XLike
Kickoff Meeting
Bled, Jan 18th 2012
Marko Grobelnik (JSI)
Outline

- Genesis of the proposal
- Previous collaborations
- Rationale, scope and objectives
- Expected results and their use
- Possible extensions of the project
Small or medium scale focused research project (STREP)

Date of preparation: January, 18 2011

Version number: 1.0

Work programme topic addressed:
Objective ICT-2011.4.2 – Language Technologies
b) Information access and mining
   (i) cross-lingual information search and retrieval
   (iii) text mining and information extraction from multilingual collections

Name of the coordinating person: Marko Grobelnik
E-mail: marko.grobelnik@ijs.si
Fax: +386 (1) 251 93 85

<table>
<thead>
<tr>
<th>Participant no.</th>
<th>Participant organisation name</th>
<th>Participant short name</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Coordinator)</td>
<td>Jožef Stefan Institute</td>
<td>JSI</td>
<td>Slovenia</td>
</tr>
<tr>
<td>2</td>
<td>Karlsruhe Institute of Technology</td>
<td>KIT</td>
<td>Germany</td>
</tr>
<tr>
<td>3</td>
<td>Universitat Politècnica de Catalunya</td>
<td>UPC</td>
<td>Spain</td>
</tr>
<tr>
<td>4</td>
<td>University Zagreb</td>
<td>FFZG</td>
<td>Croatia</td>
</tr>
<tr>
<td>5</td>
<td>Tsinghua University</td>
<td>THU</td>
<td>China</td>
</tr>
<tr>
<td>6</td>
<td>iSOCO</td>
<td>ISOCO</td>
<td>Spain</td>
</tr>
<tr>
<td>7</td>
<td>Bloomberg</td>
<td>BB</td>
<td>UK</td>
</tr>
<tr>
<td>8</td>
<td>Slovenska Tiskovna Agencija</td>
<td>STA</td>
<td>Slovenia</td>
</tr>
</tbody>
</table>
Genesis of the proposal (the very beginning)

- Initial idea for the project proposal came in a discussion between BT and JSI in Sep 2011
  - ...to cover social media for Olympic games in 2012
  - ...and Reuters with financial news
  - Both cases had strong need for cross-lingual technologies, knowledge extraction and scalability

- At the same time, from the research side, there was an innovative solution between Cycorp and JSI for translating text to logic
  - ...this inspired proposal beyond statistical approaches

- Altogether seemed perfect fit to the Call 7 objectives
Genesis of the proposal (evolution)

- ... KIT joined the core group
- The idea evolved, BT had to step out
- Case studies got replaced with Bloomberg (financial news), IHT/NYT (general news) and Wikipedia (social media)
- Wikipedia couldn’t join, IHT/NYT was slow getting internal confirmation, STA joined instead
Genesis of the proposal
(final developments)

- As the core linguistic partner **UPC** joined due to its extensive experience and technology in multilingual parsing.
- **University Thsingua** appeared as a very relevant partner to cover Chinese language together with semantic and mining technology.
- **University Zagreb** for good experience in dealing with machine translation technology and evaluation.
- **iSOCO** joined as experienced SME in semantic and language technology taking care for engineering parts.
JSI, KIT and iSOCO have long tradition of collaborating on several FP6 and FP7 projects in semantics, language and machine learning.

University Zagreb and JSI have long term relationship.

UPC worked on several EU linguistic projects.

Bloomberg has a few year collaboration with JSI and one of its spin-offs.

For University Thsingua XLike is the second FP7 project.

STA, and Bloomberg are first time on EU projects.
There are two key research objectives:

1. To extract and integrate formal knowledge from multilingual texts with cross-lingual knowledge bases, and

2. To adapt linguistic techniques and crowdsourcing to deal with irregularities in informal language used primarily in social media
Increasing proportion of non-English articles in Wikipedia: the need for cross-linguality
Research areas dealing with text: the need for integration of ideas and approaches

THE ELEPHANT METAPHOR OF REALITY

Text Understanding

mappings

Machine Translation

Computational Linguistics

Semantic Web

Text Mining

Information Retrieval

language

community

interoperability

extraction

statistics

search

Crowdsourcing

Machine Learning

XLike
The goal of the project: logic and statistical based interlingua
The architecture of the project: processing pipeline

WP1
Multilingual data sources:
Mainstream news, blogs, Twitter, Facebook

WP2
Multilingual: Named entity extraction, POS tagging, dependency parsing, informal languages

WP3
Cross-lingual: Disambiguation, semantic annotation, fact extraction, crowd-sourcing

WP4
Cross-lingual: Document linking, semantic graphs, event extraction

WP5
Cross-lingual: Multi-document summarization, visualization, news bias, trend analysis, complex event extraction

WP6
Toolkit:
Architecture API, exploration, front-end

Multilingual data feeds
Multilingual linguistic processing
Cross-lingual semantic annotation
Cross-lingual semantic integration
Reporting & analytics components
Toolkit & interface
Execution of the project: fast development cycles

Year 1
- M3: Requirements Analysis
- M9: Agile Development Metrics / Benchmarking
- M12: Evaluation Assessment

Year 2
- M15: Requirements Analysis
- M21: Agile Development Metrics / Benchmarking
- M24: Evaluation Assessment

Year 3
- M26: Req’ Analysis
- M30: Agile Development Metrics / Benchmarking
- M36: Evaluation, Assessment Final Roll-Out
Project work-plan: main segments

- Early prototypes
- Demonstrators and components
- Integrated system

Deliverables:
- Validated early prototypes
- Validated demonstrator
- Use case evaluation report

Analysis and Development:
- Needs analysis and requirements refinement
- Prototype evaluation
- Field tests

Evaluation:

Fast development cycles between test and redesign
Case studies

“Slovenian Press Agency” covers general news with focus on Slovenia in English and Slovenian language

…while “Bloomberg” delivers mainly Financial and business news in English

Both have strong need for multi- and cross-linguality
Expected results and their use

In terms of research:
- The key result would be using semantic and statistical interlingua for linking information across languages.

In terms of technology:
- The result is XLike Toolkit, open source environment for cross-lingual language processing.
- Easy reusability and potential further commercial exploitation.

Source: “Galileo Galilei was an Italian physicist and astronomer.”

Learn Logic: ($\&$ and ($\&$isa $\&$GalileoGalilei $\&$ItalianPerson)
($\&$isa $\&$GalileoGalilei $\&$Physicist)
($\&$isa $\&$GalileoGalilei $\&$Astronomer))

Fact: Galileo was an Italian, a physicist, and an astronomer.
Possible extensions of the project (1/2)

- Participation of IIT Bombay to cover Hindi language on their own cost
  - ...group of prof. Pushpakk Bhattacharyya developing language resources for Hindi
  - ...from XLike side some minor costs should be covered

- Participation of NYTimes R&D department
  - General news scenario
  - ...from XLike side, one trip per meeting should be covered
Possible extensions of the project (2/2)

- Active Advisory board to assist in expert advice
  - Active group of invited experts bringing operational expertise to assist critical parts of XLike research
  - XLike should cover on average one or two trips per expert per meeting
Some administrative info

- Web site:
  - [http://xlike.org](http://xlike.org)

- Document exchange:
  - DropBox shared XLike folder

- Administrative contact:
  - Mojca Kregar Zavrl ([mojca.kregar@ijs.si](mailto:mojca.kregar@ijs.si))

- Project Manager:
  - Blaz Fortuna ([blaz.fortuna@ijs.si](mailto:blaz.fortuna@ijs.si))

- EC Project Officer:
  - Hanna Klimek ([Hanna.KLIMEK@ec.europa.eu](mailto:Hanna.KLIMEK@ec.europa.eu))
<table>
<thead>
<tr>
<th>Milestone no.</th>
<th>Milestone name</th>
<th>Work packages involved</th>
<th>Expected date</th>
<th>Means of verification</th>
</tr>
</thead>
</table>
| 1            | Inception check         | WP1, WP2, WP6, WP8, WP9 | M3            | Project Management: Project satisfactorily commenced, all partners fully resourcing project  
Dissemination and Exploitation: Website established  
Requirements: Initial Requirements captured |
| 2            | Project establishment   | All                    | M12           | RTD: Initial components available, Pipeline functional  
Dissemination and Exploitation: Initial presentation material  
Use cases: Initialized, First data evaluations finished  
Project Management: Deliver initial progress report |
| 3            | Project refinement      | All                    | M24           | RTD: Demonstrator components available, published and documented, Pipeline refined and open for external development  
Use cases: Experience from initial feedback round fully incorporated, demonstrator version evaluated  
Project Management: Carry out risk assessment |
| 4            | Project validation      | All but WP1            | M30           | RTD: Final version of pipeline components completed  
Use cases: Use case experiences fed back into component and use case development |
| 5            | Project impact          | WP6, WP7, WP8, WP9     | M36           | Dissemination and Exploitation: Exploitation plans from all industrial partners in place  
Use cases: Evaluation finished and reported  
Project Management: Completion of all project objectives, technical, commercial, and disseminational |
## Workpackage load

<table>
<thead>
<tr>
<th></th>
<th>JSI</th>
<th>KIT</th>
<th>UPC</th>
<th>FFZG</th>
<th>THU</th>
<th>ISOCO</th>
<th>BB</th>
<th>STA</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP1</td>
<td>10*</td>
<td>10</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>8</td>
<td>8</td>
<td>57</td>
</tr>
<tr>
<td>WP2</td>
<td>0</td>
<td>6</td>
<td>48*</td>
<td>12</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>WP3</td>
<td>9</td>
<td>30*</td>
<td>12</td>
<td>24</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>WP4</td>
<td>33*</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>WP5</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>18*</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>54</td>
</tr>
<tr>
<td>WP6</td>
<td>9</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>11</td>
<td>38*</td>
<td>0</td>
<td>0</td>
<td>67</td>
</tr>
<tr>
<td>WP7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>18*</td>
<td>18</td>
<td>46</td>
</tr>
<tr>
<td>WP8</td>
<td>5</td>
<td>8</td>
<td>2</td>
<td>4*</td>
<td>2</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>WP9</td>
<td>26*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
<td>81</td>
<td>80</td>
<td>56</td>
<td>64</td>
<td>72</td>
<td>32</td>
<td>32</td>
<td>521</td>
</tr>
</tbody>
</table>
Interdependencies among WPs

WP1: Definition and Data Provision

WP2: Multi-lingual Linguistic Processing

WP3: Cross-lingual Semantic Annotation

WP4: Cross-lingual Semantic Integration

WP5: Reporting and Analytics

WP6: Integration and Toolkit

WP7: Use Cases and Evaluation

WP8: Dissemination, Exploitation and Community Building

WP9: Management