

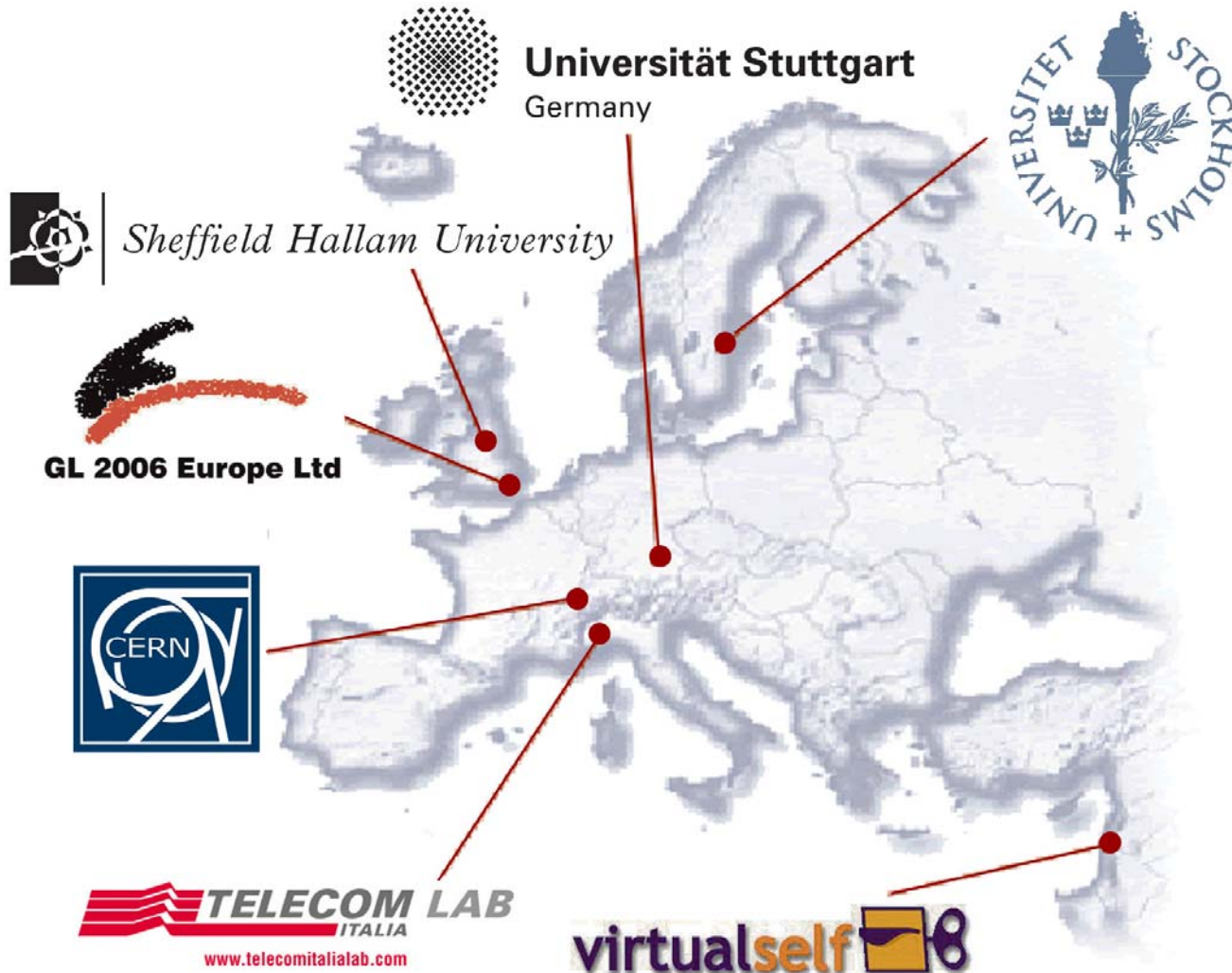


Information Retrieval on the Grid? Results and suggestions from Project GRACE

Werner Stephan
Stuttgart University Library

IATUL 2005

Project Participants



What is Information Retrieval?

Retrieval of unstructured, textual information typically stored in various document formats

- Unstructured: does not include, for example, metadata in Dublin Core Simple
- Textual: does not include numeric data (like produced by experimental instrumentation in High Energy Physics)
- Document formats: does not include database federation

Information Retrieval

- Typical approach: indexing
 - Pre-processing text applying - at least to some extent - natural language processing
 - Resulting index stored in a format optimized for rapid querying
- Exotic approaches:
 - Post-retrieval processing (typically in meta-search)
 - Concept indexing (similar to manual keyword annotation, only automatic)
- Information retrieval is “text crunching”



What is an ontology?

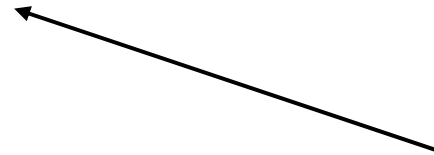
A list, sometimes hierarchical, of agreed upon subject headings

Example:

Title: Travel in Britain
Subject: **Tourism--UK**

Title: Sightseeing around England
Subject: **Tourism--UK**

A uniform
descriptor for a
single concept



What is concept based indexing?

- Concept based indexing is similar to manual keyword annotation
- Automatic process on paragraph level
- Uses an existing ontology

Is Grid a Solution?

- “Text crunching” is a heavy computational task
- Resulting indices of enormous size
- What is grid good at?
 - Batch pre-processing (e.g., text indexing)
 - Massive data storage (e.g., indices)
- Grid - an ideal solution for pre-processing (indexing) documents for information retrieval



Limitations of the Current Grids

- Interaction with end-user
- Complex installation and maintenance
- Complex certification
- Grid job failure rate
- Weak monitoring
- Primitive distributed data management

Grid in the GRACE Context

- Infrastructure for Enabling Grids for E-Science in Europe (EGEE) in Italy
 - GILDA (Grid INFN Laboratory for Dissemination Activities)
 - Run by Istituto Nazionale Fisica Nucleare (INFN)
 - Based on the Large Hadron Collider Computing Grid LCG-2
- Grace testbed INFN-Grid 2.0.0
 - 2 nodes
 - Turin (5 CPUs, 300 GB Storage Space)
 - Milan (4 CPUs, 250 GB Storage Space)

Content Sources on Grid

- Scarce content sources on Grid
- Abundant content sources on WWW
- Web-based content sources distributed
- Meta-search is required to use web-based content sources
- In meta-search scenario document texts not available for pre-processing (not accessible without submitting a query)

How to Acquire Content Sources?

- Harvesting instead of crawling
- Ontology concepts used as queries submitted to multiple content sources
- Downloaded document processed
- Indices stored on grid
- Queries repeatedly submitted to keep information updated

GRACE

- High Energy Physics Keyword Index (HEP)
 - Not a full-fledged ontology
 - Useful terminology
- Targeted content sources:
 - CERN Document Server (CDS)
 - Google Scholar

Pre-processing Content

- Keyword indexing (like regular search engines)
- Concept indexing (using HEP terminology)
- Categorization
 - Extracting additional lexical patterns
 - Clustering documents accordingly

Concept based indexing in GRACE -- Knowledge Domains

The screenshot displays the GRACE web interface. At the top left is the GRACE logo, which features a globe with the word "GRACE" overlaid. To the right of the logo are two yellow buttons labeled "Language" and "Utilities", each with a double arrow icon. Below the logo is a navigation bar with four buttons: "search steps", "Search", "My Profile", and "My Collection". The "search steps" button is highlighted in yellow and contains three numbered steps: 1. selected content source, 2. define search criteria, and 3. launch search. Below the "search steps" button is a "next" button with a double arrow icon. To the right of the "search steps" button is a "Knowledge Domains" section. This section has a tree structure with a "General" folder and a "Physics" folder. The "Physics" folder is expanded, showing two sub-items: "Google.com" and "CERN Document Server". An arrow points from the text on the right to the "Physics" folder.

The user can choose from a variety of subjects and content sources

User enters a search term

Language

GRACE

Search steps

selected content source

define search criteria

launch search

next

previous

Search

My Profile

choose terms from a classification

Quick-Search

xenon

Reset

Search results are indexed according to the HEP (High Energy Physics) Keyword Index

Language

Utilities

GRACE

HEP Keywords

Sort by Name

Sort by No of Documents

black hole	[98]
mass	[82]
Delta	[61]
horizon	[58]
angular momentum	[56]
spectra	[52]
Schwarzschild	[52]
OMEGA	[50]
approximation	[50]
particle	[49]
spin	[43]
matter	[43]

Listing 100 out of 100 Documents

1

Shapes and Positions of Black Holes
Takahashi, R
Can we determine a spin parameter for a black hole? In order to answer this question, we must first determine the shadow cast by a rotating black hole. We have found parameters for the shadow size and a shape of a black hole named a shadow axis. For a rotation axis of a black hole shadow axis is finite. An inclination angle between a shadow axis and a center of a black hole in a plane is finite. Source: CERN Document Server
Url: <http://doc.cern.ch/Archive>

Automatic Categories in addition to concept based indexing

Context indices	Suche	Mein Profil	Meine Sammlungen
<ul style="list-style-type: none"> --A0 [3] --accelerator [1] --acceptance [1] --angular distribution [1] --angular resolution [1] --any-dimensional [1] --associated production [1] --ATLAS [1] --B [8] --baryogenesis [1] --baryon [1] 	<p>Listing 101 out of 101 Document(s) found</p> <p>1 Little Higgs Phenomenology Logan, H E Recently a new class of models has emerged, the Standard Model Higgs boson acquires mass radiatively only through required to give the Higgs a mass. This one-loop. These models contain new vector bosons, the little Higgs models, focusing on colliders. Url: http://doc.cern.ch/archive/electronic/hep-ph/0603087 Source: CERN Document Server</p> <hr/> <p>2 Probing the Radion-Higgs mixing at hadron colliders Cheung, K; Kim, C S; Song, J In the Randall-Sundrum model, the radion features a sizable three-point vertex. We study the possibility of probing the radion-Higgs mixing at the CERN LHC in probing the radion-Higgs mixing. We also studied all the partial decay widths of the KK gravitons into a pair, with the branching ratio of order 0.1. Url: http://doc.cern.ch/archive/electronic/hep-ph/0603087 Source: CERN Document Server</p> <hr/> <p>3 Exclusive Double Diffractive Higgs Boson Production Petrov, V; Ryutin, R</p>		
<p>Major Topics</p> <ul style="list-style-type: none"> --no text [11] --Higgs bosons [7] --Higgs sector [6] --Higgs boson production [5] --production cross section [5] --Z boson [5] --Unclassified Documents [1] 			

GRACE also creates its own categories based on the content of the resulting documents

Concept based indexing plus categorization as a suitable retrieval concept for a grid based application

- Interoperability layer: single-point access to multiple content sources (similar to database federation)
- Unified presentation of information originating from multiple document formats and languages
- Structured and concise view of large amounts of information (similar to data warehouse)

Benefits

- Extends sharing of computational and storage resources to **knowledge resources**
- Allows members with limited resources to join forces in order to build powerful and hand-tailored information retrieval solutions
- Brings the collaboration to the level of “thinking together”



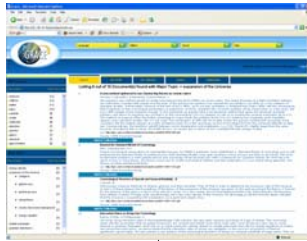
Related Initiatives

- **GGF GIR-WG**
 - Promotes information retrieval as issue for grid
 - Assumes that content sources are available on grid
- **IBM “Masala”**
 - IBM DB2 on grid
 - Similar: distributed data sources
 - Different: structured data (data warehousing)

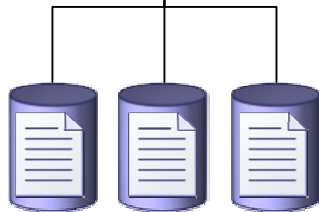


DB2 Masala Preview

Content Index & Categorization



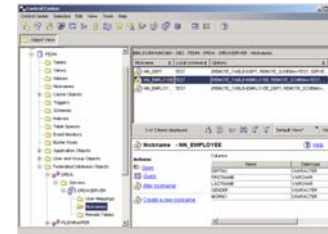
GRACE Search Engine



Distributed Content
Sources



Data Warehouse



Distributed Databases

Expectations from Future Grids

- Improved interactivity with end-user (high hopes in WSRF: grid services to be used just as Web Services are today)
- Simple installation and maintenance
- Simple certification
- Bullet-proof grid jobs
- Effective monitoring
- Highly efficient distributed data management (storage elements exposed as databases with uniform schema)

Future Research Directions

- Maximize use of ontologies for information retrieval on grid
 - Extend use of ontologies for information presentation
 - Improve harvesting through use of ontologies
- Tighten integration with grid infrastructure
 - Follow closely advances in standardization
 - Dependence on the grid infrastructure
 - Extend integration of GRACE with Replica Manager