Semantic Search in P2P-based Digital Libraries

Hao Ding

Information Management Group,
Norwegian Univ. of Science and Technology
JCDL 2005 Doctoral Consortium
June 7th, 2005, Denver

NTNU
Innovation and Creativity

Agenda

- Motivation
- Research Questions
- Significant Problems
- Existing Solutions
- Proposed Approach
- System Architecture
- Current Status
- Future Work

NTNU
Innovation and Creativity

Motivation

The Digital Libraries (DL) of the future shall enable users to access collections in various forms at any time, from anywhere, and in an efficient and effective way.

Research Questions

- Is it feasible to interlink largely distributed DLs by applying innovative system infrastructures, such as P2P?
- If could, is it then applicable to search across dispersed and heterogeneous metadata records by adapting semantics-oriented approaches, such as metadata crosswalks and ontologies?

NTNU
Innovation and Creativity

Significant Problems

- From the perspective of system infrastructure:
  - Largely dispersed and isolated DLs, especially, *small and medium-sized DLs.*
    - Requirement in *Independence*
    - Requirement in *Extensibility*
  - DLs are generally autonomous
    - Requirement in *Modularity*
- From the perspective of information :
  - Heterogeneous metadata schemas
    - Requirement in *Heterogeneity*

NTNU
Innovation and Creativity

The State of Existing Solutions

- System infrastructure
  - Client/server
  - Peer-to-Peer
  - Hybrid approach. i.e., Super-peer based architecture
- Data integration
  - Design-time analysis:
    - Schema integration
    - Data Warehouses
    - E-Commerce
  - Run-time analysis:
    - Semantic query processing
      - i.e., user specifies the output of a query. (SELECT...FROM...WHERE...)

The State of Existing Solutions

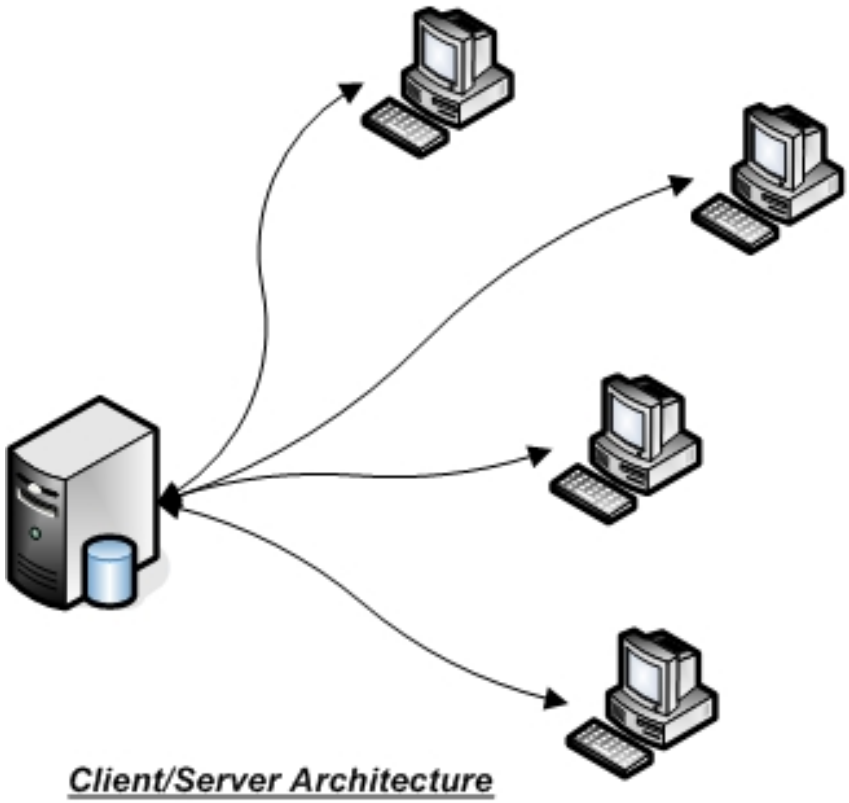
- Much research is carried out in the database and information retrieval communities.
  - i.e., table-level schema mapping in P2P network
  - i.e., keyword-based search in P2P network
- Little focuses on semantic search in digital library community, especially in the setting of P2P network.
- Related projects.
  - Edutella (Wolfgang Nejdl)
  - P2PIR (Norbert Fuhr)
  - Bibster (Steffen Staab)
  - FreeLib(Kurt Maly)
  - Metadata3 (md3)

Proposed Approach

- System Infrastructure
  - Peer-to-Peer (P2P)
    - *Extensibility*: Scalable
    - *Modularity*:
      - flexible
      - Allow loosely coupled peers
    - Low cost
    - ...
- Heterogeneity in metadata schemas

NTNU
Innovation and Creativity

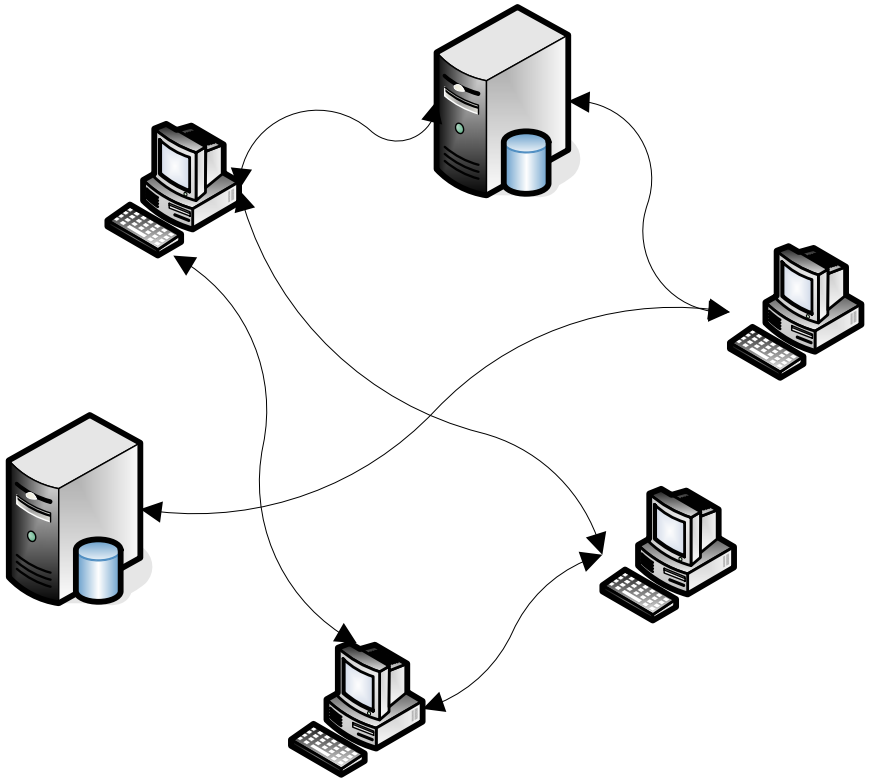
Proposed Approach (Con'd)



Client/Server Architecture

NTNU
Innovation and Creativity

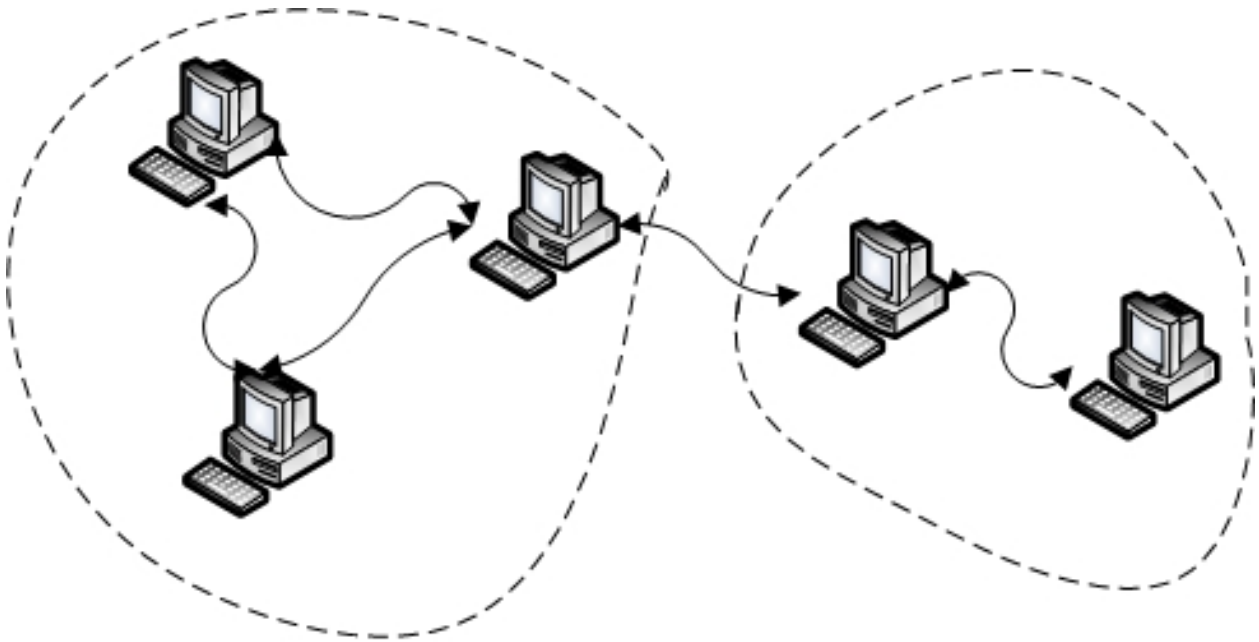
Proposed Approach (Con'd)



Pure P2P Architecture

NTNU
Innovation and Creativity

Proposed Approach (Con'd)



Super-peer Architecture

NTNU
Innovation and Creativity

Proposed Approach (Con'd)

- Platform: JXTA-based infrastructure, not DHT-based one.
  - Modularity:
    - DHT: requires a global distributed hash.
    - JXTA: Peers are rather independent
  - Search:
    - DHT: good at locating peers, not in multi-keywords search.
    - JXTA: does not have such a limitation.
  - JXTA supports super-peer approach
    - We focus on ONE super-peer (peer community) in current stage.
    - (System is extensible in this direction)

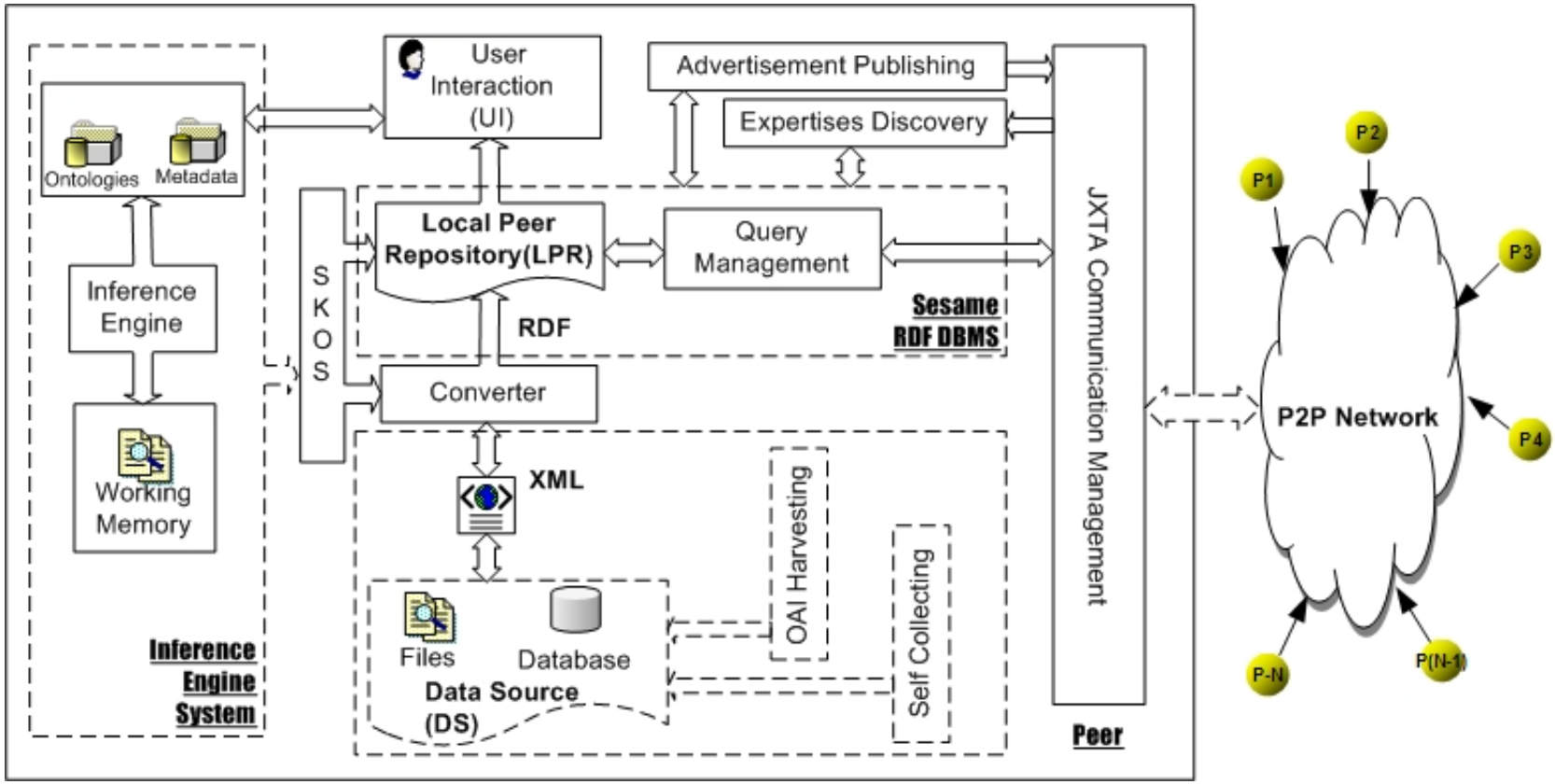
NTNU
Innovation and Creativity

Proposed Approach (Con'd)

- Heterogeneity in metadata schemas
  - Structure (not in this thesis)
  - Syntax (not in this thesis)
  - **Semantics**
    - Metadata crosswalks
      - mapping to Dublin Core (DC)
      - mapping to qualified DC is supported for complex metadata. (optional)
    - Ontology approach
  - Pragmatics? (not in this thesis)

NTNU
Innovation and Creativity

System Architecture



Semantic Search

- Goal: to improve traditional ‘search’ technologies by using machine understandable data.
- Presumption: libraries contain records with relation among each other
- In this thesis, search is conducted over RDF formatted repositories by applying domain specific ontologies. A RDF database (SESAM) with support for RDF schema inferencing and querying is applied.

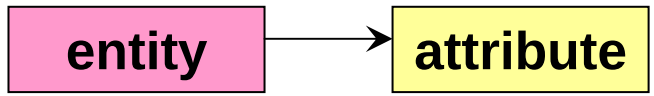
Semantic Search -Sample Query

- **Example:** Given the search term “Harry Potter”
  - Is a book (i.e., dc.title = “harry potter”)
  - Written by Joanne Kathleen Rowling (i.e., onix.Author=“Joanne Kathleen Rowling ”)
  - One of the DVD movies on eBay’s auction

NTNU
Innovation and Creativity

Analysis - Views on Metadata

Attribute View:



- attribute view – simplest, most direct:
  - i.e., dc.identifier.isbn “0439784549”
  - i.e., dc.creator= “J. K. Rowling”
- values may be strings, IDs etc.

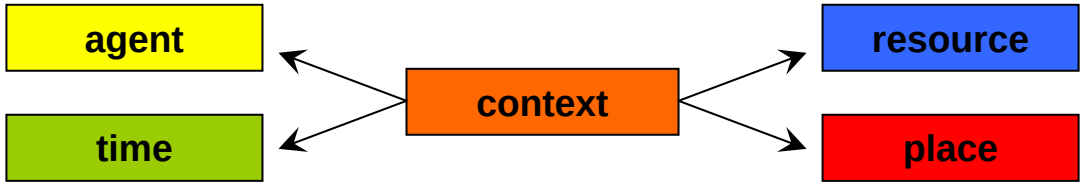
Analysis - Views on Metadata

Relationship: 

- association or relationship view – richer, more indirect:
  - book “0439784549” hasTitle “Harry Potter”
- treats attributes as defined entities
- and others e.g.
  - book “0439784549” hasAuthor “J.F. Rowling ”
- allows multiple occurrences

NTNU
Innovation and Creativity

Analysis - Views on Metadata

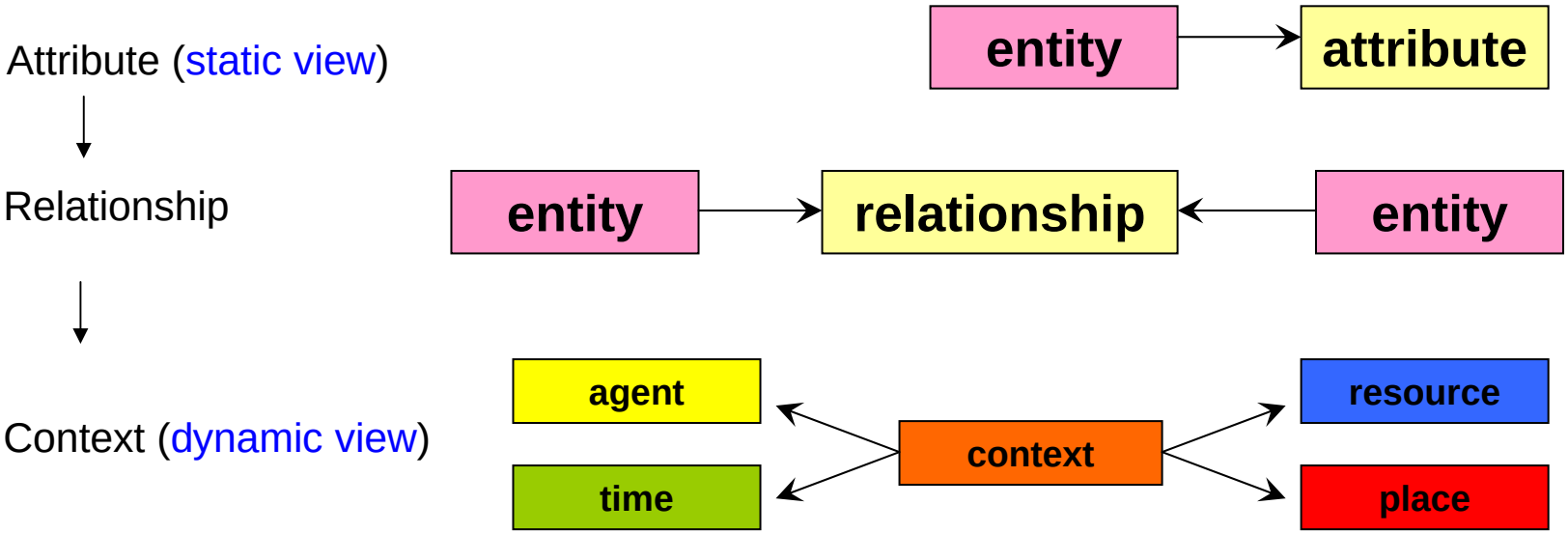


- context view – richest, most indirect
  - i.e., publishingEvent hasPlaceType placeOfPublication “US”
- Analysis moves from attribution to attribution process (Event)
- Most efficient handling of complex multiple metadata
  - i.e., ABC Ontology, CIDOC, etc.
- Allows analysis of complex relationships and meaning

NTNU
Innovation and Creativity

Analysis - Views on Metadata

- Three levels of attribution, moving from simple (static) to richer (dynamic events):



Tentative Results

- An ontology approach uses the deeper view of metadata
  - A upper-level ontology in specific domain,
    - i.e., bibilographic records.
    - Can not be too complicates, such as MARC.
    - Can not be too simple, such as that in Simple Knowledge Organization System (SKOS).
    - Support dynamic view (context)
  - Mapping
    - Crosswalk
  - Inference

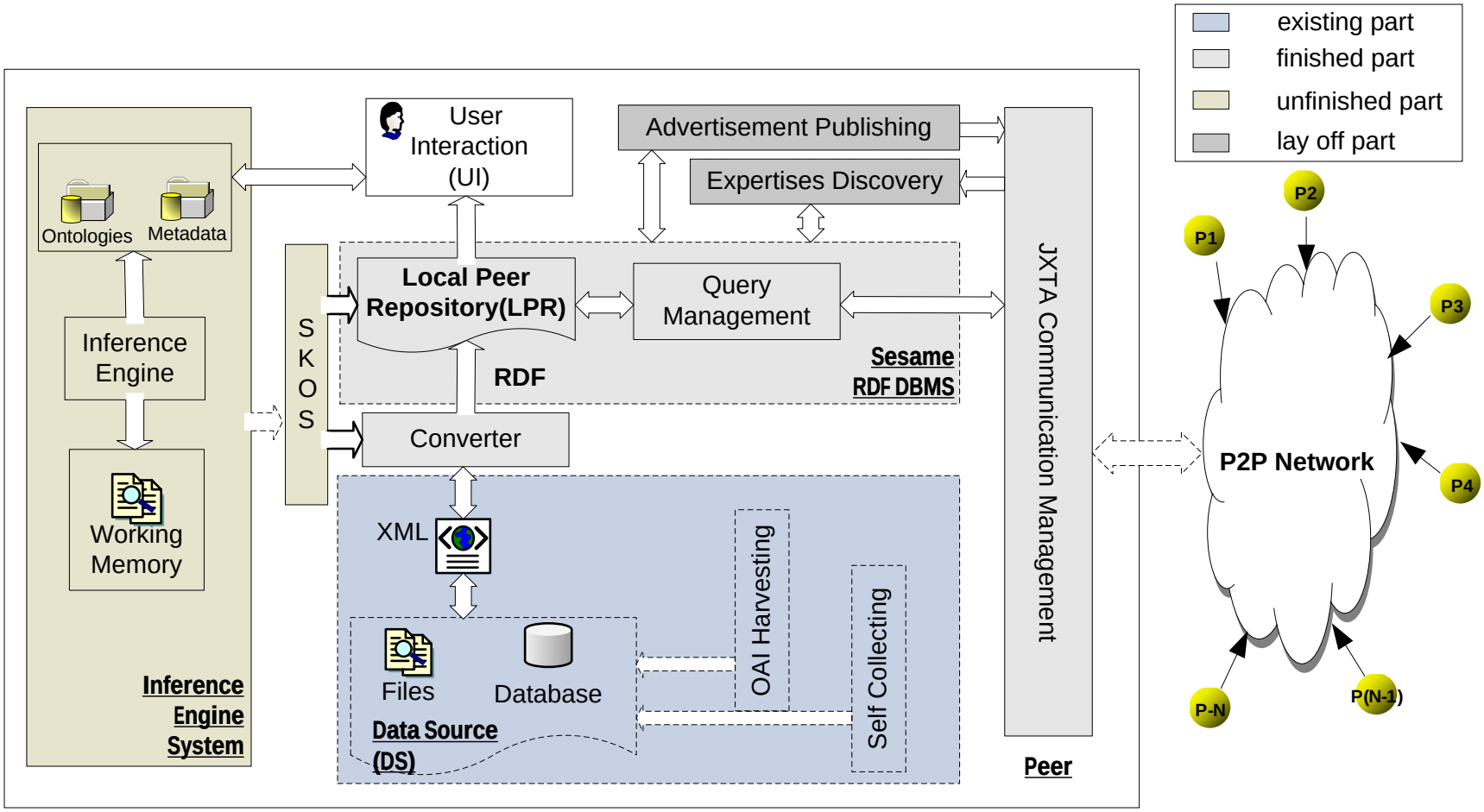
NTNU
Innovation and Creativity

Current Status

- JXTA framework has been implemented
- Applying a RDF database – Sesame for storage and query
- Homogeneous metadata – DC is supported currently.
- Experimental data:
  - OAI harvester tool oaiarc-0.97.2 (conversion required)
  - INEX collection

NTNU
Innovation and Creativity

Current Status (in picture)



The Way Forward

- More heterogeneous metadata are to be imported into the system by exploiting corresponding crosswalks.
- Deeper analysis on how to execute inference in search procedure.
- Applying Simple Knowledge Organization System (SKOS) [W3C] as schema integration language.
- Implementing inference engine component.
- Evaluation

NTNU
Innovation and Creativity

Questions?

NTNU
Innovation and Creativity