Experiment

“The ONE”

“The Semantic Web In One Day”

Schloss Dagstuhl, October 2004

Johanna Völker
Nenad Stojanovic
Peter Haase
Max Völkel

AIFB

FZI

ontoprise

We are …

- Johanna Völker
- Nenad Stojanovic
- Peter Haase
- Max Völkel

Our Idea

- What can we do in one day?
- Build a new application from scratch?
- Extend an existing application?
- Integrate two applications?
- → **Integrate three applications!**

TextTo nt

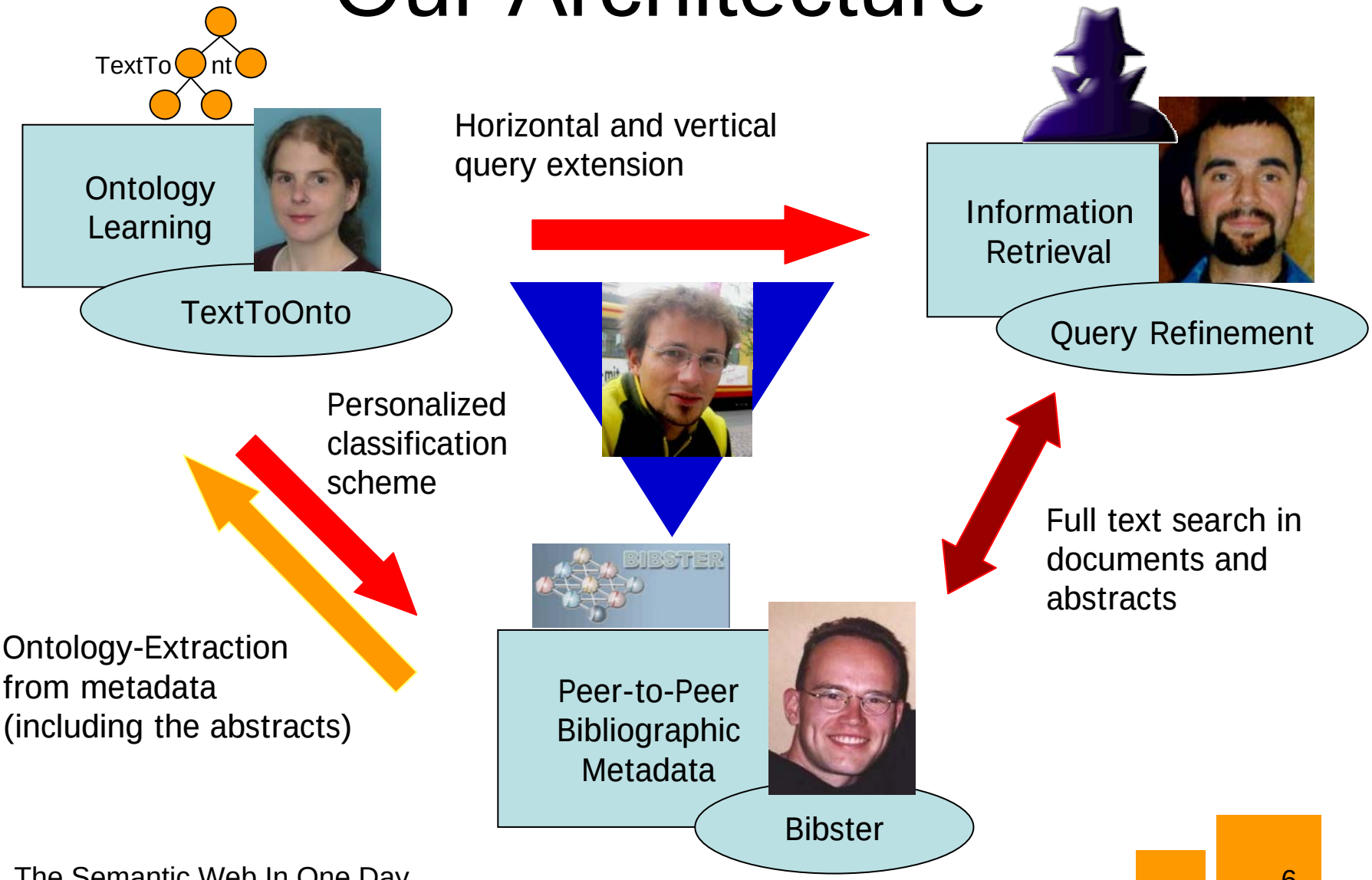
Motivation

- Bibliographic data is semi-structured:
  - Structured metadata
  - Unstructured full text
- Need for better browsing
- Need for better querying
- Need for integrated approach
  - Integrated retrieval
  - Integrated view on data

Goals

- Ontology-based browsing with a custom-learned ontology
- Personalized ontology-driven query refinement
- Efficient, integrated management of metadata and full texts

Our Architecture



Data Preparation

CiteSeer.IST
Scientific Literature Digital Library

6000 files with 100 XML-fragments each = 600.000 entries

- Wrap XML.bat
- Parse to DB

DB2

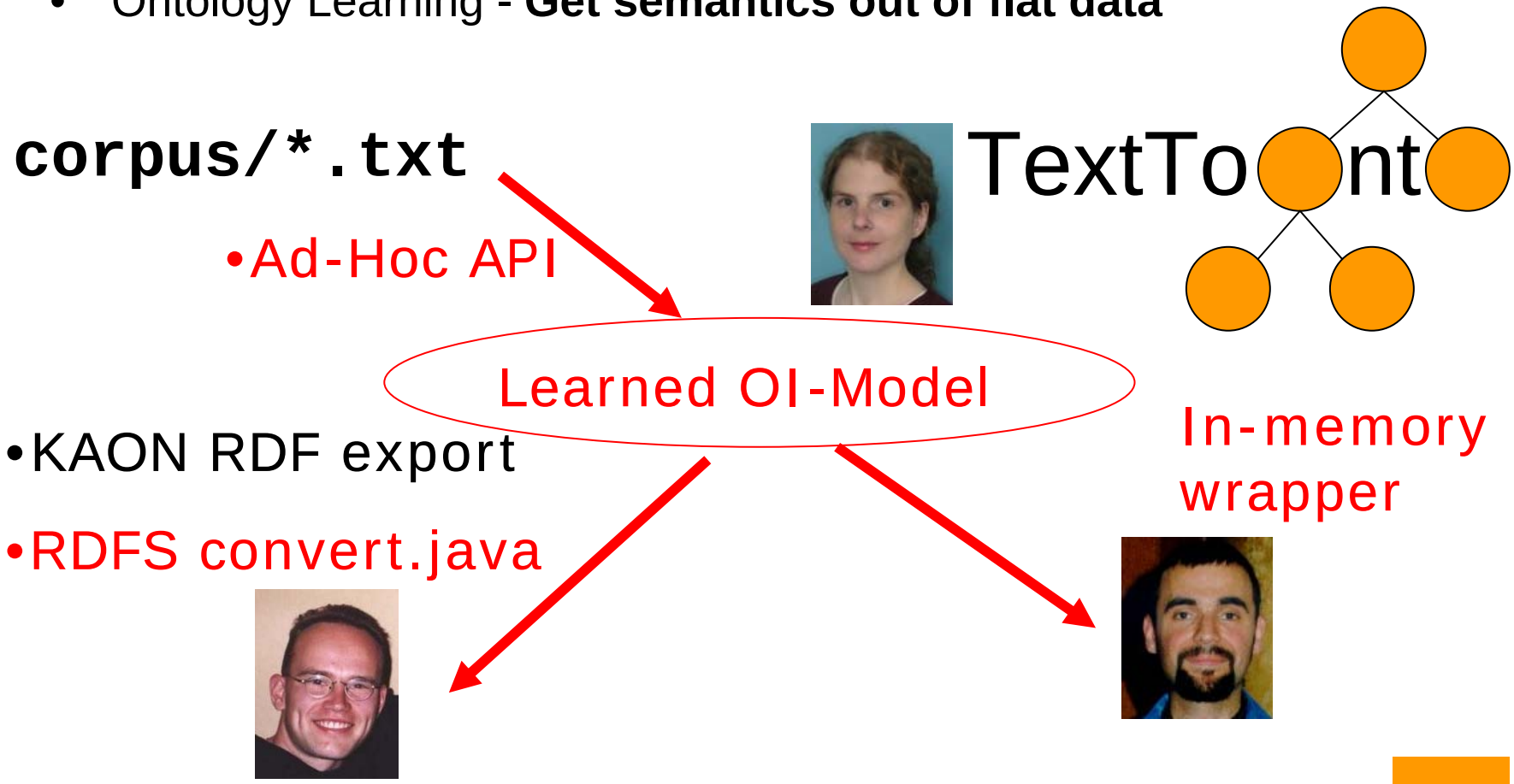
- Export

- Export

→ **corpus/*.txt**

Used Technologies

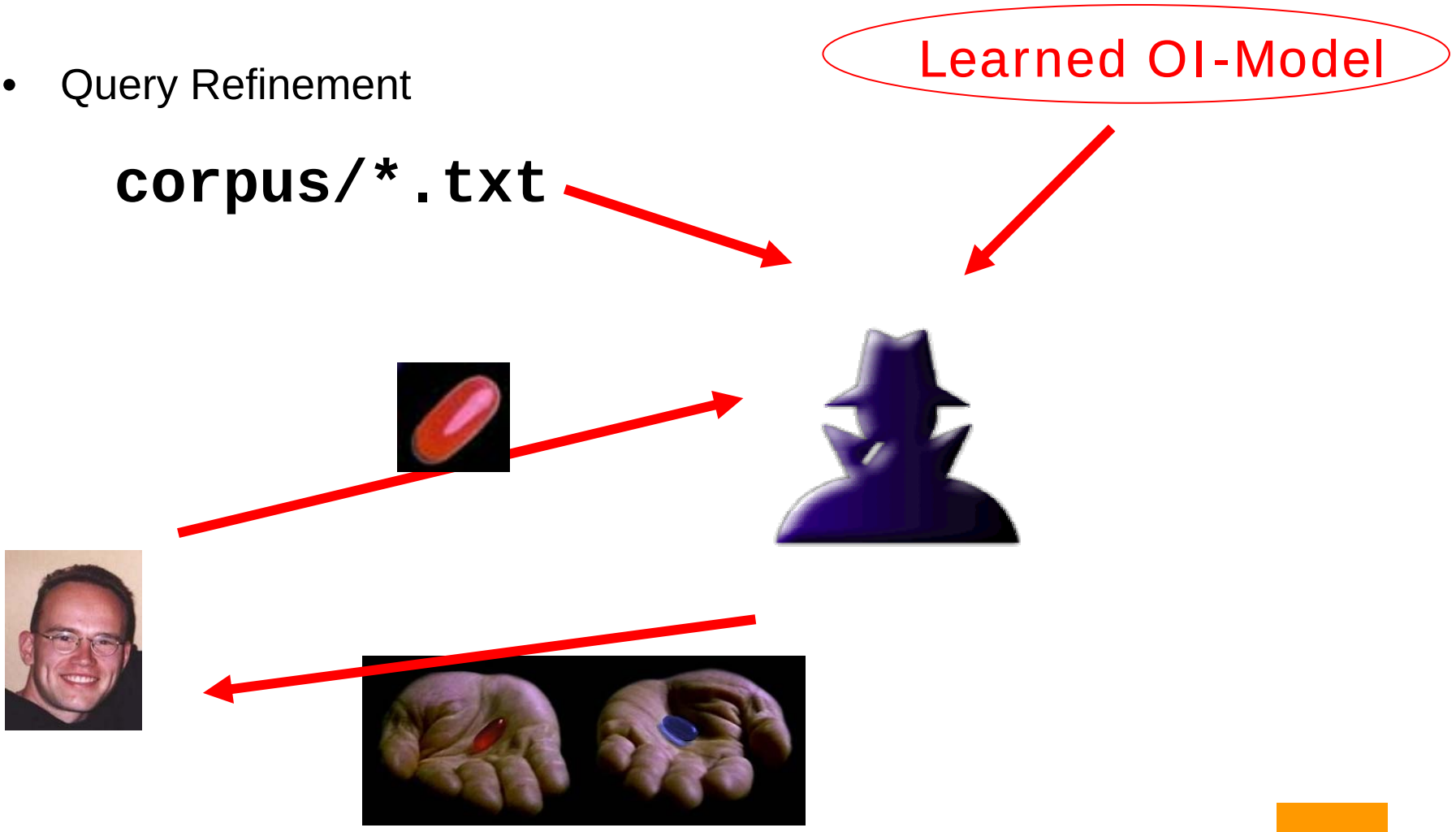
- Ontology Learning - **Get semantics out of flat data**



Used Technologies

- Query Refinement

```
corpus/*.txt
```



Our Lessons Learned

- Integration is possible in 24 hours even with a large data set: 0.6 million entries
- Trouble in the team
  - Common agreement on integration goals
  - Common language
- Design
  - Mapping real-world task to an integration architecture
  - Which data sources are integrated? → Quality requirements depend on data consumer
  - Amount of data? Performance → Create meaningful subsets
  - Data format(s)? → Use exiting converter! → Might be buggy
- Data integration easier by using semantic data
  - Nenad uses relations as pairs
  - Peter uses strict taxonomy
- Syntactic transformations → **hard**
- Code integration → **not that hard**
- Create interfaces
- Simulate integration steps
- Test early, test often
- Don‘t forget GUI adaption
- Check library versions
- Trouble with configuration of logging, tools

Outlook

- Explore Cooperative Answering with learned ontology from corpus
- Ontology evolution suggestions
  - Usage-driven: browsing log
  - Data-driven: corpus changes → runtime performance
- Second generation tools make integration easier
  - Text2Onto, KAON2, …
- Peer-to-peer aspects:
  - Collaboration
- Links from metadata to Amazon
  - Sell proceedings!