



International Chemical Information Conference Nice, France, 20 October 2008

Professional IR - Cultivating The Corpus




Francisco Eduardo De Sousa Webber

Matrixware and the Database of Ideas




- Founded 2005 in Vienna
- 80 Employees
- >15 Academic Partners Worldwide
- 5 Product Lines:
 - Alexandria Repository
 - Leonardo Framework & Workbench
 - Machine Translation
 - Semantic Supercomputing
 - Professional Services
- Targeting: global industry top1000, Government, IPR Service Providers

Current „Consumer“ IR Methods

- 
- A horizontal bar composed of several rectangular segments in various shades of blue, ranging from light to dark, positioned below the title.
- **Simplicity → Google – like Interface**
 - **Hide the Data away from the user → „why“ for a hit is not important as long as the hit is good**
 - **Recall is not relevant → Only Top ranks are of interest**
 - **Dynamic precision → Always find something**
 - **Hurry! → Always respond within a second**
 - **All the users want the same → not much about the user is known – An average user is assumed**

Needs of Professional IR

- 
- A horizontal bar composed of several rectangular segments in various shades of blue and grey, spanning the width of the slide.
- **Efficiency over Simplicity**
 - **Transparency on Methods and Data**
 - **High Recall, every thing that counts count**
 - **High Precision, Time is Money...**
 - **Quality over Speed**
 - **Every User is special**

There is no single IR formula for every case

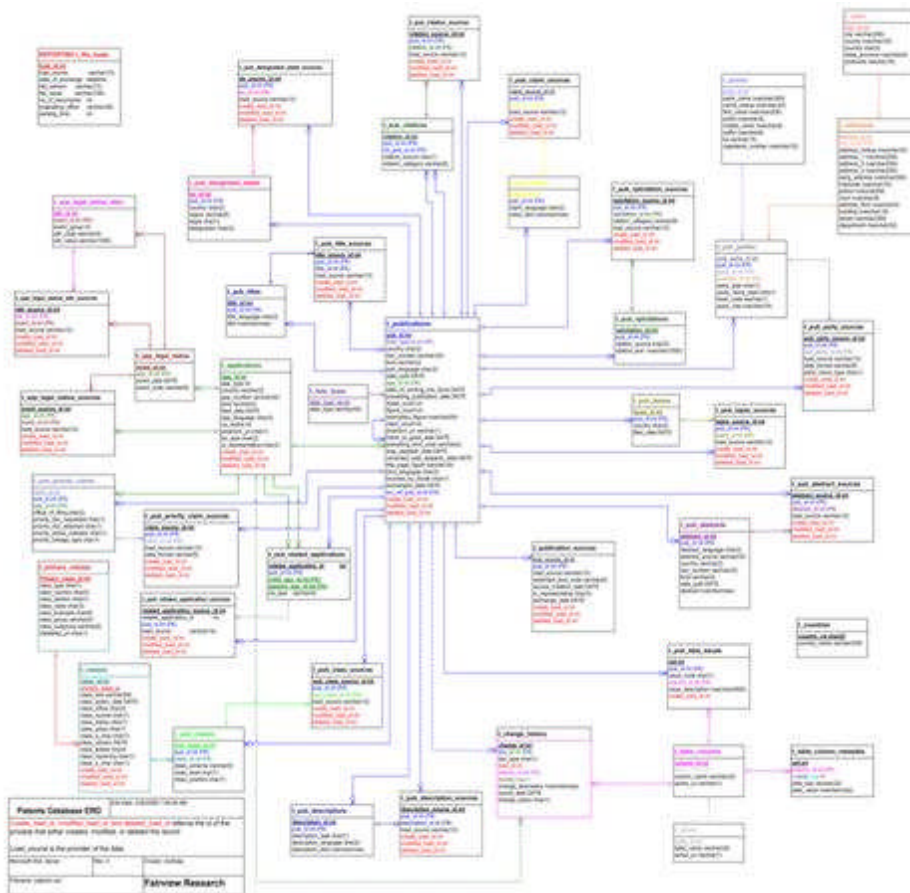
- **FEEDBACK**
- **PIPELINES**
- **COMPOUND ALGORITHMS**
- **depending on the problem to solve:**

Boolean Retrieval (Yes we do need it!)
Probabilistic Retrieval
Language Modeling
Term Extraction
Entity Extraction
Classification
Clustering
Dimensionality Reduction
Summarization
Translation
ETC...

A horizontal bar composed of several rectangular segments in various shades of blue, ranging from dark to light, positioned across the top of the page.

**“OUR FIRST TASK
IS THE SEPARATION OF THE DATA ISSUE
FROM THE APPLICATION ISSUE“**

ALEXANDRIA - PRODUCT OVERVIEW



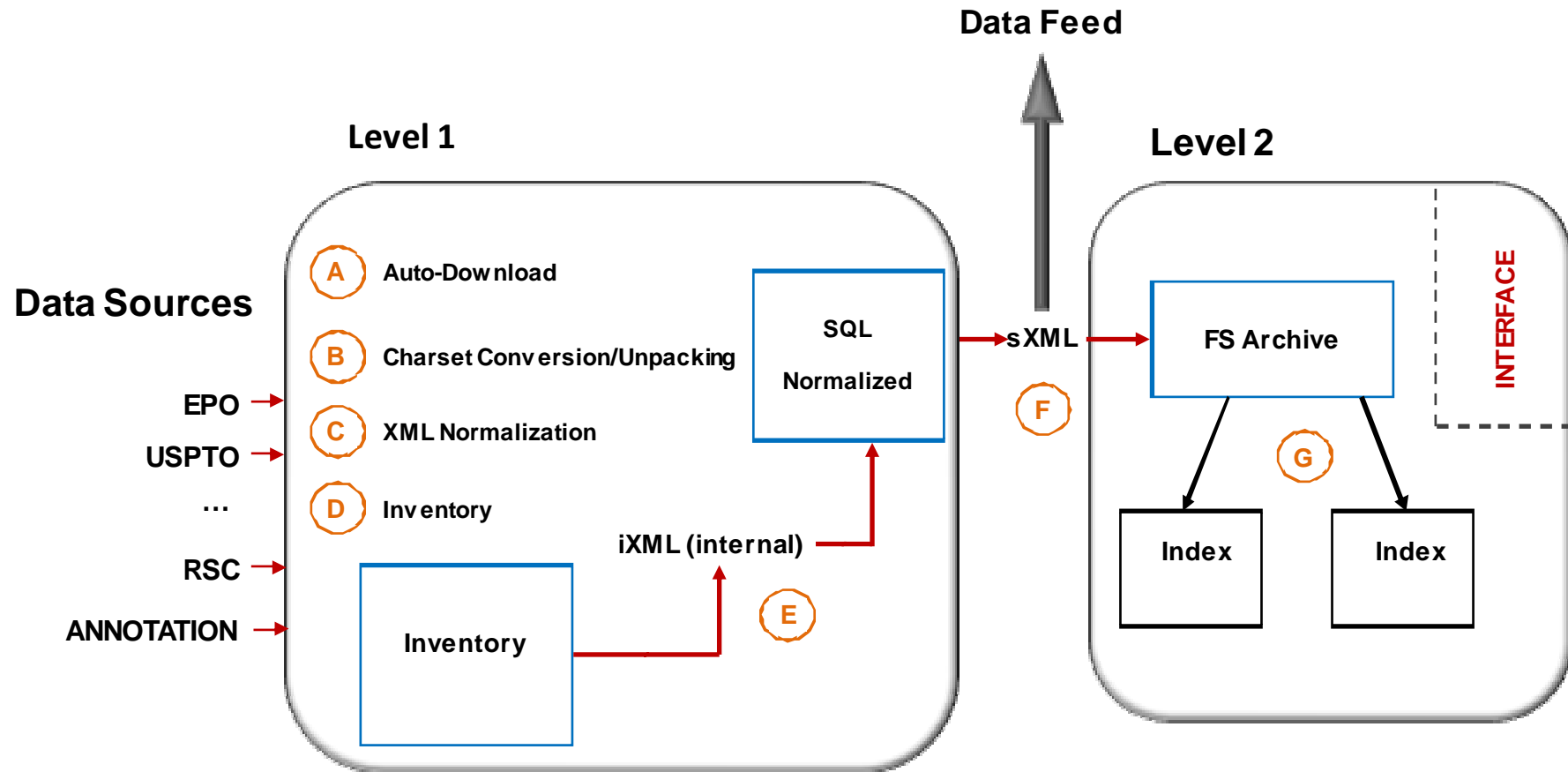
Patent repository

- Goal: high quality SciTec, Business and Intellectual Property documentation archive

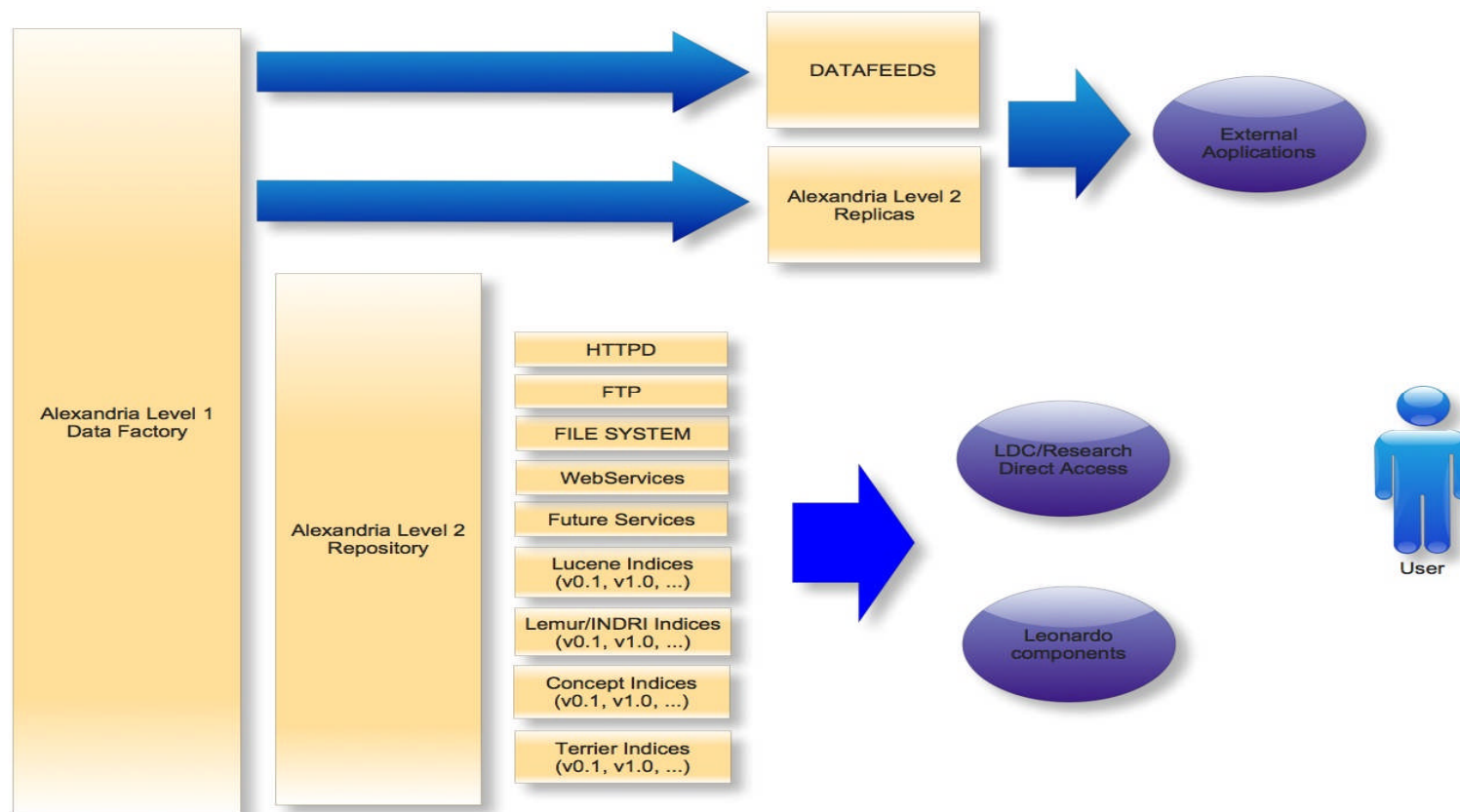
Matrixware & Fairview Research

- Acquiring, analyzing, normalizing, integrating, processing, loading and exporting of patent data
- Advanced Quality Assurance tools
- Provide data access for Leonardo Connector

THE ALEXANDRIA DATA FACTORY



THE ALEXANDRIA ARCHITECTURE



Introduction to Alexandria: Consolidation and Quality

A decorative horizontal bar consisting of a series of colored segments in various shades of blue and grey, spanning the width of the slide.

■ Data Preparation

- > from raw data into a single XML format and harmonised field names
- > corrections (OCR)

■ Standardise and normalise representation of data

Advanced QA

Gaps analysis in coverage and field content:

missing patent number, country codes missing in citations or incorrect, missing IPCs, misclassification, missing inventor or assignee names

Number standardisation, assignee standardisation

■ Revision History

All info captured and stored allowing on-demand revision history and tracking of changes for classification, re-assignments, legal status etc

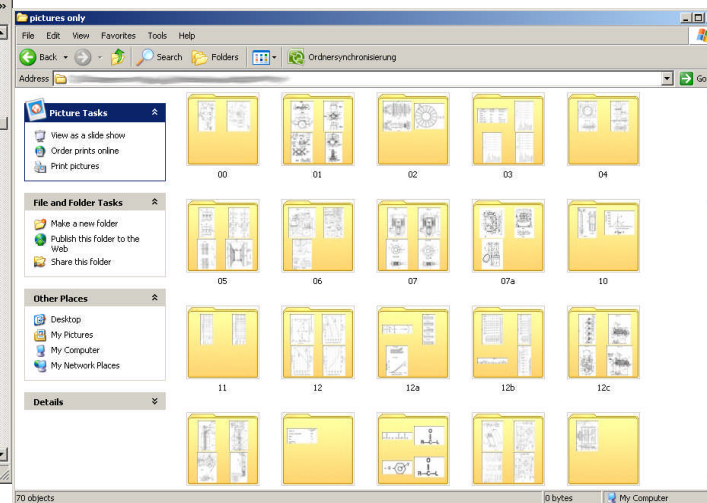
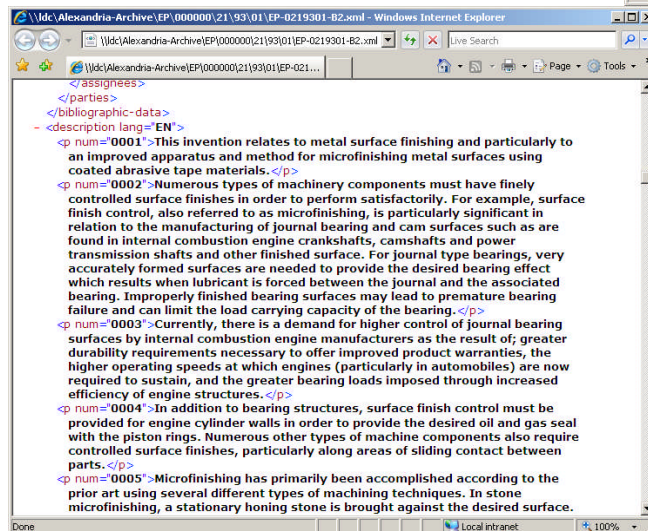
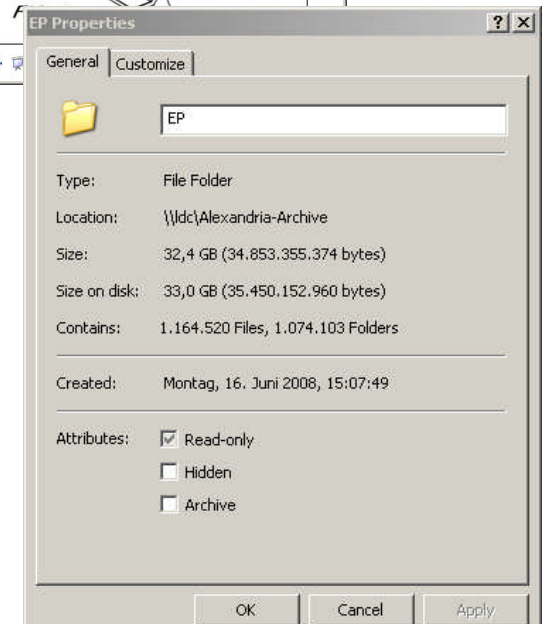
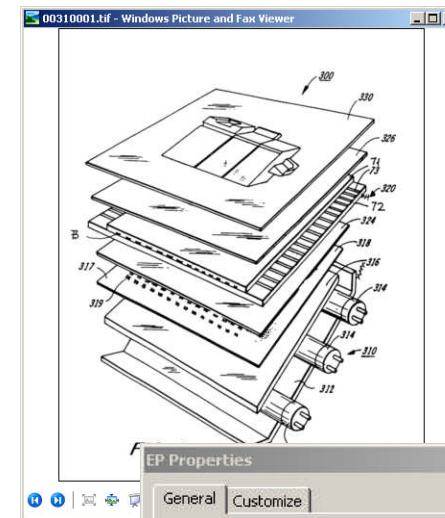
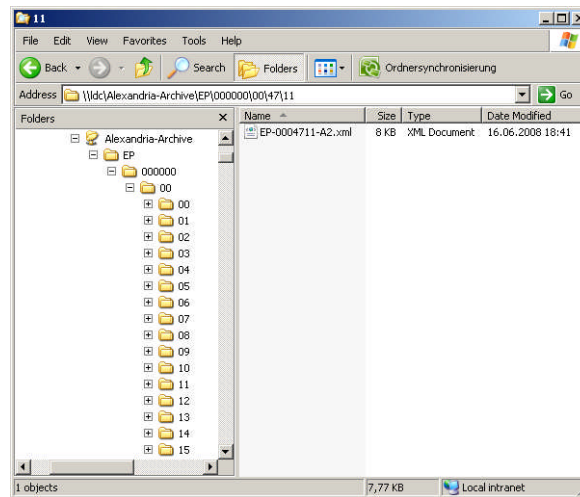
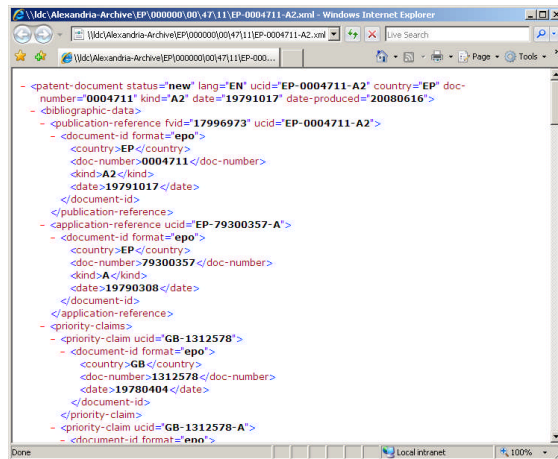
■ DOCDB merged with fulltext and national bib data and revision history

Introduction to Alexandria: Cultivating the Corpus

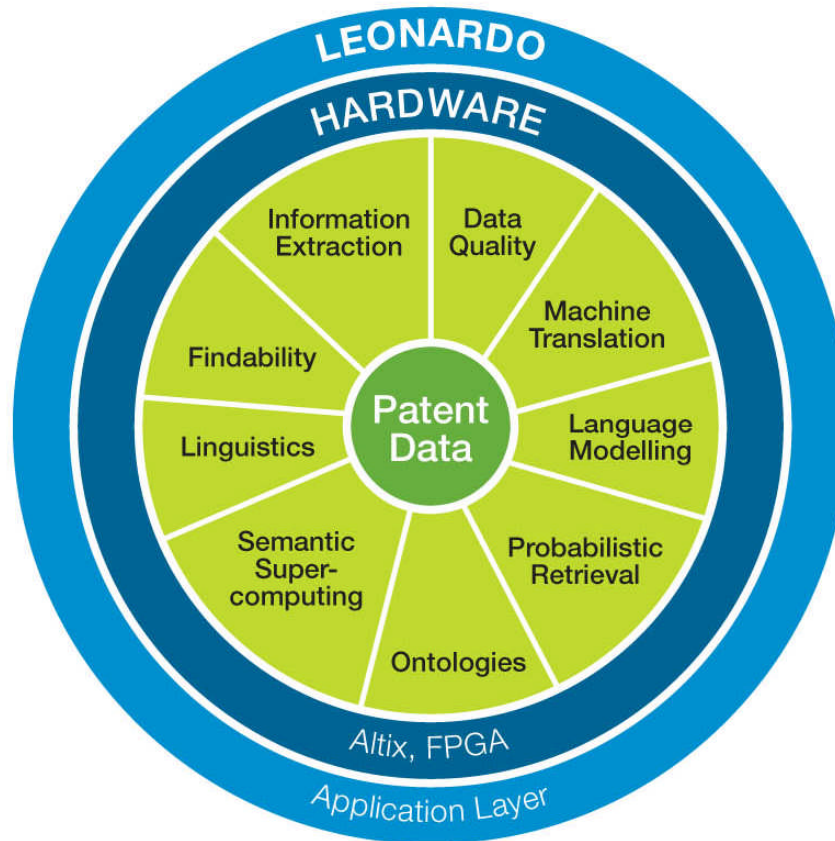
A decorative horizontal bar composed of several rectangular segments in various shades of blue, ranging from light to dark.

- **Ongoing data enrichment through results of various NLP approaches**
 - > capture and store data, metadata, semantic annotations, tags...
 - > ontologies, thesauri, linguistic corpora
- **Represent the data in more “intelligent” ways and increase retrievability with state of the art IR methods**
- **Keep the soil fertile for ongoing enrichment - Accessibility**

ALEXANDRIA - USER INTERFACE CONCEPTS

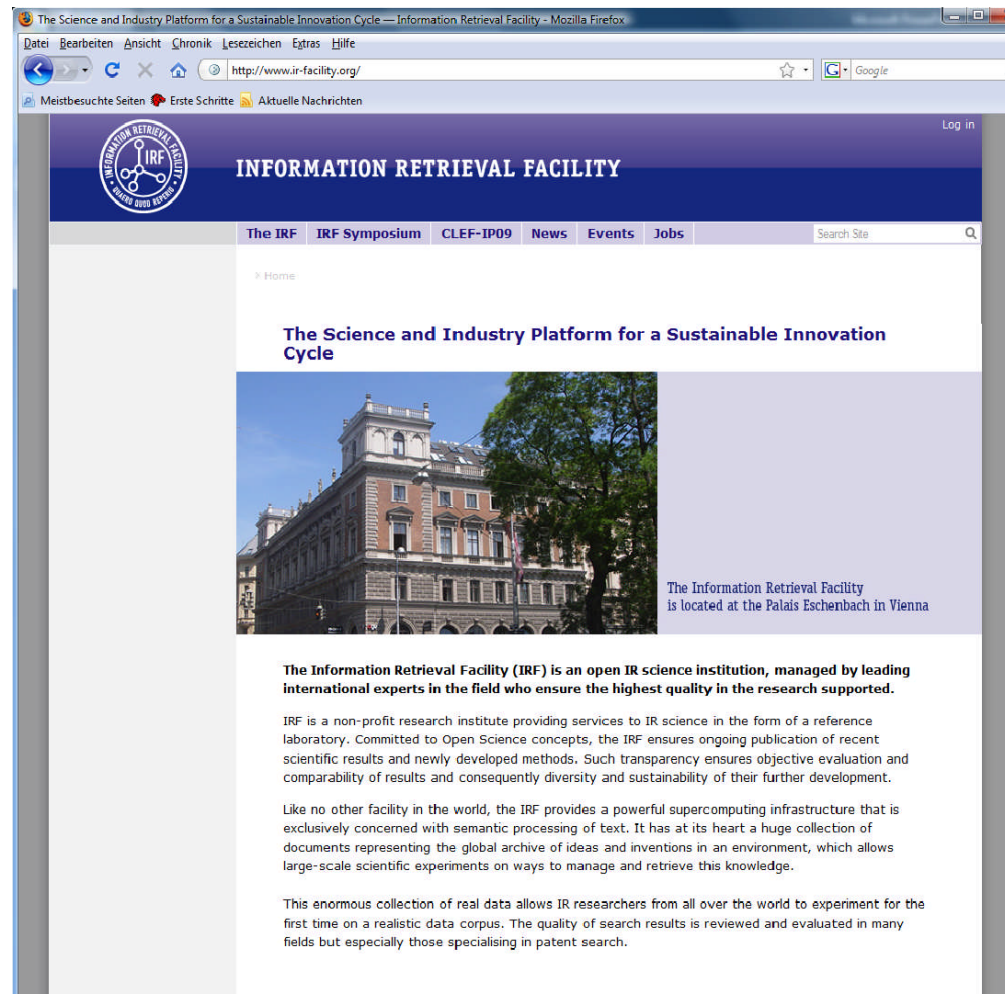


IR LANDSCAPE AND CURRENT PROJECTS



- LEONARDO** Application Layer
- HARDWARE** Altix, FPGA
- ALEXANDRIA** Patent Data Repository
- Scientific Projects

The Information Retrieval Facility - IRF



The Science and Industry Platform for a Sustainable Innovation Cycle — Information Retrieval Facility - Mozilla Firefox

http://www.ir-facility.org/

Meistbesuchte Seiten Erste Schritte Aktuelle Nachrichten

Log in

INFORMATION RETRIEVAL FACILITY

The IRF IRF Symposium CLEF-IP09 News Events Jobs Search Site

Home

The Science and Industry Platform for a Sustainable Innovation Cycle

The Information Retrieval Facility is located at the Palais Eschenbach in Vienna

The Information Retrieval Facility (IRF) is an open IR science institution, managed by leading international experts in the field who ensure the highest quality in the research supported.

IRF is a non-profit research institute providing services to IR science in the form of a reference laboratory. Committed to Open Science concepts, the IRF ensures ongoing publication of recent scientific results and newly developed methods. Such transparency ensures objective evaluation and comparability of results and consequently diversity and sustainability of their further development.

Like no other facility in the world, the IRF provides a powerful supercomputing infrastructure that is exclusively concerned with semantic processing of text. It has at its heart a huge collection of documents representing the global archive of ideas and inventions in an environment, which allows large-scale scientific experiments on ways to manage and retrieve this knowledge.

This enormous collection of real data allows IR researchers from all over the world to experiment for the first time on a realistic data corpus. The quality of search results is reviewed and evaluated in many fields but especially those specialising in patent search.

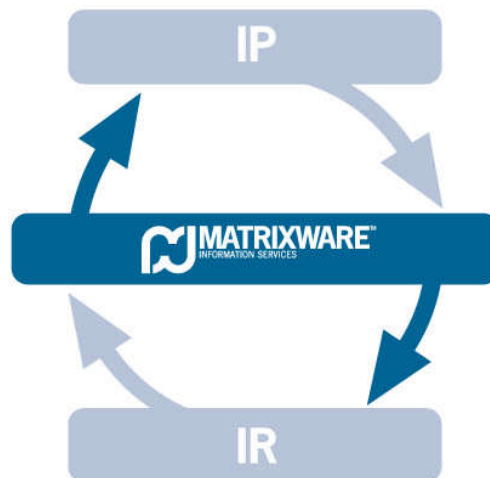
THE IRF SYMPOSIUM 2007

SCIENCE MEETS INDUSTRY
IRFS 2007 Vienna
Information Retrieval Facility
SYMPOSIUM



For the first time, IR-scientists and IP-professionals were engaged in a dedicated dialogue with a focus on improving productivity in their established workflows by applying computational models.

THE INNOVATION CYCLE



- **The IP professionals**, who need technological support in reducing their work load and in raising the overall level of quality to ensure competitiveness.
- **The information scientists**, who can continuously provide conceptual solutions to IP pain-points and who need “real world cases” to push their research limits higher.
- **Matrixware**, an essential component in the innovation cycle, acts as a mediator and facilitator. Matrixware enables a fruitful and moderated dialogue, guaranteeing maximum relevancy for both science and business.

THE IR LANDSCAPE AND CURRENT PROJECTS



PROBABILISTIC RETRIEVAL
LOGRET

INFORMATION STORAGE
MONET DB/X100

LANGUAGE MODELLING
PRIOR ART SEARCH

INFORMATION EXTRACTION
SAM-2

LINGUISTICS
TM4IP

VISUALISATION
VISIPAT

ACCESSABILITY IN IR
DOCUMENT FINDABILITY

SEMANTIC SUPERCOMPUTING
FPGA

TOOLKIT ADVANCES 1
TERRIER

CLASSIFICATION/CLUSTERING

CONCEPT-BASED IR
DIRECTED CRAWL

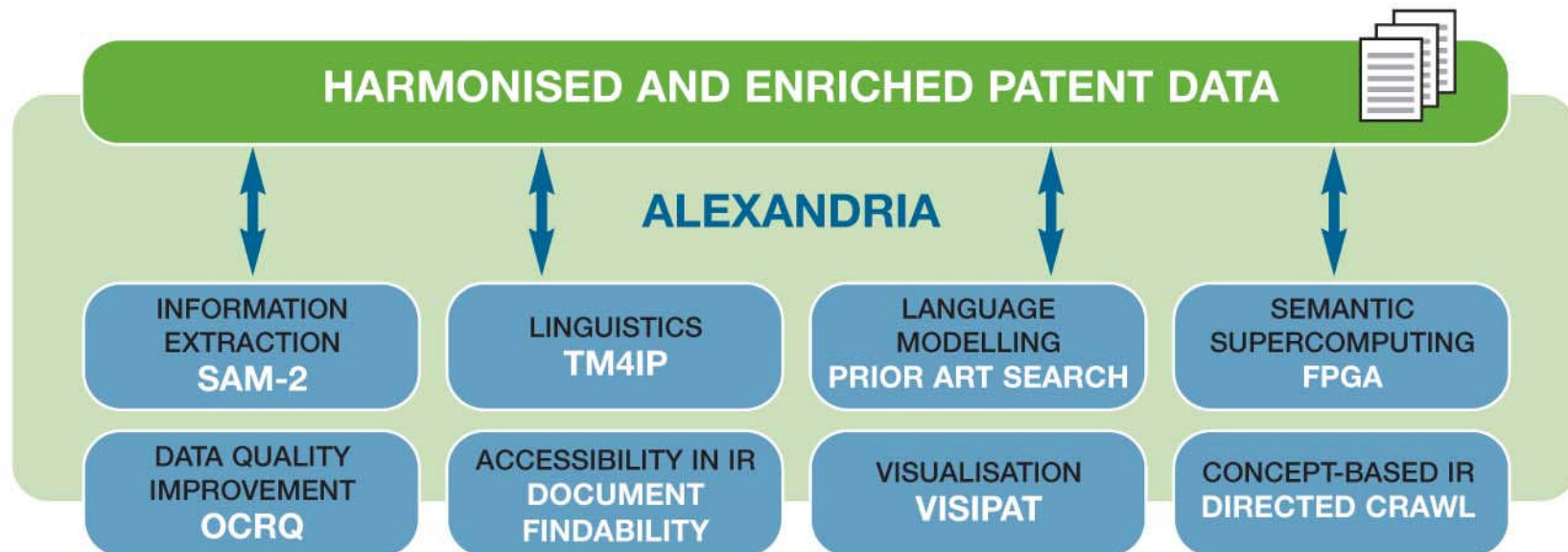
TOOLKIT ADVANCES 2
LEMONA, LEMUR

TEST COLLECTIONS
NIST/CLEF

DATA QUALITY
OCRQ

MACHINE TRANSLATION

THE IR LANDSCAPE AND CURRENT PROJECTS



ALEXANDRIA - Quality Assurance Tools

Alexandria PATENTS Database QA

Help Search

Search Results Search Test Plans

Source XML

```
<?xml version="1.0" encoding="UTF-8"?>
<exch:exchange-document xmlns:exch="http://www.epo.org/exch"
<exch:bibliographic-data>
<exch:publication-reference data-format="A2"
<document-id lang="en">
<country>EP</country>
<doc-number>0700000</doc-number>
<kind>A2</kind>
<date>19960306</date>
</document-id>
</exch:publication-reference>
<exch:publication-reference data-format="A2"
<document-id lang="en">
<doc-number>EP0700000</doc-number>
</document-id>
</exch:publication-reference>
<exch:classification-ipc>
<edition>6</edition>
<main-classification> 6G 06F 12/02 A</main-classification>
</exch:classification-ipc>
<exch:classifications-ipc>
<classification-ipc sequence="1">
<text>G06F 15/00 20060101CFI20
</classification-ipc>
<classification-ipc sequence="2">
<text>G06F 15/00 20060101AFI20
</classification-ipc>
<classification-ipc sequence="3">
<text>G06F 9/44 20060101CLI20C
</classification-ipc>
<classification-ipc sequence="4">
<text>G06F 9/44 20060101ALI20C
</classification-ipc>
<classification-ipc sequence="5">
<text>G06F 9/46 20060101CI I200
</classification-ipc>
<classification-ipc sequence="6">
<text>G06F 9/46 20060101AI I200
</classification-ipc>
<classification-ipc sequence="7">
```

Search Results Previous Next

##	Publication ...	Publica...	Application ...	Filing D...		
1	20...	EP0700000A2	06 Mar...	EP95480098A	24 Jul ...	Check Sources
2	71...	EP0700000A3	no Date	EP95480098A	24 Jul ...	Check Sources
3	20...	EP0700000A3	02 Jan...	EP95480098A	24 Jul ...	Check Sources
4	20...	EP0700000B1	04 Oct...	EP95480098A	24 Jul ...	Check Sources

Extract XML

```
<?xml version="1.0" encoding="UTF-8"?>
<patent-document status="new" lang="en">
<bibliographic-data>
<publication-reference fvid="2007000000">
<document-id format="epo">
<country>EP</country>
<doc-number>0700000</doc-number>
<kind>A2</kind>
<date>19960306</date>
<lang>EN</lang>
</document-id>
</publication-reference>
<application-reference ucid="EP07000000">
<document-id format="epo">
<country>EP</country>
<doc-number>95480098</doc-number>
<kind>A</kind>
<date>19950724</date>
<lang>EN</lang>
</document-id>
</application-reference>
<legal-status>
<legal-status-events>
<legal-event country="EP" code="A2"
<attributes>
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
</attributes>
</legal-event>
<legal-event country="EP" code="A2"
<attributes>
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
</attributes>
</legal-event>
<legal-event country="EP" code="A2"
<attributes>
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
<attribute attr-group="4" value="19960306" />
<attribute attr-group="4" value="19950724" />
</attributes>
</legal-event>
```

Sources Results

EP-0700000-A2

Exchan...	Produc...	Product Name	Source File	
28 Feb ...	14.7	DOCDB	DOCDB-2008...	View
06 Mar ...	14.1	EBD	/patpool/Prim...	View
06 Mar ...	14.4	EPA Abstracts	/patpool/Prim...	View
15 Mar ...	14.1	EBD	/patpool/Prim...	View

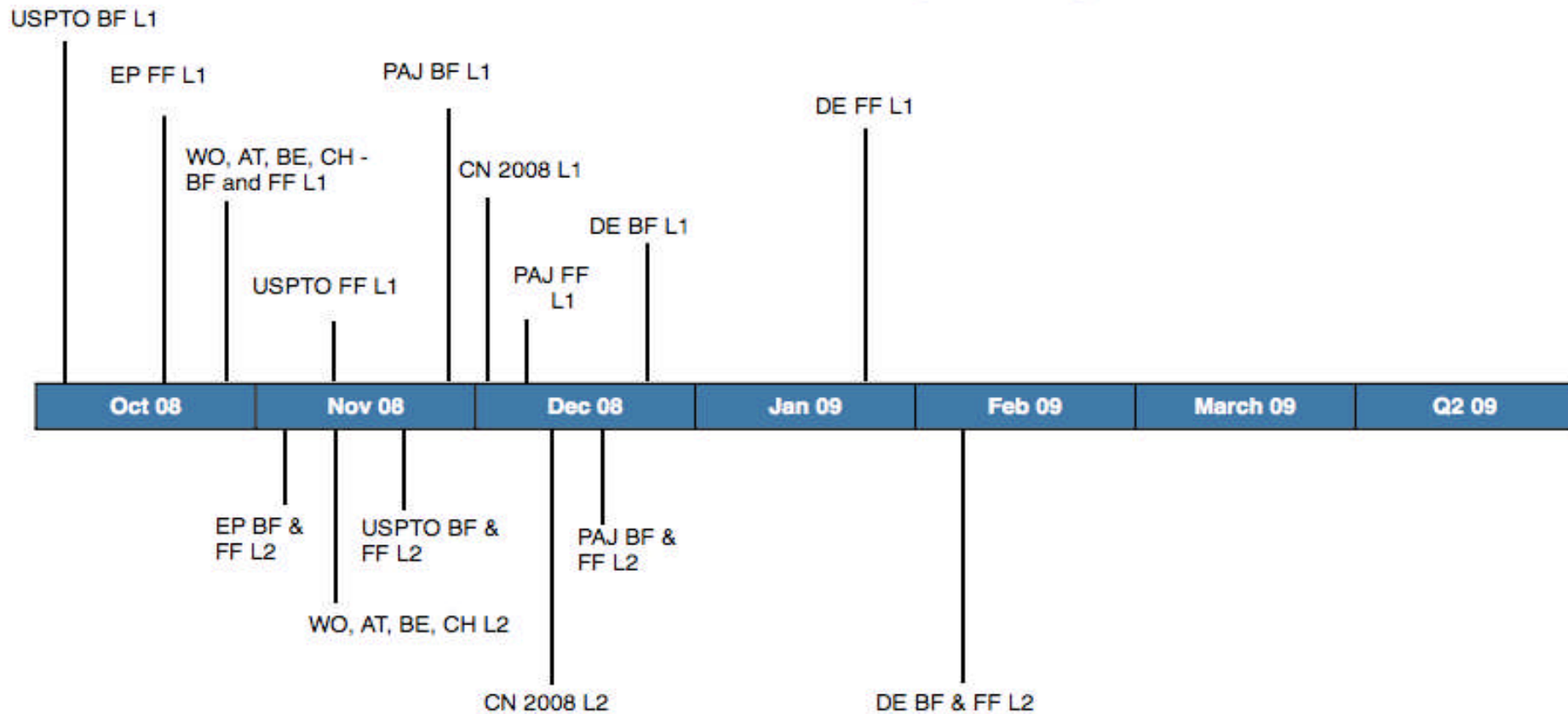
ALEXANDRIA - ACCESS

A decorative horizontal bar consisting of a series of colored segments in shades of blue and grey, spanning the width of the slide.

Alexandria data sets are/will be available as:

- **Data Subsets**— Collections of XML files selected from the XML archive. These are created by individual users.
- **Data Extracts**— XML files prepared from the Level 1 database. These are available on demand.
- **Data Feeds**— XML files provided periodically by subscription. These are available Q1 2009
- **Alternative Formats**— Data in a configuration other than the normalised XML. These can be prepared on demand.

ALEXANDRIA - Timeline to full release



Lucene Index

INDRI/Lemur Index
Concept-Based

Test Plan
Completed

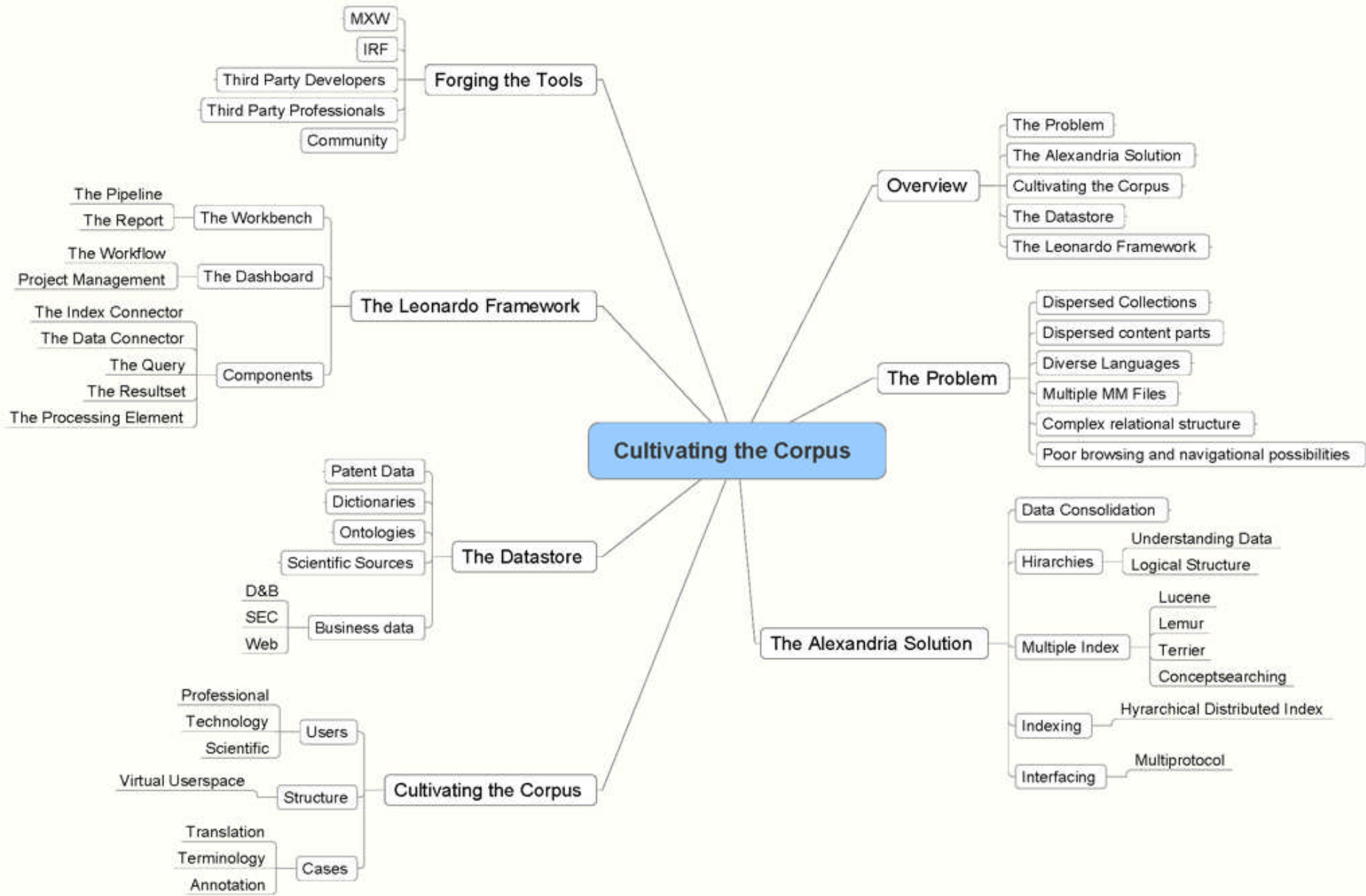
RSC Journals In
the Repository

CN Data Feed

Integration of Annotation Tools TBD

Annotation Store

- Annotations in patent documents
 - Numeric values: length, mass, time, temperature
 - Named entities
 - Part-of-speech
- Small example
 - 10^6 documents x 10^3 annotations per document
 - **10^9 annotations**
- Real life
 - 10^8 documents x $10^?$ annotations per document
 - Part-of-speech: more annotations than terms in document!
 - Semantic connections between annotations?
- How to store and how to query?

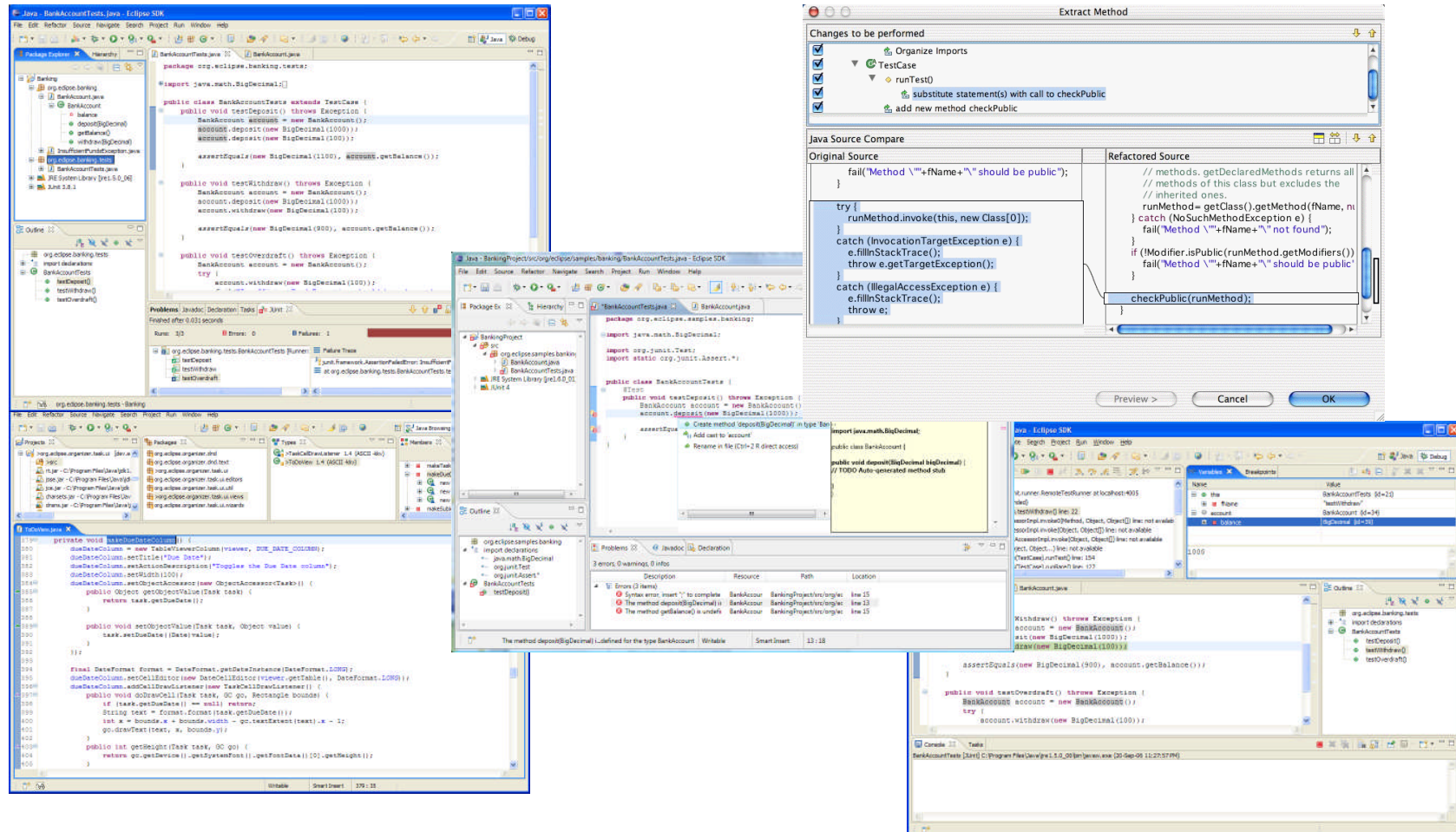


THE LEONARDO FRAMEWORK

A horizontal bar composed of several rectangular segments in various shades of blue, ranging from light to dark, positioned below the section header.

- Workbench for the IP professional to access enhanced data and take advantage of sophisticated tools for searching and retrieving patent data
- Technology transfer environment between IP and IR
- Creation of end-user search interfaces based on preset configuration by IP professionals
- Development framework for IP and 3rd parties to create their own solutions including their own knowledge base (ontologies, data sources etc)
- Open, transparent, extensible framework for the continuous introduction of new data and information retrieval technologies

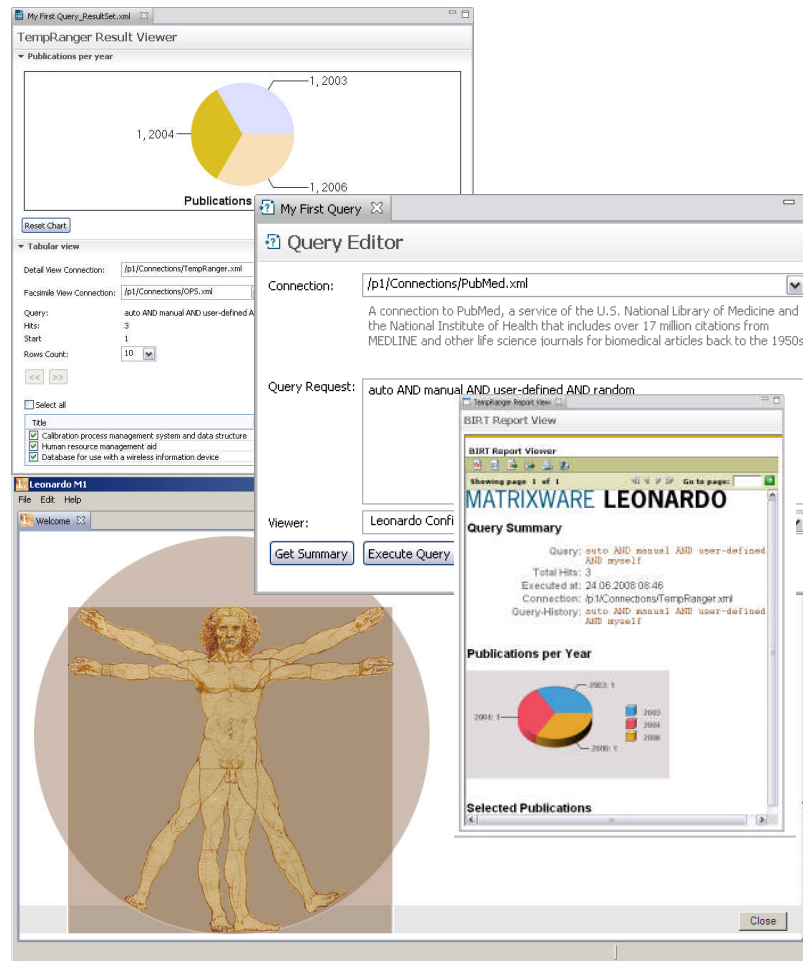
THE LEONARDO FRAMEWORK I



The image displays several screenshots from the Eclipse IDE illustrating the Leonardo Framework's capabilities:

- Top Left:** A Java class file showing a test case for a `BankAccount` class. The code includes imports for `org.junit.Test` and `org.junit.Assert`, and methods for `testDeposit`, `testWithdraw`, and `testOverdraft`.
- Top Right:** The "Extract Method" dialog box. It shows "Changes to be performed" with options like "Organize Imports", "Test Case", "runTest()", "substitute statement(s) with call to checkPublic", and "add new method checkPublic". Below, it compares the "Original Source" (a try-catch block) with the "Refactored Source" (a call to a new `checkPublic` method).
- Middle:** A screenshot of the IDE showing a Java class with a `deposit` method. A yellow tooltip indicates a "Create method depositBigDecimal" refactoring action.
- Bottom Left:** A screenshot of a Java class with a `testDeposit` method. A yellow tooltip indicates a "Rename in file" refactoring action.
- Bottom Center:** A screenshot of the IDE showing a Java class with a `testDeposit` method. A yellow tooltip indicates a "Rename in file" refactoring action.
- Bottom Right:** A screenshot of the IDE showing a Java class with a `testDeposit` method. A yellow tooltip indicates a "Rename in file" refactoring action.

THE LEONARDO FRAMEWORK II



The screenshot displays the Leonardo Framework II interface with several overlapping windows:

- TempRanger Result Viewer:** Shows a pie chart titled "Publications per year" with three segments representing the years 2003, 2004, and 2006, each with a value of 1.
- Query Editor:**
 - Connection: /p1/Connections/PubMed.xml
 - Query Request: auto AND manual AND user-defined AND random
 - Buttons: Get Summary, Execute Query
- BIRT Report View:**
 - Header: MATRIXWARE LEONARDO
 - Section: Query Summary
 - Text: Query: auto AND manual AND user-defined AND random; Total Hits: 3; Executed at: 24.06.2008 08:46
 - Section: Publications per Year
 - Section: Selected Publications
- Leonardo MI:** The main application window, featuring a large image of Leonardo da Vinci's Vitruvian Man.

Thank You!



International Chemical Information Conference Nice, France, 20 October 2008

Professional IR - Cultivating The Corpus

Francisco Eduardo De Sousa Webber

THE IRF SYMPOSIUM 2008

IRFS2008

**INFORMATION RETRIEVAL
FACILITY SYMPOSIUM**

**5-8 NOVEMBER 2008
VIENNA, AUSTRIA**



5th November 2008, 2pm - 5pm
"IP for IR" and "IR for IP" Tutorial

6th & 7th November 2008
Information Retrieval Facility Symposium

8th November 2008
Outreach Exhibition Day

Location
Austria Trend Hotel Savoyen, Vienna, Austria

www.ir-facility.org/irfs2008



ALEXANDRIA - Sample Data Extracts



- To request a data extract, send an e-mail to Alexandria.Support@Matrixware.com describing:
- Your project.
- The goals you want to achieve.
- The data extract content (e.g., patent authorities, language, classes).
- Format and medium.
- The expected size of the data set.
- A deadline indicating when you need the data.
- Any requirements for updated data sets.
- Alexandria team experts will contact to discuss your requirements.