WebLab in the land of Big Data
What is WebLab?

An integration platform

- based on recognised standards (SOA, Web Services, Semantic Web)

- allowing the integration of a selection of software components (search engine, information extraction, translation, knowledge management, graphical representation using maps/networks, etc.)

- allowing the interoperation of the selected components.

A set of media mining services for document processing to be reused in all the WebLab projects.
What is WebLab?
A range of projects in the fields of media-mining, NLP & KM
WebLab in the land of Big Data

WebLab Scope

From unstructured data...

… to structured and actionable knowledge
WebLab in the land of Big Data

WebLab Scope

Structured Data
- Resides in formal data stores
- RDBMS
- Data Warehouse
- Spreadsheet
- OpenData
- ≈10%

Semi-structured Data
- Type of structured data that doesn’t conform with a model
- XML
- RSS
- ≈10%

Unstructured Data
- Raw text, written human language, images, video, audio
- Document
- Image
- Video
- Audio
- Text message
- Blogs
- Social media
- Logs & clicks
- Sensors/IoT
- ≈80%
WebLab Layers

Applications dedicated to specific domains
- Business intelligence
- Knowledge management
- Military intelligence
- Medias Analysis

Integration of unstructured information processing softwares
- COTS
- Open source
- Advanced research components

Open source SOA infrastructure for unstructured information processing components
- Enterprise Service Bus (ESB)
- Portal
- Orchestrator
- Exchange data model

Resources and Means for WebLab technology implementation
- Tutorials & Guides
- Services and portlets directory
- Tools for chain edition & construction
- Tools for application deployment
- Tools for services & portlets development
- Tools for system tests and performance assessment
- etc.
WebLab in the land of Big Data

WebLab Offer

- **WebLab Dummy Demo**
- **WebLab for Competitive Intelligence**
- **WebLab for Media Mining**
- **WebLab for Intelligence**

**Customer Care**
- Technical Support
- Training
- Consulting
- Knowledge Engineering
- Support, etc.

**Demo Application for Training**

**Competitive Intelligence**

**Other Business Cases:**
- Medias
- E-reputation
- Enterprise Search, etc.

**OSINT/HUMINT Integrated System**

**Technical Services with a low added value**
- (FOSS basic integration)

**Business and technical services with a high value added**
- (Enhanced FOSS integration, COTS integration, Business models and GUI, etc.)

**Middleware**
- (Exchange Model, Helpers, ESB, Orchestration, Monitoring, Logging, etc.)

**Open Source**

**Proprietary**

**WECLAB APPLICATIONS**
- *BUSINESS*

**WECLAB SERVICES**
- (WS + PORTLETS)

**WECLAB CORE + STUDIO**
An Example of a Processing Chain: From TV to annotations

**Video**

Collect

Video Segmentation

Audio

Audio Cleaning

Audio vocal

**Enriched text**

An international Greenpeace alpine team delivers messages of support and hope for the victims of the nuclear disaster at Fukushima Daiichi to the summit of Mt Fuji. Collected from thousands of people in Japan and all over the world, Greenpeace hopes that these messages will help unite the people...

**Translated text**

An international Greenpeace alpine team delivers messages of support and hope for the victims of the nuclear disaster at Fukushima Daiichi to the summit of Mt Fuji. Collected from thousands of people in Japan and all over the world, Greenpeace hopes that these messages will help unite the people...

**Text**

国際グリーンピース高山チームは富士山の頂上への支援と福島第一原子力災害の被害者のための希望のメッセージを配信します。日本と世界中の何千人もの人々から集めましたグリーンピースは、これらのメッセージは、原子力発電に反対する日本の人々を団結に役立つし、日本当局はそれらに耳を傾けることを奨励することを期待しています。

**Alert**

**information Extraction**

**Translation**

**Audio Transcription**
General Pattern of a WebLab Processing Chain

WebLab in the land of Big Data
Big Data technologies in a WebLab Processing Chain

- **Collection**
- **Processing**
- **Content Storage**
- **Indexing**
- **Knowledge Management**
- **Search**
- **Analytics**
- **Decision & Reporting**

- **mongoDB**
- **CouchDB**
- **Not Only SQL**
- **elasticsearch**
- **Virtuoso**
- **H2O**
- **Datameer**
- **SAS**
- **OWLIM**
- **OPENLINK SOFTWARE**
- **AIRBUS DEFENCE & SPACE**

WebLab in the land of Big Data
WebLab in the land of Big Data

The WebLab Triple « V »

OLUME

WebLab enables to exploit the vastest repository of unstructured data available for structuring

- 1992: 26 Web sites
- 1996: 100 000 web sites in January
  about 230 000 in June
- 2000: 19 823 296 Web sites
- 2011: 312 693 296 Web sites
- 2014: 1 000 000 000 Web sites

Getting information off the Internet is like taking a drink from a fire hydrant.

Mitchell Kapor
The WebLab Triple « V »

WebLab collects and processes various data from various sources

Natural Language processing technologies allow extraction of structured information (dates, locations, keywords, personal attributes, events, etc.) and data set buried deep in informal documents (lists, tables, form fields, etc.)
WebLab in the land of Big Data

The WebLab « Triple V »

Volume

Variety

Velocity

WebLab tries to absorb increasing data flow to process it « on the fly ».
WebLab in the land of Big Data

Thanks for your attention!

The reproduction, distribution and utilization of this document as well as the communication of its contents to others without express authorization is prohibited. Offenders will be held liable for the payment of damages.
All rights reserved in the event of the grant of a patent, utility model or design.