iMeMex
Escapes from the Personal Information Jungle

Jens-Peter Dittrich **Marcos A Vaz Salles** **Donald Kossmann** **Lukas Blunschi**

ETH Zurich, Switzerland

Problem

1. **Data Silos**
 - Desktop Computers use file based storage
 - Each file is a data cage
 - Each application reinvents its own ways of storing and searching information
2. **Lack of Query Processing Capabilities**
 - How to do queries that go beyond simple keyword search?
 - How to search for stuff exploiting structure/schema information?
 - How to join data from different files?
3. **Lack of Information Management**
 - How to store structured information?

iMeMex Benefits

- Provides logical information management layer on top of OS
- Brings information management, IR and data mining capabilities to your desktop
- Platform independent: works on Windows, Linux, Mac
- Extensible through plugins
- Extensible through operators
- Manages your entire data space
- Provides physical data independence
- Decouples content addressing from content storage

Vision

1. Do not build yet another application on top of the operating system!
2. Manage the users entire data space with a single system: iMeMex.

DOs

1. Handle unstructured, semi-structured and structured data pieces in a single system.
2. Make long-matured DBMS, IR and data integration technology available on the desktop.

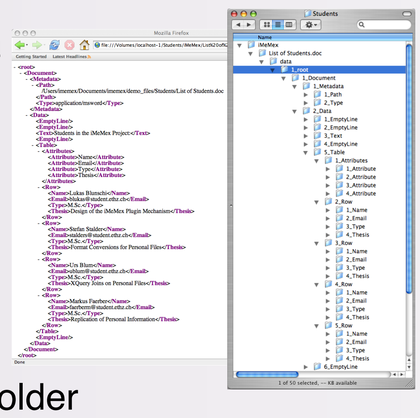
DONTs

1. Do not force application providers to change their code.
2. Do not force the user's data into schemas.

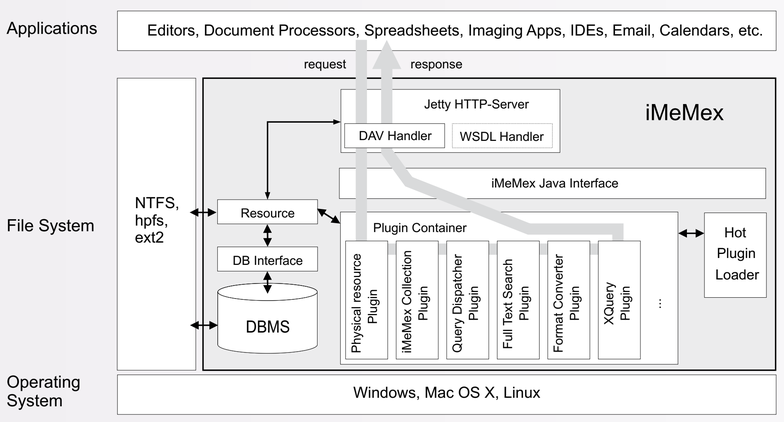
Plugins

Physical Resources
 what the file system does
Metadata
 display metadata of resources
Data
 browse through hierarchical data (XML)

iMeMex Collection
 display iMeMex folder



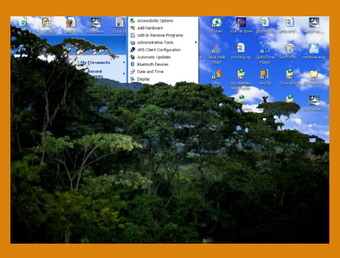
Architecture



Views on your Desktop

```
<?xml version = '1.0' encoding = 'utf-8'?>
<imemex-query>
    <alias>
        <realname>file:///C:/tests/pim/ROOT/bla/student projects.xls</realname>
        <name-in-query>file1</name-in-query>
    </alias>
    <alias>
        <realname>file:///C:/tests/pim/ROOT/bla/student emails.doc</realname>
        <name-in-query>file2</name-in-query>
    </alias>
    <xquery><![CDATA[
        <result>{
            for $a in doc("file1")//Student,
            $b in doc("file2")//Name
            where $a/text() = $b/text()
            return  <person>
                    <name> {$a/text()}</name>
                    <projectType> {$a/../Type/text()} </projectType>
                    <email> {$b/../Email/text()} </email>
                </person>
        }</result>
        ]]>
    </xquery>
    <sql/>
    <search/>
    <output-format>xls</output-format>
</imemex-query>
```

Consequences



- A desktop computer is a jungle of information and data processing solutions.
- It is hard to find information in this jungle.
- It is hard or impossible to formulate queries against multiple sources.
- No central instance to handle structured content
- No central solution for recovery/synchronization
- No central solution for backup/distribution
- No central solution for versioning/encryption
- Search restricted to keyword search

More Plugins

Query Dispatcher
 central query handling
Full Text Search
 keyword search/intelligent folders
XQuery
 desktop XQuery integration & join processing
History
 versioning on selected resources
Format Converter
 provide conversion routines
DataSpaces (Future Work)
 Physical Data independence, P2P
RSS/ATOM (Future Work)
 RSS/Desktop integration

Related Work on PIM

VLDB 2003 panel
Lowell Report 2003
Both keynotes at SIGMOD 2005
Misc papers at CIDR&SIGMOD

Other systems

SEMEX [Halevy et.al.]
Bloomba [Stata et.al.]
MyLifeBits [Bell et.al.]
Stuff I've seen [Dumais et.al.]
...

Operating systems

MAC OS X Tiger: full text search
Win FS (to appear after Longhorn)

iMeMex.org