Towards Open Government
Metadata Harvesting
Regione Lazio Experience
Claudio Biancalana, Dante Chiroli
February 2014
LAit (LAzio Innovazione Tecnologica) is the Regione Lazio in-house Company dedicated to ICT innovation within the regional context.

According to this perspective LAit promotes the adoption of ICT solutions for:

- **Citizens and Businesses**, through the creation of value-added services for the simplification of administrative procedures and the promotion of *e-democracy*
- **Regional Administration**, through the development of computing and communication infrastructure, as well as basic application services
- **Regional Development**, by the promotion of initiatives to support new forms of local development, as the districts high-tech companies

In order to achieve its mission LAit designs, builds and manages the Information System of the Regione Lazio (SIR), for the 'digital administration'
The Information System of Regione Lazio (SIR) covers several regional domains of ICT, like Health, Instruction, Territory, Connectivity and so on.

The actual consistencies of the Regional SIR are

<table>
<thead>
<tr>
<th>Information Systems</th>
<th>Total Storage Capacity</th>
<th>CPUs (Socket)/vCPU</th>
<th>Operating systems Databases</th>
</tr>
</thead>
<tbody>
<tr>
<td>78 Application Projects 13 Infrastructural Projects</td>
<td>113 TB net 350 TB raw</td>
<td>493 / 807 x64</td>
<td>Windows Server/ Linux/ Oracle, MySQL, Postgres, SQLSERVER</td>
</tr>
</tbody>
</table>

In this context we aimed at deploying the open data possibility to enable third parties to leverage the potential of government data, as the development of applications and services that address public and private demands.

The Metadata collections should be considered as an harvesting activity rather than a static collection, in order to ensure their accuracy.
The opening of public databases to other government departments and agencies, citizens, businesses and organizations is a good way to improve the transparency and efficiency of public administration and an opportunity to create value-added services.
data as service

ICT Infrastructure

Raw Data

Aggregation/Organization

Publish Processing

Delivery

Final Service

Data as a Service
The reference framework consists of two components:

- **Data Provider**: it consists in a store that supports a protocol for accessing metadata contents.

- **Service Provider**: the archives of data providers, receives the metadata exposed and builds value-added services.
Collecting Metadata: taking the big picture

http://facweb.cs.depaul.edu/sgrais/collage.htm
David Hockney
Anzio: cittadinanza onoraria a Roger Waters


Waters è autore e voce solista, contiene dodici pezzi ispirati al rifiuto della guerra e dedicati alla figura del padre morto durante la seconda guerra mondiale, in seguito allo Sbarco di Anzio.

Per maggiori informazioni sull’evento:
Indirizzo: Villa Corsini Sarsina - Via Gaspare Ambrosini
Tel: +39 06 98499314
Visualizza il programma completo della manifestazione

<XML>
<TITLE>Anzio: cittadinanza onoraria a Roger Waters</TITLE>
<DATE>18 Febbraio 11.00</DATE>
<LOCATION>Villa Corsini Sarsina</LOCATION>
<TAGS>Pink Floyd, Roger Water, Musica, Cittadinanza</TAGS>
</XML>
Language Analysis Stack

- **Domain Specific**
  - Events & Facts
  - Entities
    - Candidates, Resolution, Normalization
  - Basic NLP
    - Noun Groups, Verb Groups, Numbers Phrases, Abbreviations
  - Metadata Analysis
    - Title, Date, Body, Paragraph
  - Sentence Marking
  - Morphological Analyzer
    - POS Tagging (per word)
    - Stem, Tense, Aspect, Singular/Plural
    - Gender, Prefix/Suffix Separation
  - Tokenization

Copyright 2014 LAit S.p.A.
Metadata Extraction – General Architecture

Service Provider

Output API

XML

DB Output

Metadata Extractor

Entity, Fact, Event Extraction
Categorizer
Language ID

Input API

Web Crawlers (Agents)
File Based Connector
RDBMS Connector
Programmatic API (SOAP web Service)

Control API

Data Provider

Console
Data Extraction and Collection - POC

RAW DATA

Text Normalization

Text Segmentation

Named Entity Recognition

GeoCoding

KEA Based Extractor

Source Specific Extractor

LocalDB

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th>Sub-category</th>
<th>City</th>
<th>Address</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Plain Tags: (wine, Pakistani, Indian, Haveli, naan, tika, Curry, samosa)

Semantic Tags: (Take out: yes, Meal served: dinner, ...)

KEA

Keyphrase extraction algorithm

WEKA

The University of Waikato
Conclusions and Future Works

- Proposed Framework
  - Harvesting
  - Metadata Extraction

- Where we go?
  - Metadata Refresh
  - Metadata Merging
  - Personalization in e-Government
  - Context-aware Open Data
Towards Open Government
Metadata Harvesting
Regione Lazio Experience
Claudio Biancalana, Dante Chirolı
February 2014