

User Generated Contents

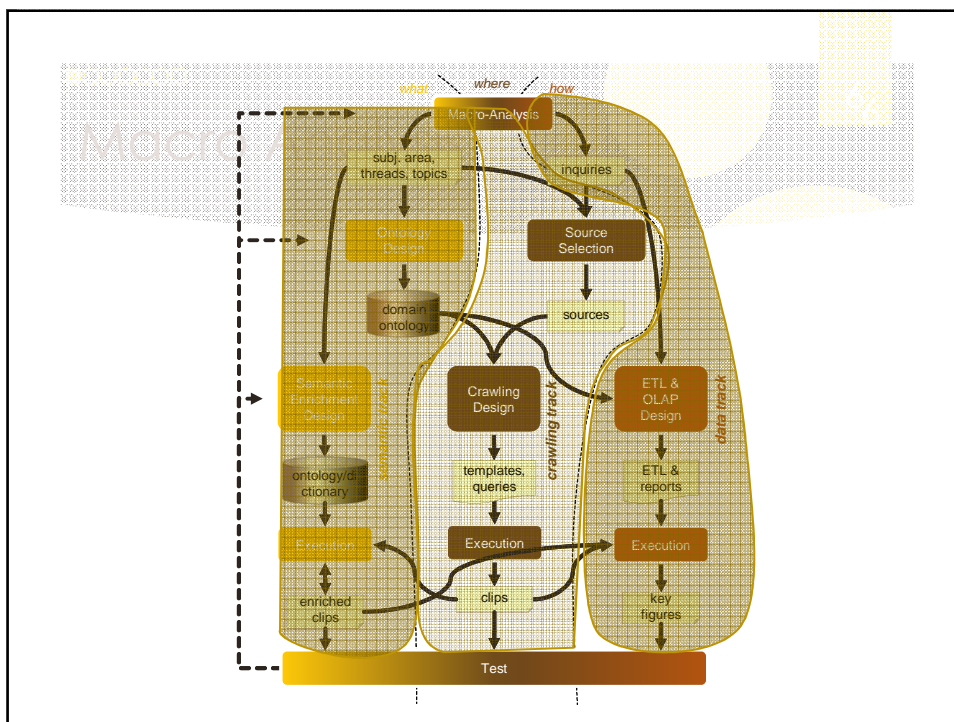
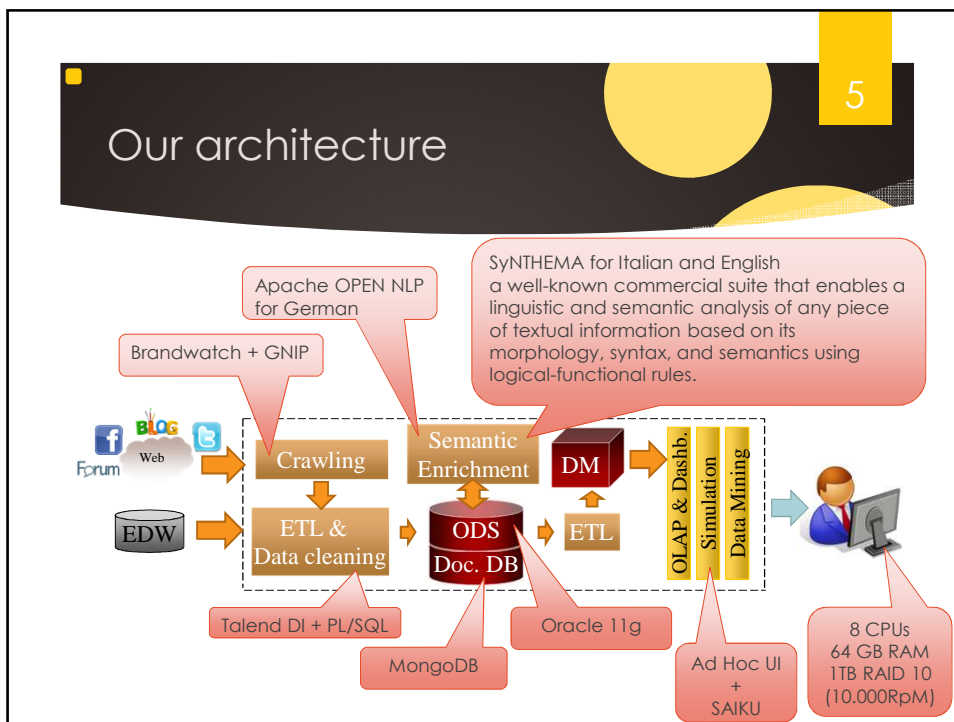
3

- ▶ **User-generated content (UGC)** refers to a variety of media content available in a range of modern communications technologies. UGC is often produced through open collaboration
- ▶ UGC is raising an increasing interest from decision makers
 - ▶ Give a **fresh** and **timely** perception of the **market mood**
 - ▶ Can be used to **deliver** important messages to potential **customers**
 - ▶ **Social events** are perceived by:
 - ▶ traditional information systems when they impact on the company processes (e.g. sales reduction).
 - ▶ SBI systems when they start happening

Social BI: a Definition

4

- ▶ **Social Business Intelligence (SBI)** is the discipline that applies DW and OLAP approaches to the analysis of user-generated content to let decision-makers improve their business based on the trends perceived from the environment.
- ▶ As in traditional BI the goal of **SBI** is to **enable** powerful and flexible **analysis** even for decision makers with limited technical skills.



7

Macro Analysis

Goal:

- ▶ **Project Scope**
 - ▶ domain of interest for the users
- ▶ **Inquiries**
 - ▶ captures an informative need of a user (What? How? Where?)
 - ▶ drive the definition of Themes and Topics
- ▶ **Activities:**
 - ▶ Interview/non technical meeting with users

8

Ontology Design

Goal:

- ▶ Describes the project scope.
- ▶ Key input for almost all process phases

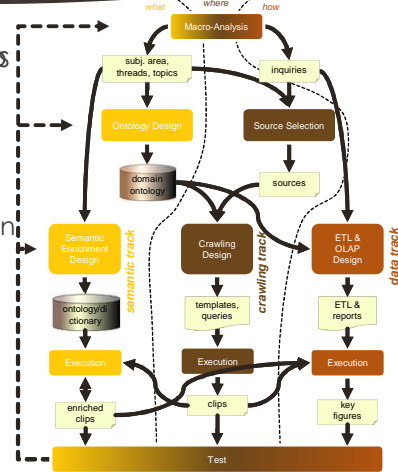
Activities:

- ▶ Detecting domain-relevant
 - ▶ topics
 - ▶ alias
 - ▶ themes
- ▶ and organizing them into a hierarchy

9

Source Selection

- ▶ **Activities:**
 - ▶ Identify **relevant** domains as possible to crawl
 - ▶ Backlinks analysis
 - ▶ Primary sources:
 - ▶ searching the Web using as keywords communication channels
 - ▶ themes
 - ▶ Generalist sources (online version of the major publications)
 - ▶ Choosing sites is a trade-off
 - ▶ **satisfying coverage** produce valuable information with high informative value
 - ▶ **optimizing the effort** for analyzing the retrieved clips
 - ▶ very focused on the project scope

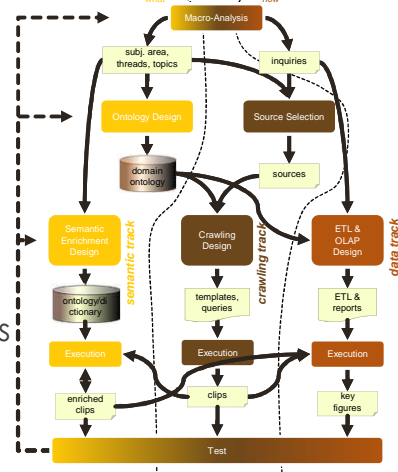


The flowchart illustrates the Source Selection process. It starts with 'Micro-Analysis' (what, where, how) leading to 'Inquiries' and 'sources'. 'Inquiries' leads to 'Ontology Design' (subj. area, threads, topics) and 'domain ontology'. 'Sources' leads to 'Crawling Design' (templates, queries) and 'Execution'. 'Ontology Design' leads to 'Semantic Enrichment Design' (ontology/d. ctionary) and 'Execution'. 'Crawling Design' leads to 'Execution'. 'Execution' leads to 'enriched clips' and 'clips'. 'enriched clips' leads to 'Test'. 'clips' leads to 'Test'. 'Test' leads to 'ETL & OLAP Design' (key figures) and 'ETL & reports'. 'ETL & OLAP Design' leads to 'ETL & reports'. 'ETL & reports' leads to 'Test'. The process is divided into three tracks: 'semantic track', 'crawling track', and 'data track'.

10

Crawling Design

- ▶ **Goal:**
 - ▶ Retrieving **in-topic** clips by filtering **off-topic** clips out
- ▶ **Activities:**
 - ▶ Template Design (clipping)
 - ▶ Query Design
 - ▶ Content Relevance Analysis
 - ▶ Sometimes is useful to release some constraints
 - ▶ filter clip at a later stage



The flowchart illustrates the Crawling Design process. It starts with 'Micro-Analysis' (what, where, how) leading to 'Inquiries' and 'sources'. 'Inquiries' leads to 'Ontology Design' (subj. area, threads, topics) and 'domain ontology'. 'Sources' leads to 'Crawling Design' (templates, queries) and 'Execution'. 'Ontology Design' leads to 'Semantic Enrichment Design' (ontology/d. ctionary) and 'Execution'. 'Crawling Design' leads to 'Execution'. 'Execution' leads to 'enriched clips' and 'clips'. 'enriched clips' leads to 'Test'. 'clips' leads to 'Test'. 'Test' leads to 'ETL & OLAP Design' (key figures) and 'ETL & reports'. 'ETL & OLAP Design' leads to 'ETL & reports'. 'ETL & reports' leads to 'Test'. The process is divided into three tracks: 'semantic track', 'crawling track', and 'data track'.

11

Semantic Enrichment Design

- ▶ **Goal:**
 - ▶ Increase the **accuracy** of **text analytics**
- ▶ **Activities:**
 - ▶ Dictionary enrichment
 - ▶ entity
 - ▶ alias
 - ▶ entity/multi-word polarization
 - ▶ Inter-word relation definition

The flowchart for Semantic Enrichment Design is divided into two main tracks: the **semantic track** and the **data track**.

- **Macro-Analysis** (top) branches into **what** (sub-area, threads, topics) and **how** (Inquiries).

- **what** leads to **Ontology Design**, which produces a **domain ontology**.

- **how** leads to **Source Selection**, which produces **sources**.

- The **semantic track** includes **Semantic Enrichment Design** (using the domain ontology) to create an **ontology/dictionary**, followed by **Execution** to produce **enriched clips**.

- The **data track** includes **Crawling Design** (using templates and queries) to produce **clips**, followed by **Execution** to produce **key figures**.

- **ETL & OLAP Design** (using sources) leads to **ETL & reports**, followed by **Execution** to produce **key figures**.

- All tracks converge into a final **Test** phase.

12

ETL & OLAP Design

- ▶ **Goal:**
 - ▶ Design and develop the **analytics front-end** and specific **analysis metrics**
- ▶ **Activities:**
 - ▶ **ETL&OLAP design**, depends on
 - ▶ semantic engine features
 - ▶ presence of specific data acquisition channels (CRM, enterprise db, etc.)
 - ▶ **KPI design**, depends on
 - ▶ Users informative needs
 - ▶ Both depends on metadata richness and availability

The flowchart for ETL & OLAP Design is identical to the one in slide 11, showing the process from Macro-Analysis to Test, including Semantic track and data track.

13 Execution & Testing

- ▶ Has a basic role in the methodology
- ▶ **Coverage Analysis**
 - ▶ Measure the **ontology maturity level**
 - ▶ percentage of clips that include at least one ontology topic
- ▶ **Correctness Analysis**
 - ▶ Measure **actual improvements** the **overall ability** of the process in understanding a text
- ▶ **Crawling Coverage Analysis**
 - ▶ wrong query may lead to losing relevant clips
 - ▶ is a daily and critical task

14 Social BI Projects

- ▶ Social BI projects are characterized by:
 - ▶ Quickly **changing requirements** and **environment**
 - ▶ **Data sources** are not known a priori
 - ▶ Neither their structure
 - ▶ Project overall quality heavily depends on crawled content quality
 - ▶ **Keyword query** are in some situations rough tool
 - ▶ Cubes schema is **project independent**, mainly related to the **project domain**

15

Social BI Projects

▶ In the table below activities executed in projects of higher levels are carried out in lower levels too

Project Type	Crawling	Semantic Enrichment	Storing & Analysis
Level 1: Best-of-Breed	template design	dictionary enrichment, inter-word relat. def.	ETL design and impl.
Level 2: end-to-end	source selection, query design, content rel. analysis	polarization, correctness analysis, ontology coverage analysis	ontology design, KPI & dashboard design
Level 3: Off-the-Shelf	macro-analysis	macro-analysis	macro-analysis

End-To-End tuned to involved Acquire, in an aaS way, an off-the-shelf solution that usually support only a limited number of dashboard and is poorly customizable

16

Case Studies

- ▶ PR-CG:
 - ▶ Level 2 (end-to-end) project
 - ▶ Domain: large consumer goods company
 - ▶ Team guided by previous experiences (not SBI)
- ▶ PR-Pol:
 - ▶ Level 1 (Best of Breed) project
 - ▶ Domain: Italian politics
 - ▶ Metodology applied and enforced

▶ In both project the iterative approach were adopted†©

17


Case Studies

Activity / Task	PR-CG		PR-Pol	
	1st Iteration	Maint. Iteration	1st Iteration	Maint. Iteration
Macro Analysis	10	-	9	-
Ontology Design	4	0,6	7	1,5
Topics Definition	2	0,5	2	1
Inter-Topic Relation Definition	2	0,1	5	0,5
Source Selection	3	1	5	1
Semantic Enrichment Design	7	0,75	5	1
Crawling Design	10	1	29	1,5
Template Design	n.a.	n.a.	15	-
Query Design & Content Relevance Analysis	10	1	14	1,5
ETL & OLAP Design	15	-	24	-
ETL Design & Implementation	5	-	10	-
KPI Design	5	-	7	-
Dashboard Design	5	-	7	-
Execution & Test	3	-	5	-
Total	52	3,35	84	5
In charge of the customer	15	0,85	84	5

18

Outcomes


- ▶ **Responsiveness** in an SBI project is not a choice but rather a necessity
 - ▶ the frequency of changes requires
 - ▶ Continuous and tight **involvement** of domain experts
 - ▶ Change in **project managing**
 - ▶ huge **effort** to both end users and developers
 - ▶ If a proper methodology is not adopted the main **problems** are:
 - ▶ a **lack of synchronization** between the activities, that reduced their effectiveness
 - ▶ an **insufficient control** on the effects of changes (side effects)
 - ▶ With our methodology we tried to solve such problems through:
 - ▶ A clear **organization** of goals and tasks for each activity.
 - ▶ A protocol and a set of templates (not for brevity) to record and share information between activities to support collaboration
 - ▶ A set of tests to be applied



Outcomes

19

- ▶ Big Data raises many questions
 - ▶ Storing
 - ▶ OLAP with Big Data is far to be an explored topic
- ▶ Deep semantic analysis may largely increase the size of the data to be handled (70x)
- ▶ The polarization correctness has still a statistic value
 - ▶ is typically less than 70% when web/social sources are involved
 - ▶ May be higher than 90% on very specific sources, topics and vocabulary



Thank you
for your attention!

matteo.francia3@unibo.it

More informations and demos on:
big.csr.unibo.it/SBI