Russian Academy of Sciences
Program Systems Institute

Text Mining: Knowledge Extraction and Management

Overview

Why text mining?

Information is overwhelming

Information is crucial to decision-making

Only part of relevant information can be found in the form of *structured data* (databases, spreadsheets, Web forms etc.)

Unstructured data, or texts, remain the premier source of vital information

Problem Statement

Large amounts of unstructured text content in various fields

- e-Science
- Business Intelligence

Natural language processing (NLP) is computationally intensive

Data mining approach applies to structured content only

Efficient means of unstructured content management and analysis are required

Goals

Combine different NLP techniques to provide an efficient knowledge management solution

- text categorization
- information extraction
- information retrieval

Use GRID technology for distributed natural language text processing and storage

Enhance existing text mining technology

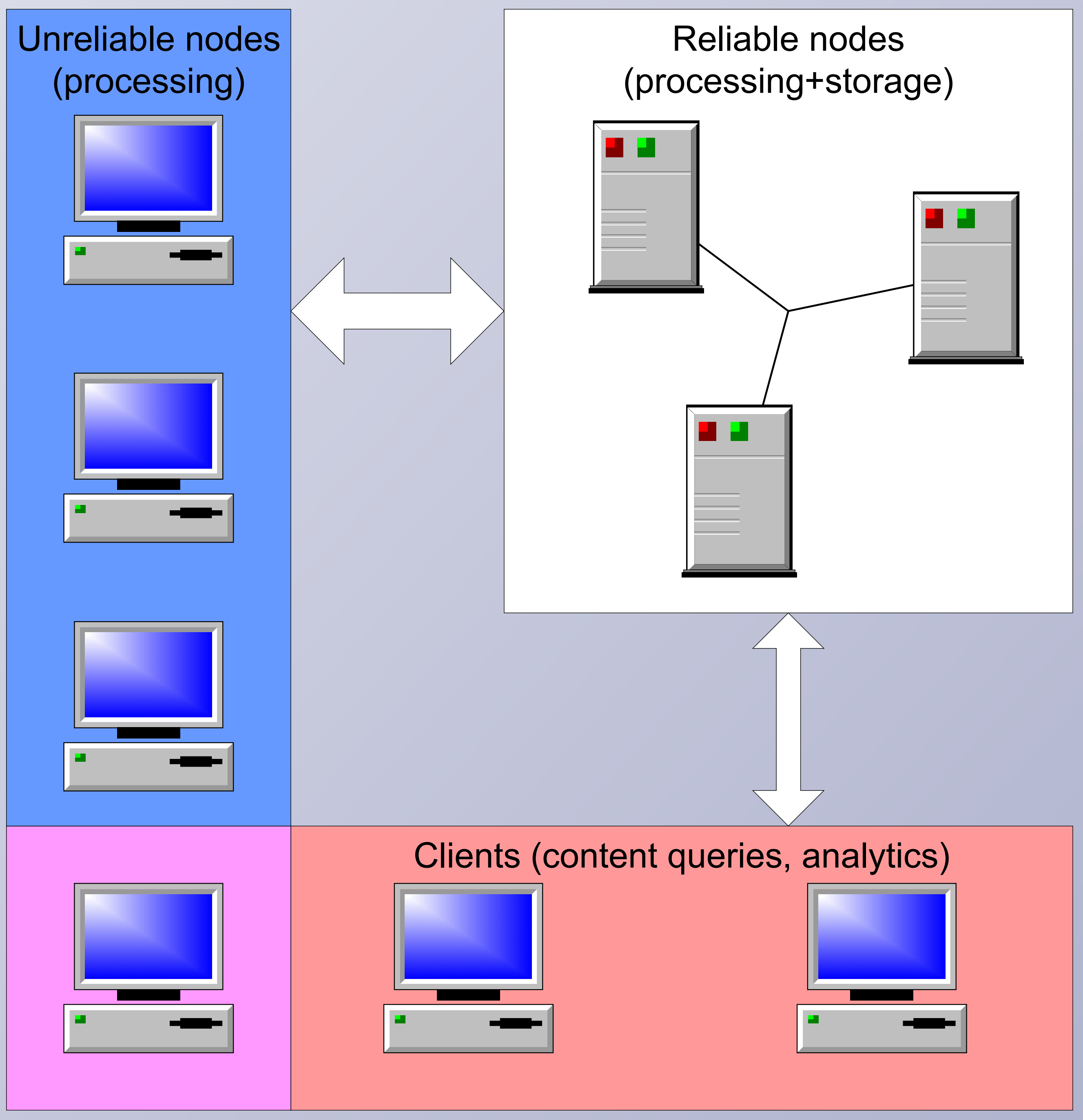
Issues

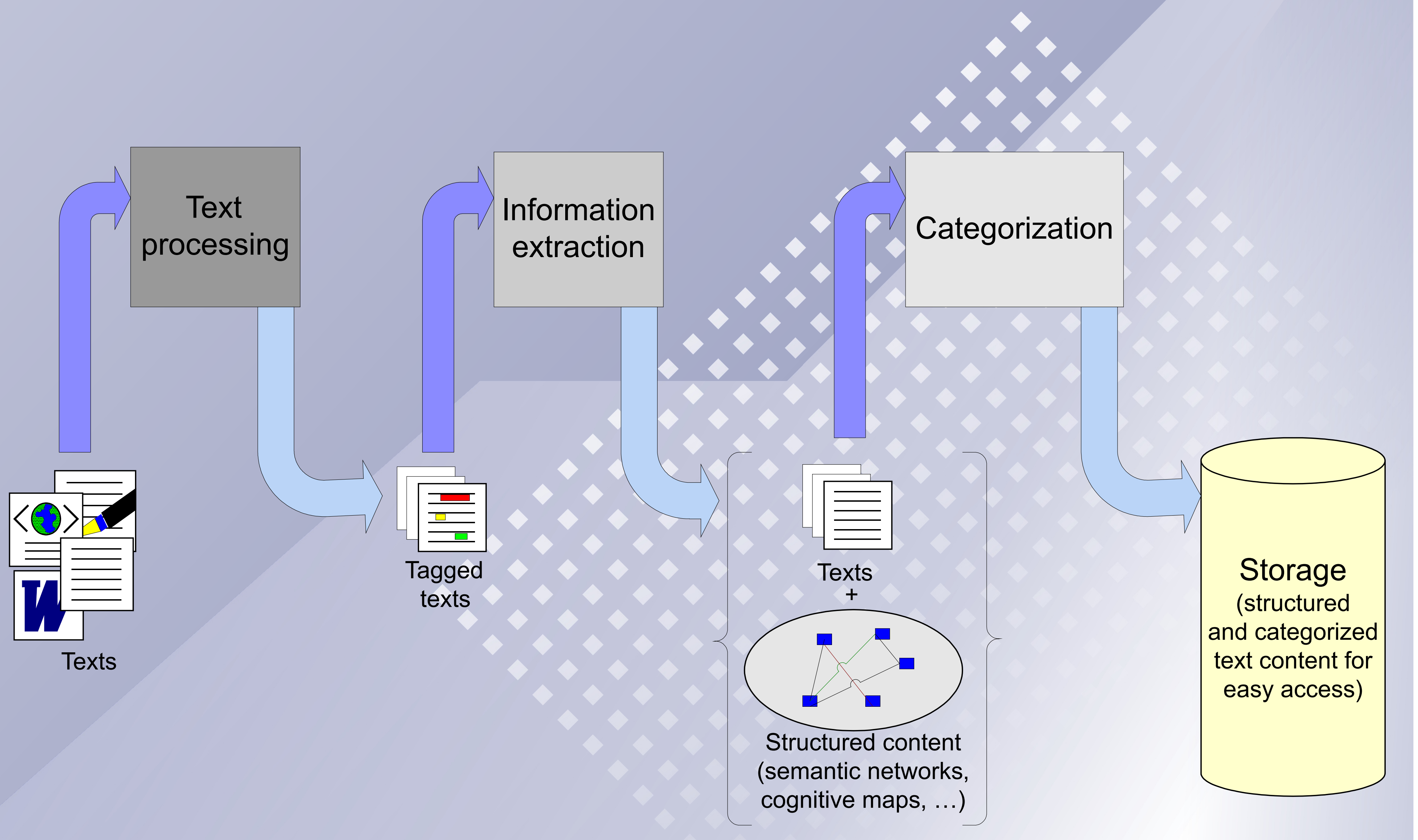
- Portability between domains
  (e.g. documents in e-Science and Business Intelligence domains differ a lot in language, concepts and writing style)
- Handling multilingual content
- Poor quality of real-life texts
  (documents may include speech transcripts, optically recognized texts etc.)
- Software portability

Status

Technologies and software tools developed:

- INEX a portable system for information extraction
- AKTIS a portable text categorization tool
- TGRID distributed computations in a cross-platform network environment





ADDRESS

Artificial Intelligence
Research Center
Program Systems Institute
Russian Academy of Sciences

Pereslavl-Zalessky
Yaroslavl Region
Russia, 152020
Tel/Fax: **+7 (48535) 98065**
E-mail: airec@botik.ru
Web-site: http://www.botik.ru/PSI/AIReC