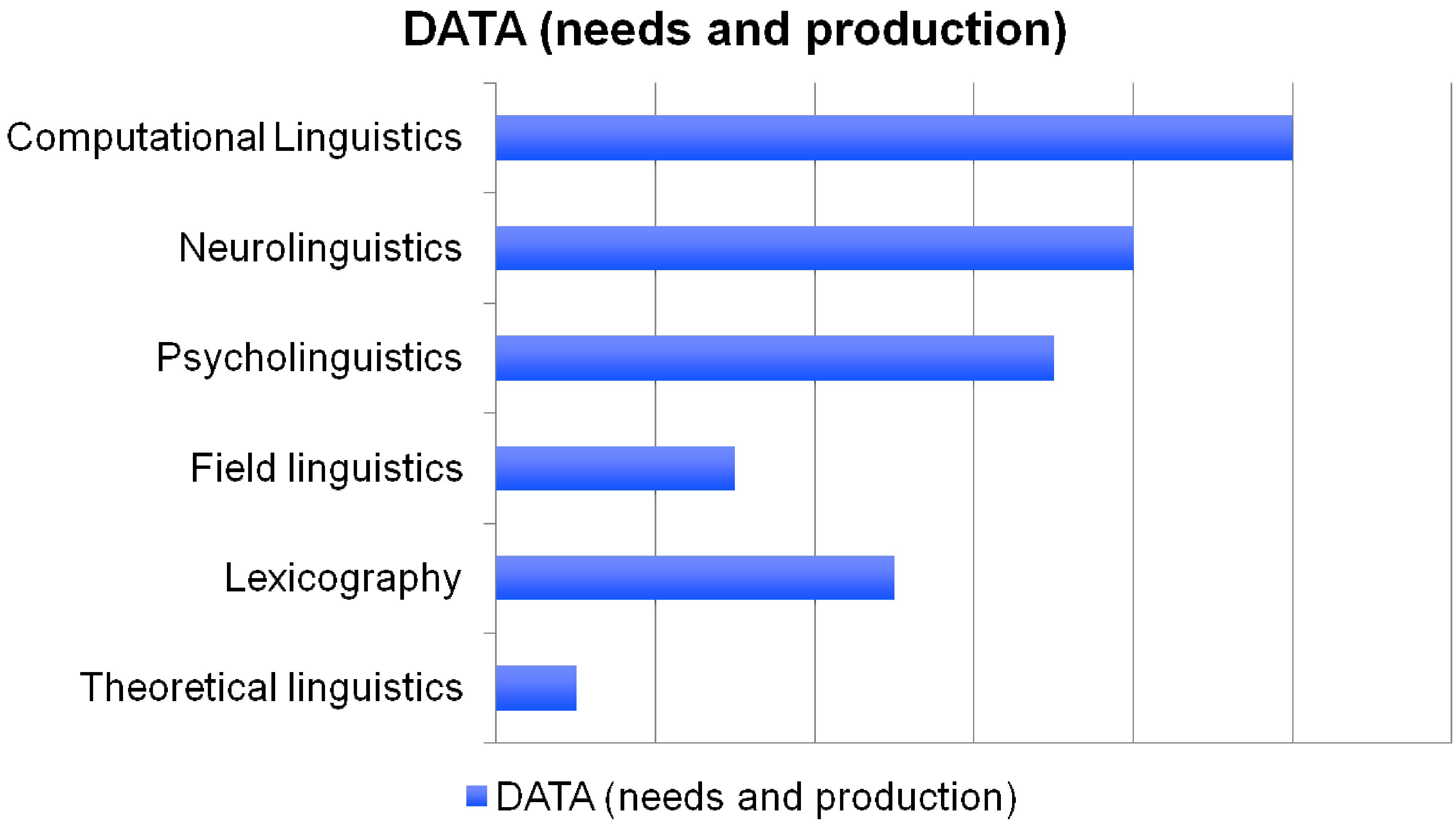
Linguistics

- Theoretical linguistics (morphology, syntax, semantics, …)
- Lexicology and lexicography
- Field linguistics (documenting languages)
- Psycholinguistics
- Neurolinguistics (FMRI, EEG, MEG)
- Computational Linguistics (Natural Language Processing)

Linguistics

- A study of language
- **A data science**
  - *Always has been*




DATA (needs and production)
Computational Linguistics
Neurolinguistics
Psycholinguistics
Field linguistics
Lexicography
Theoretical linguistics
DATA (needs and production)

Data collection and processing

- An integral part, a foundation of linguistics
- Used to be prohibitively expensive
  - Small scale
  - Manual processing
  - Time consuming
  - Inferences based on limited samples
- Computer Science changed it all:
  - Massive data (billions of words per language) available
  - Automatic processing possible
    - NLP – Natural language processing
- Linguists are not Computer scientists
- Linguistics departments don’t have the infrastructure

Data Needs

- Statistical methods: “More data is better data” (training)
- Evaluation of new methods in NLP
    - Assuring the same data are used
        - Availability of the data (licensing)
        - Exact identification of the version (PID)
    - Same with tools (segmenters, analysers, synthesisers)

Most linguists don’t have

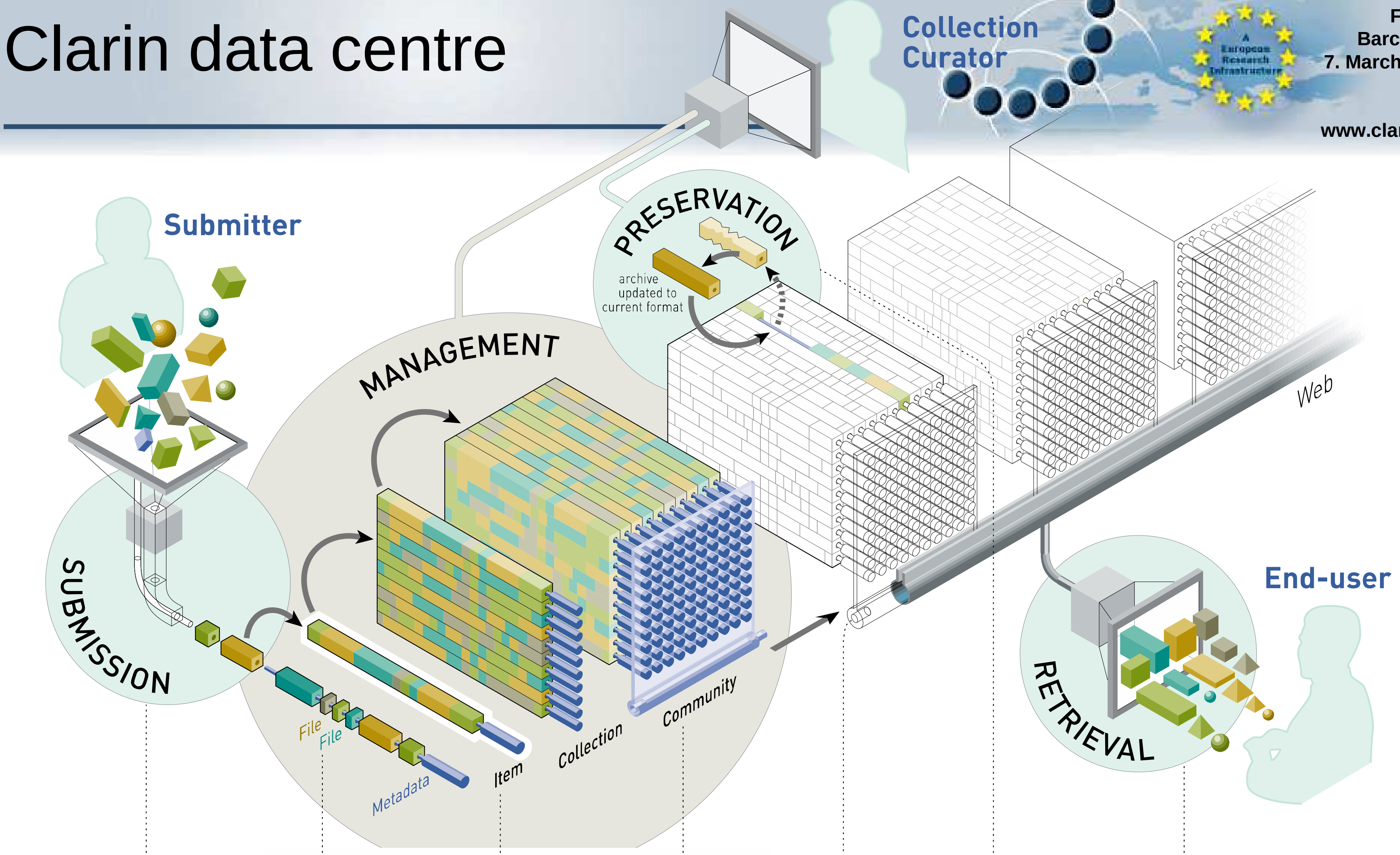
- Access to data: being able to find out what exists, and get it
- Facilities for large and safe storage and replication
- Computers for running demanding applications with big data
  - Big processing power, but sometimes also a lot of memory for a shared model, or fast storage, etc.
- Expertise for effective acquisition and processing of data

- **Enable eHumanities**:
  - **integrated:** the resource and service centres are connected
  - **interoperable:** to overcome format, structure and terminological differences
  - **stable:** the resources and services are offered with a high availability
  - **persistent:** the resources and services to be accessible for many years
  - **accessible:** the resources and services accessible via the web; different access methods and training possibilities are offered tailored to the needs of the communities
  - **extendable:** the infrastructure is open; new resources and services can be added easily

Clarin data centre

1st EUDAT User Forum
Barcelona
7. March 2012

www.clarin.eu



1 Web-based interface makes it easy for a submitter to create an archival item by depositing files. DSpace was designed to handle any format from simple text documents to datasets and digital video.

2 Data **files**, also called bitstreams, are organized together into related sets. Each bitstream has a technical format and other technical information. This technical information is kept with the bitstreams to

3 An **item** is an "archival atom" consisting of grouped, related content and associated descriptions (**metadata**). An item's exposed metadata is indexed for browsing and searching. Items are organized into **collections** of logically-related material.

4 A **community** is the highest level of the DSpace content hierarchy. They correspond to parts of the organization such as departments, labs, research centers or schools.

5 DSpace's modular architecture allows for creation of large, multi-disciplinary repositories that ultimately can be expanded across institutional boundaries.

6 DSpace is committed to going beyond reliable file preservation to offer **functional preservation** where files are kept accessible as technology formats, media, and paradigms evolve over time for as many types of files as possible.

7 The end-user interface supports browsing and searching the archives. Once an item is located, Web-native formatted files can be displayed in a Web browser while other formats can be downloaded and opened with a suitable application program.

Visual Explanation by Dynamic Diagrams
www.DynamicDiagrams.com

Language Processing Services

- Computer Science doing Linguistics
  - Speech recognition (speech-to-text) and synthesis (TTS)
  - Machine translation
    - Also for multilingual IR
  - Grammar checking
    - Everybody needs it (to work)
  - Information retrieval (search engines)
  - Information extraction
    - “What new topics have appeared in particle physics in the last 2 years?”
  - Question answering
- Very much Statistics and Machine learning (from Data)

Produce…

- Annotated text corpora and video archives
- Aligned multi-modal and multi-lingual resources
- Methods for creating and searching the above ...
  - for anyone, not just linguists
- Everyone works with (and using) language:
  - historic archives
  - medical documents
  - scientific literature

Linguistic Applications (Services)

- Smart spell checking and grammar checking
- Machine translation
- Speech recognition (dictation, subtitling)
- Speech synthesis (text-to-speech)
- Dialog systems
- Automated indexing of audio and video files for searching

EUDAT can help with:

- Identification and availability of resources
  - Many linguists: many places to search, not up-to-date
  - Unified portal to get (language) data
  - Data from other communities often still is or includes language data.
    - Interesting for inf. retrieval and inf. extraction

EUDAT can help with:

- Data hosting and replication
  - MALACH video history: 135 TB. We can only store and present 10-20TB (most of Czech, Slovak and Polish)
- Number crunching
  - SMT (Czech-English): 200/230 mil. words
    - 3 days to get to a transl. model (on a modest cluster)

EUDAT can help with:

- Workspaces and web applications for data annotation and searching
  - Current annotation tools are developed ad hoc
  - Many tools, little to no long-term support
  - Exactly the same for search tools

EUDAT can help with:

- Running workflow services
- End-to-end experience currently unreachable:
  - Choose the data (searchable data repository)
  - Choose the analysis (services, workflow system)
  - Run the analysis (big cluster, big memory...)
  - Present the results and store them (persistently)

Thank you for your attention