

UNIVERSITY OF HELSINKI

PULS

University of Helsinki, Finland

Content Collection and Analysis in the Domain of Epidemiology

Göteborg, 26 May 2008

Roman Yangarber, Peter von Etter, Ralf Steinberger

doremi.cs.helsinki.fi/puls/databases/medical

Outline

- Introduction
- Information and Intelligence
- MedISys: Information Retrieval
- PULS: Information Extraction
- MedISys/PULS Integration
- Aggregation of Information
- Performance evaluation
- Current work
- Context of development
- Technical description
- Users
- Outputs
- Coverage: sources and languages
- Interoperability features
- Current collaboration with other systems
- Upcoming developments

Context of development

- PULS: University of Helsinki, Department of Computer Science
 - Natural language processing group
 - TEKES funded project: ContentFactory (June 2008 – end 2010)
 - Text mining
 - Ontology creation
- “Algodan: Algorithmic Data Analysis”
 - National Center of Excellence of the Academy of Finland (2008-2013)
 - Machine Learning and Data Mining
 - Tight collaboration among teams
- Our focus: language analysis and text mining
- **Colaboration:**
 - EC’s Joint Research Center – JRC
 - Users

Pascal

Users and motivation

Users:

- Organizations: ECDC, national Health Authorities/Agencies

Goals:

- Create a global portal for epidemic surveillance
- Automatic labeling of resources
 - Help users label resources
- We want to deliver tools to help users navigate medical information for **epidemic surveillance**
- A simple and effective *early-warning* system
- More ?...

Related Work

- (In paper)

Information vs. Intelligence

- Timely – no time lost from the information's initial appearance to delivery
 - everyone is 100% intelligent in hindsight
- Complete – no information missed, from any source
 - basically: high recall
- **Concise** – no information overload
 - not quite about precision, but related
- Last requirement – conciseness – has two complementary aspects:
 - no **redundancy**, no unnecessary detail:
 - deliver only the minimum necessary information
 - If the user wishes more detail, s/he can request further elaboration
 - comprehensive **background** knowledge:
 - for a system to identify *genuinely new* and *critical* information, with high confidence, it must know whether this information truly new, or has already been known, at some earlier time or from some different source

Combination of Technologies

Delivering critical intelligence entails combining following key technologies:

- Information retrieval
- Information extraction
- Information aggregation
- Information visualisation and exploration

Outline

- Introduction: Information and Intelligence
- **MedISys: Information Retrieval**
- PULS: Information Extraction
- MedISys/PULS Integration
- Information Aggregation
- Performance evaluation
- Current work

Information Retrieval: MedISys

Developed at JRC, initiated by DG-Sanco:

Key functionality:

- Keyword-based, boolean queries, select documents
- Statistics: running averages, expected hits, etc.
- Estimate threat levels: alerts

- 1500 news portals globally
- 40+ languages
- 50K items/day

Demo:

- medusa.jrc.it

Outline

- Introduction: Information and Intelligence
- MedISys: Information Retrieval
- **PULS: Information Extraction**
- MedISys/PULS Integration
- Information Aggregation
- Performance evaluation
- Current work

MedISys à Beyond IR

Rationale: **Fact Analytics** can further enhance functionality of MedISys

- Provide **specific facts**, extracted from text documents found by MedISys
- Boost **precision**
 - keyword-based queries may trigger on documents which are off-topic but happen to mention the alerts in unrelated contexts
 - pattern matching in IE provides the mechanism that assures that the keywords appear in relevant contexts only

PULS

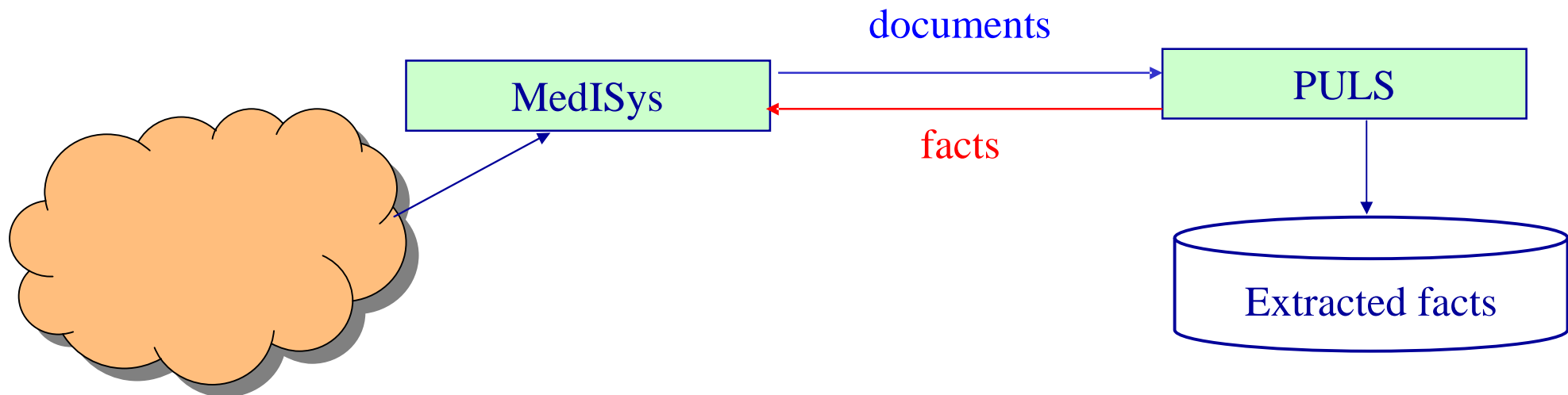
- Pattern-based Understanding and Learning System
- Customized for several application domains:
 - *epidemic surveillance*
- On-line database

Event Extraction: review

- Analyse free text à extract structured facts
- Input: plain, free text documents
 - Unstructured
- Output: facts = structured information
 - Facts extracted from text
- Operate at the **semantic** level

MedISys/PULS Integration

- RSS tunnel
- MedISys sends new documents
- PULS returns new events
- Live exchange of data every 10 minutes

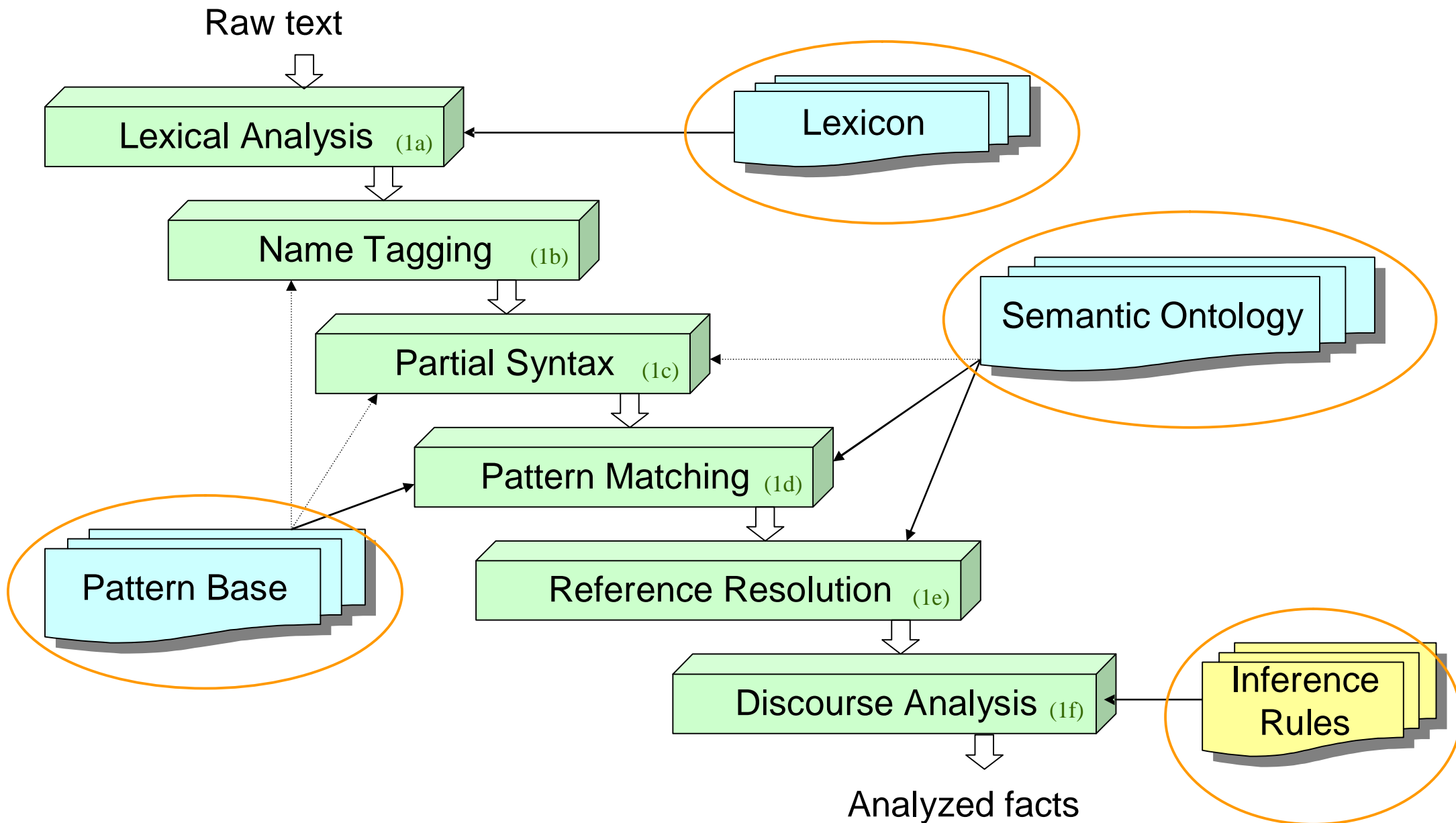


- **Demo:**
 - records that PULS returns to MedISys
 - top five: latest, most urgent
 - top fifty

Epidemics scenario

- Domain: epidemiological **news** reports
- Sources: **plain-text** documents
 - JRC/MedISys
 - ProMED-Mail list
- Extract facts (=incidents)
- An incident : atomic event = record in a database
- Attributes:
 - **Disease**
 - **Location**
 - **Date**
 - **Victims**
 - Descriptor: {people | animals | plants}
 - Number
 - Status: {affected | dead}
 - other...
- **Demo**: doremi.cs.helsinki.fi/jrc

Core IE Engine and Knowledge bases (KBs)



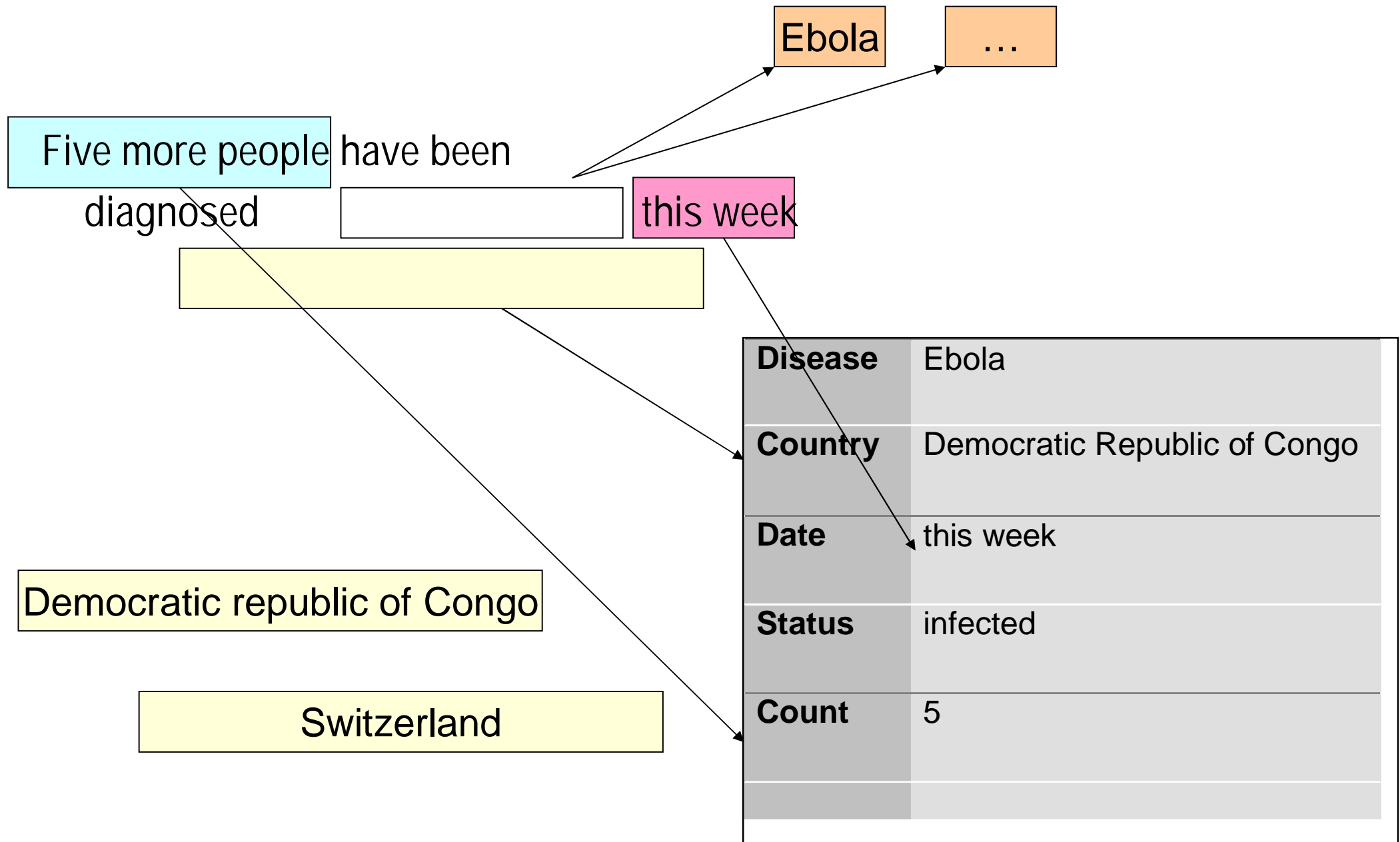
Example: Event extraction (fact extraction)

np(Victim) *vp-pass(diagnose)* 'with' *np(Disease)* [*np(Date)*] ['in' *np(Location)*]

Five more people have been
diagnosed with Ebola during this week
in the Democratic Republic of Congo

| | |
|----------------|------------------------------|
| Disease | Ebola |
| Country | Democratic Republic of Congo |
| Date | this week |
| Status | infected |
| Count | 5 |
| | |
| | |

IE and Semantics: Elided attributes



Aggregation into Group

Group/thread = chain of incidents/records

- Linking criteria:
 - "related" disease
 - "nearby" in location
 - "nearby" in time
- Simple idea: simplistic, to be refined in the future
 - in reality outbreaks are graphs of incidents
 - 10 people contracted the virus in Uganda since the start of the year
 - 5 health workers died in Gulu province last month

NB: Aggregation takes place across incidents:

- documents
- sources

Aggregation of information in PULS

- Link incidents into threads
 - across time
- Estimate **confidence**
 - Local: document level
 - Global: thread level

Result/Output of PULS

- Fill tables
 - In database, spreadsheet, ...
- Annotate resources with **meta-data**

- Main Database
 - Table view
- For aggregation of information, generates **aggregate views**
 - Threads (model epidemics/spread)
 - Geographic map
 - Chart/histogram

- **Demo:** doremi.cs.helsinki.fi/jrc

Current work

- Key research challenges:
 - **Customization** to new domains
 - **Aggregation** of information across
 - sources,
 - time,
 - Languages
- More languages
- Assessment of Resource quality
 - Reliability: à link analysis
 - Timeliness

Fin

-

Questions: User pull

- What is USEFUL to the user/community?
 - technology **push** vs. user **pull**
- "Current" vs. "Past" information:
 - value in "current" information is clear: early warning and alerting
 - what about "past" information?
 - example: ProMED-mail –
 - serves entire archive on the Web -- clearly user must see value in that
- What exactly is of value in the archive, and in serving the archive?
Not a frivolous question:
 - what does the researcher look for in the archive?
 - who is the researcher who uses the archive?
 - ...
- Must define needs/use-cases, before proceeding to refine technology
 - Else, danger: forever adding features, and asking: is this good? what about this one? ...

Questions: Level of analysis

- Do we aim for a system that knows and understands everything?
 - probably not: not fully feasible today (recall too low).
- Do we aim for a system that “just” points in the right direction, without understanding?
 - Probably not enough, (precision is too low)
- Likely, what is needed is something in between: what?

- What about modeling? System that models (more or less) completely
 - current situation?
 - past situation?
 - ... future?
- → cannot model the present without modeling past.

Questions:

- What is needed is something in between:
- This is what we try to do, e.g., with ECDC and JRC/MedISys,
 - some things we "know" via thorough analysis
 - other things we just have a hunch, though strong indicators/statistics
- Up to the user to learn to utilize the system optimally: can we help?
- Need to know what the user wants/needs/is trying to achieve
- We (technology side) are not supposed to know,
- ==> need to work together:
- That means in a joint project.
 - This is what is done when want to achieve real results,
 - The problem is truly interdisciplinary
 - Requires active, continual input and work from all sides
- **Very important: evaluation**

Acknowledgements:

- Academy of Finland,
- Algodan National Center of Excellence, "Algorithmic Data Analysis"