



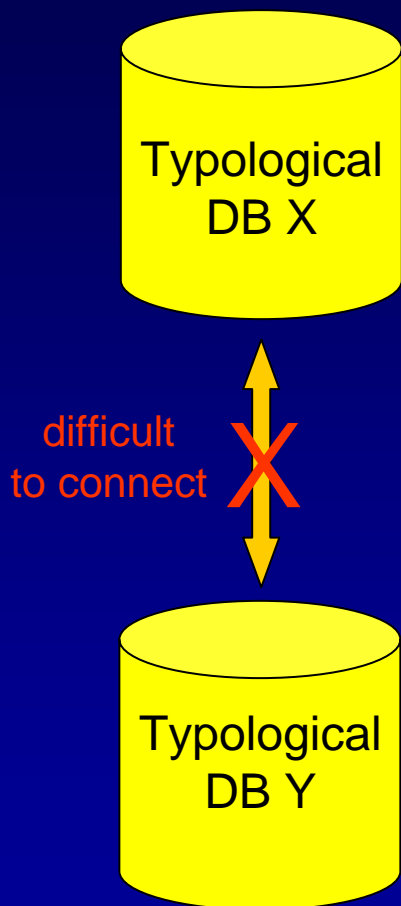
# Architectures for Distributed Language Resources

Peter Wittenburg  
Max-Planck-Institute for Psycholinguistics

[peter.wittenburg@mpi.nl](mailto:peter.wittenburg@mpi.nl)  
[www.mpi.nl](http://www.mpi.nl)



# Typological Databases Today

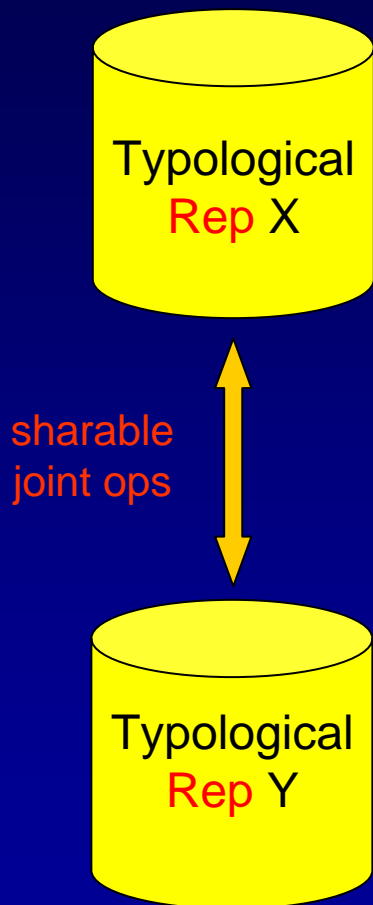


- each DB has its own individual
  - format (text, DOC, rDB, ...)
  - structure (tables, hierarchy, ...)
  - terms (labels, typo categories, CVs, ...)
  - access methods (web interfaces, ...)
- often definitions are not explicit
- many good examples today

True in many areas



# Towards a TypologyWeb: TYPEWEB



- what to do?
- how to set up a system?
- what is the price to be paid?
- exploiting rDBs not at all trivial

## similar activities in other areas

- keyword type metadata for L resources
- Math Net
- etc

Semantic Web as buzzword

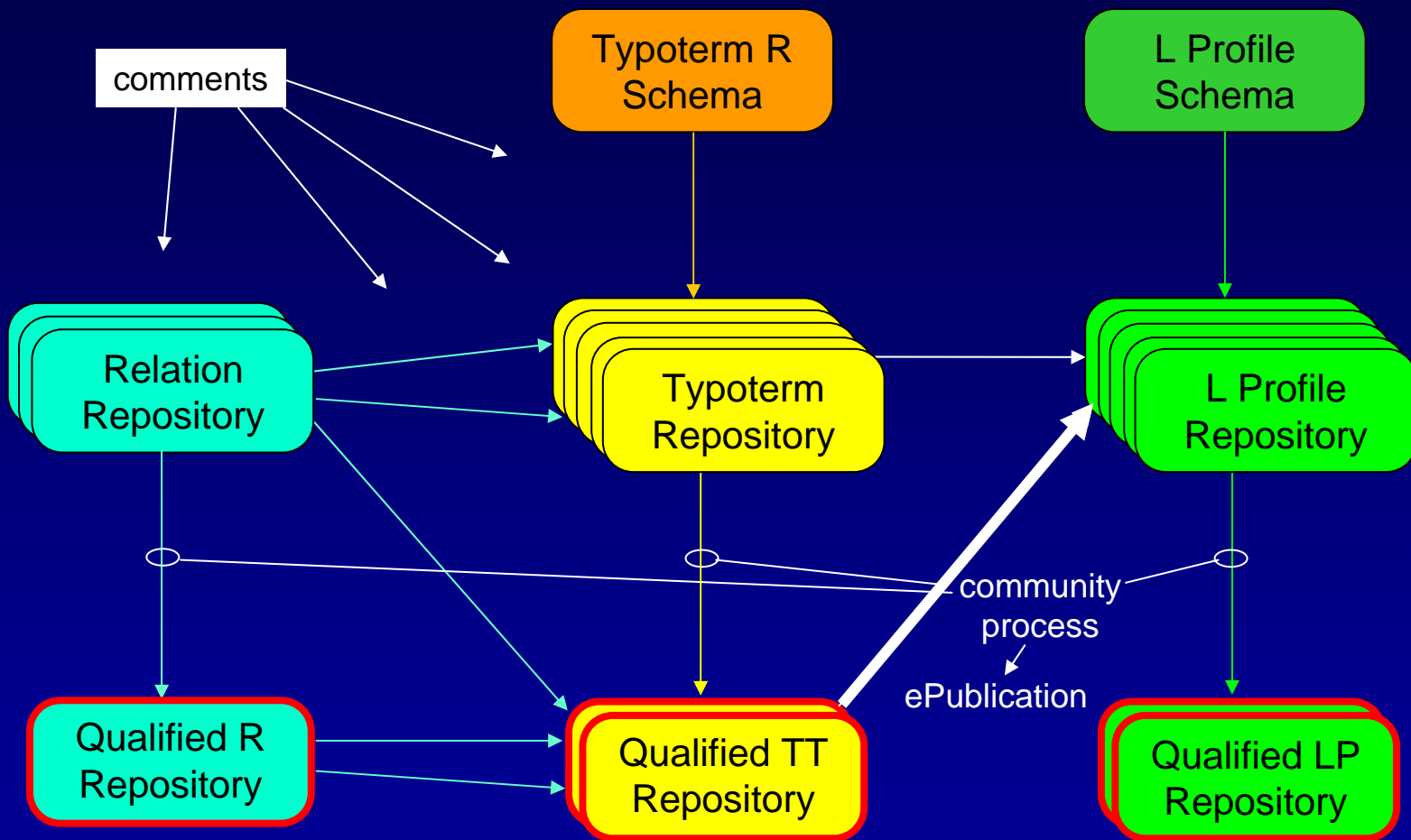


# Potential of Digital Internet Era

- almost nothing is really new
- the INTERNET has potential to bring resources together
  - primary resources  
(texts, photos, sound and video recordings)
  - all types of metadata  
(metadata is data about data, must be a relation)
    - **typological descriptions**
    - keyword type of metadata (DC, IMDI, ...)
    - annotations
    - ...
- what is different?
  - want to access the raw material (without retrieving old tapes)  
all is digital - hopefully
  - want to share and exchange data (without sending paper, CDROMs,...)
  - **want to join data** (without lots of conversion steps and more than simply create untyped hyperlinks)
  - **re-usage** (of whatever stuff from others)



# Open TypeWeb Domain



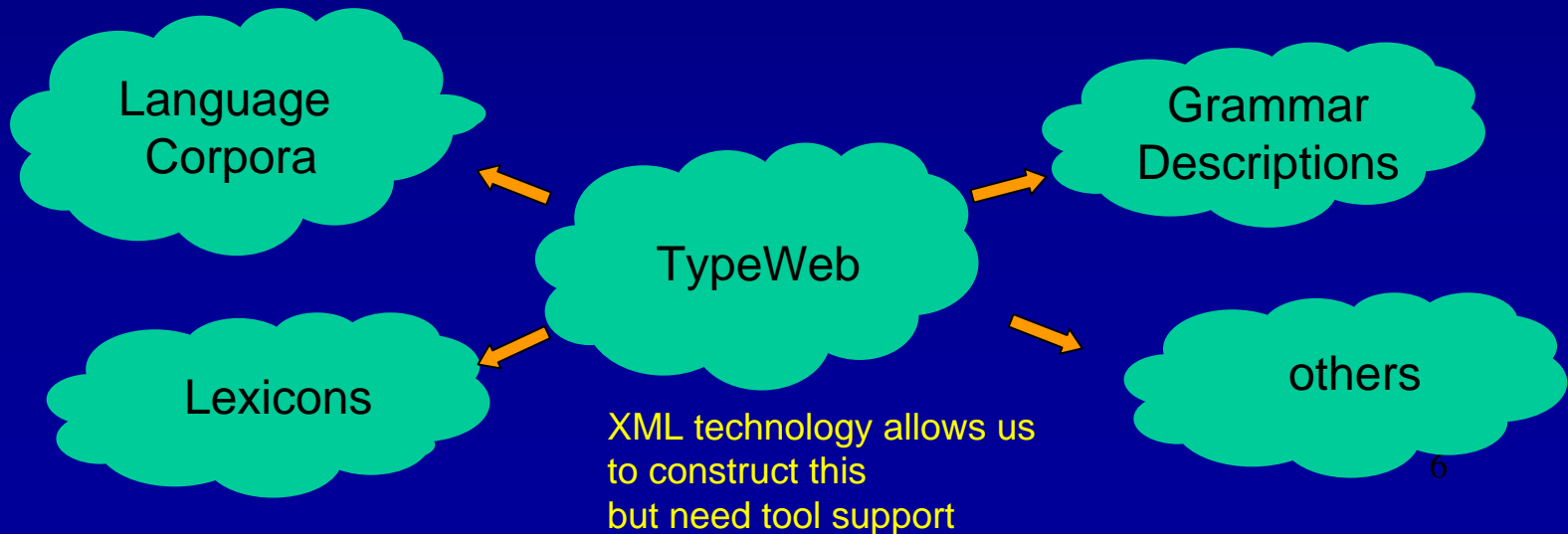
only works if we have good tools:

- editors, integrated & efficient search/browse, linkers, ...
- local & connected, caching, fast, ...



# Is that all?

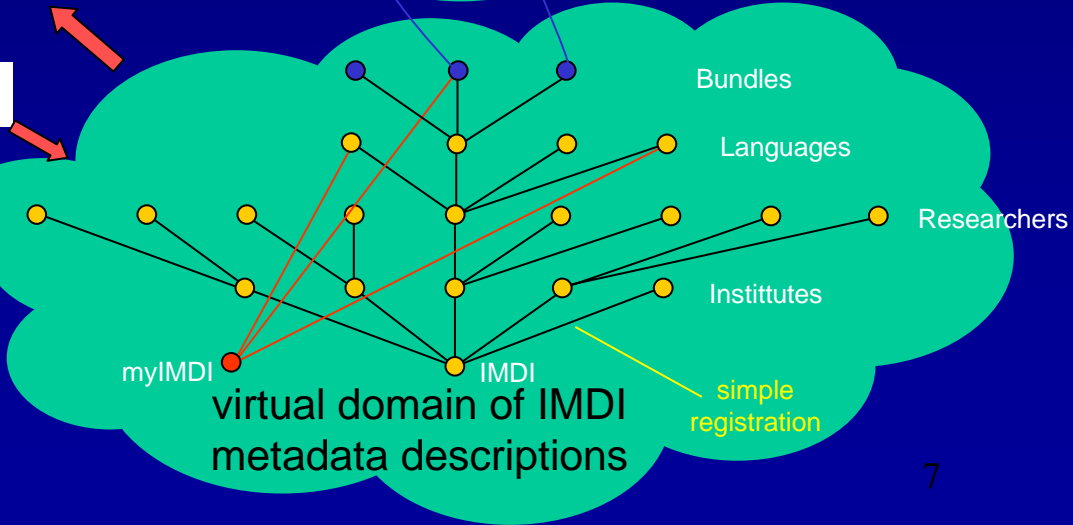
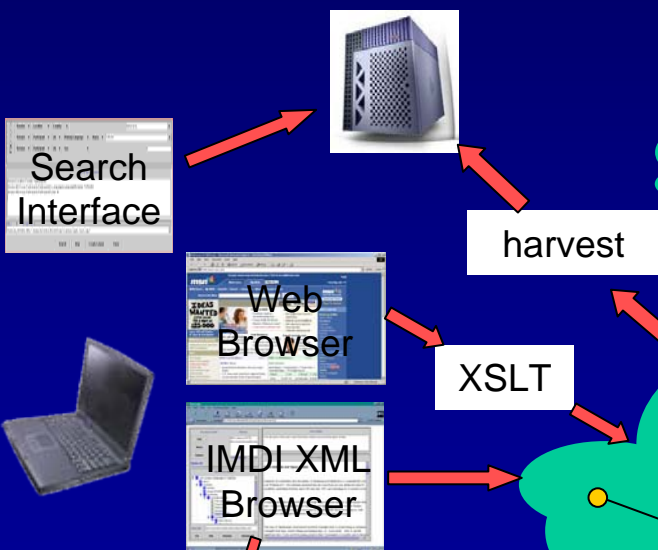
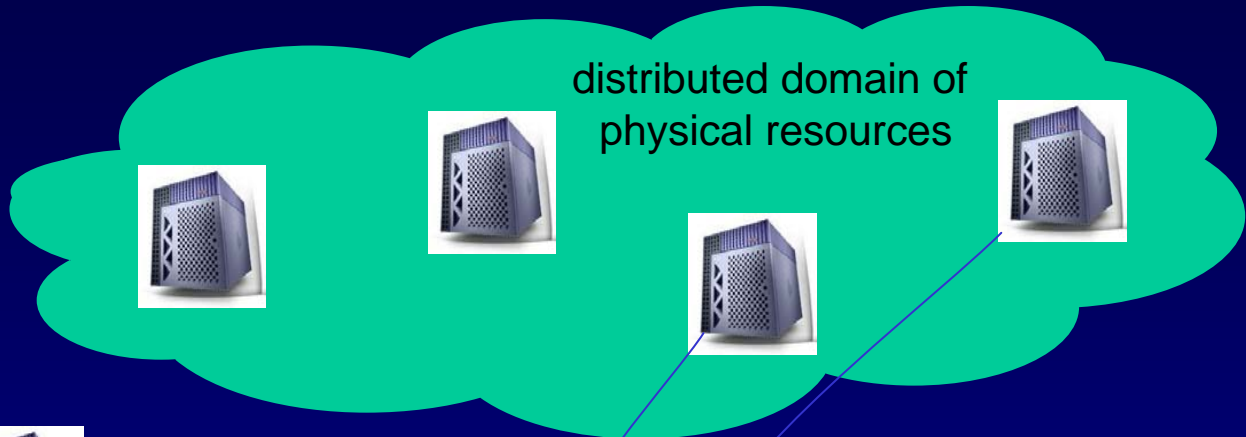
- NO
  - need framework of seamless web-services (UDDI, WSDL, SOAP)
    - transparent to linguists, i.e. support by tools
    - automatic registration if wanted
  - need to integrate the existing databases (as shown today)
    - probably not all
  - need harvesting and indexing services
  - want to relate TypeWeb to other open and shared resources





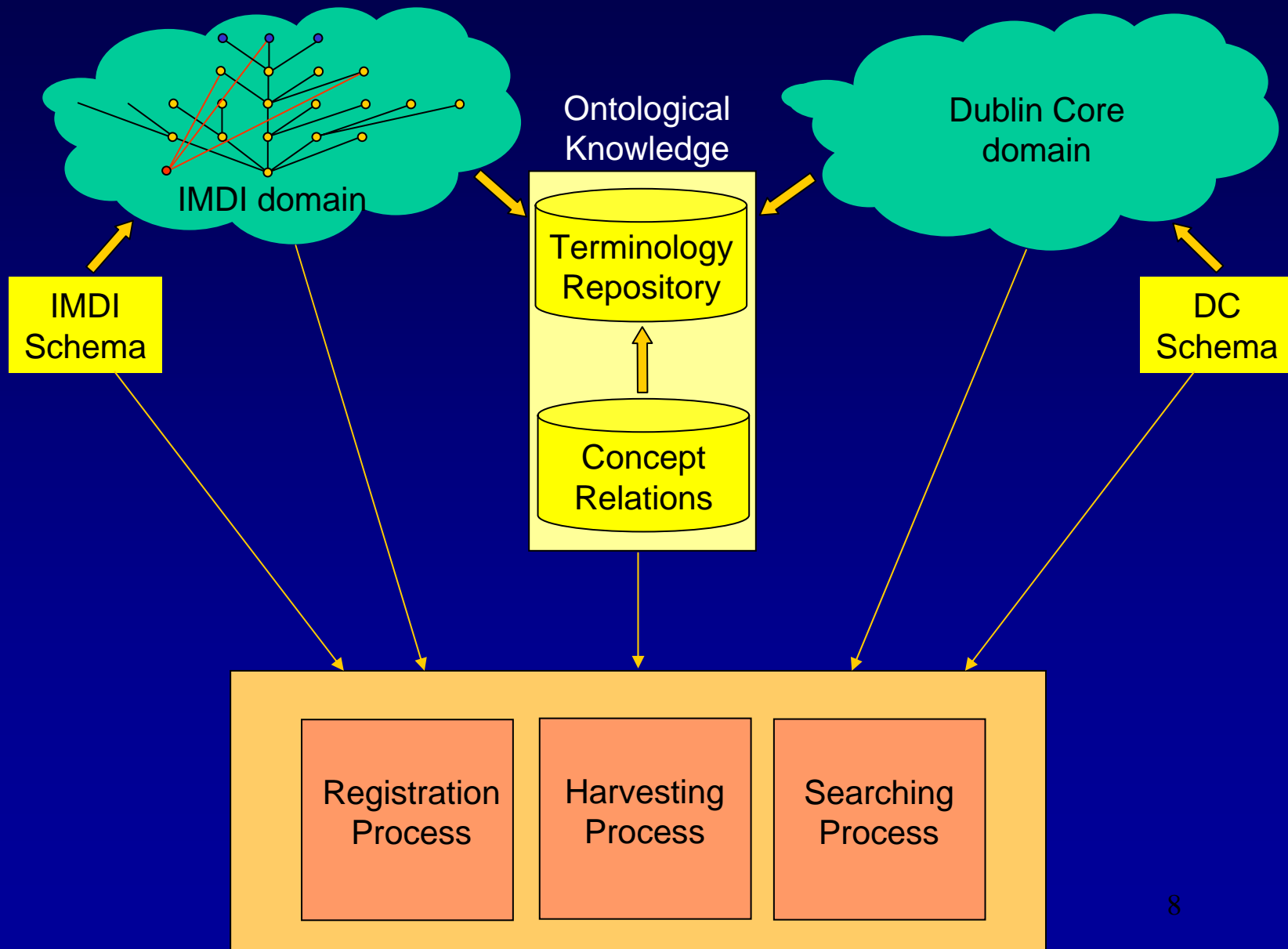
# Example IMDI Metadata Domain

this is operating





# IMDI Crossing Boundaries





# Price to be paid

- nothing comes for free - we know this
- have to do our work beforehand  
i.e. need agreements & openness

## simple ones

- for syntax of (textual) language resources (LR)
- for container/file formats
- for character encodings
- for registration

XML  
MPEGx, ...  
UNICODE  
UDDI, ...

## difficult ones

- terminology for encoding of phenomena
  - got a framework for defining semantics

various  
RDF/OWL

- all done for comparatively simple keyword type metadata descriptions of Language resources
  - selection and definition of elements
  - definition of controlled vocabularies where useful
  - internal structural relations
  - external relations (Dublin Core, OLAC, IconClass, ...)

IMDI



# Is it easy?

- NO
- people don't agree
  - different purposes
  - different granularity
  - different languages (many POS concepts)
- people hesitate to be explicit and formally describe categories
- there is no one (big) “ontology” for the domain of LR
- there will be many
  - some may have a “community” stamp
  - some will be comprehensive
  - most will be small and from small groups/individuals
  - semantics will change over time - history
  - etc etc
- community processes are very sensitive issues



# Related Web-Sites & Activities

- [www.mpi.nl/IMDI](http://www.mpi.nl/IMDI) technology for distributed LR metadata standardization project
- [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE) browsable and searchable LR domain
- [www.mpi.nl/INTERA](http://www.mpi.nl/INTERA) European Cultural Heritage Online
- [www.mpi.nl/ECHO](http://www.mpi.nl/ECHO) cross-disciplinary approach
- [www.mpi.nl/DOBES](http://www.mpi.nl/DOBES) Endangered Languages Documentation Project incl. metadata domain
- [www.mpi.nl/tools](http://www.mpi.nl/tools) MPI technology
- [www.mpi.nl/corpora](http://www.mpi.nl/corpora) metadata of MPI corpora
- ENABLER international project for LR integration
- ISO TC37/SC4 Management and Terminology of LR



# Acknowledgements & End

- TypeWeb is a result of intensive discussions with
  - **AUTOTYP** project (Balthasar Bickel, Joanna Nichols)
  - **TDS** project in NL (LOT - Utrecht, Amsterdam, Nijmegen)

***Hope that no one believes in fast solutions.***

***It will take a while to build the whole infrastructure (steps needed).  
It will take even longer to convince the community to use it.***