Research Infrastructures in the Humanities and the role of Language Technologies

CLARIN: A European Research Infrastructure Project

Marko Tadić

Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb

Croatian Academy of Sciences and Arts

marko.tadic@ffzg.hr

Teksty kultury uczestnictwa
Polish Academy of Sciences, Warsaw, 2013-11-07

Paradigm of eScience

- John Taylor
  - Director General of Research Councils, UK Office of Science and Technology
  - “eScience is about global collaboration in key areas of science and the next generation of infrastructures that will enable it”
- **Research Infrastructures** are not another set of short-term technological projects, but
  - long-term investments in services that enable the first class research by combining and sharing resources
  - contain a technological component that demands wisely defined standards for storing and accessing data
- ESF published SPB42 *Research Infrastrucures in the Digital Humanities*

What is research infrastructure?

- different meanings of term “research infrastructure” (RI)
  - a network of funding agencies (FA)
  - a network of research institutions
  - a network of professional associations or initiatives
  - a network of facilities that supports research
    - LHC, accelerators, telescopes, oceanographic ships,...
  - a network of archives of research results
    - digital libraries
    - Open Access initiative (supported by ESF)
  - a network of archives of research data and tools to access them
    - empirical data submittable to impartial examination
    - Permanent Access initiative (supported by ESF)
- this last meaning is important for Humanities (as well as for Social Sciences) (HSS)

What is research infrastructure?

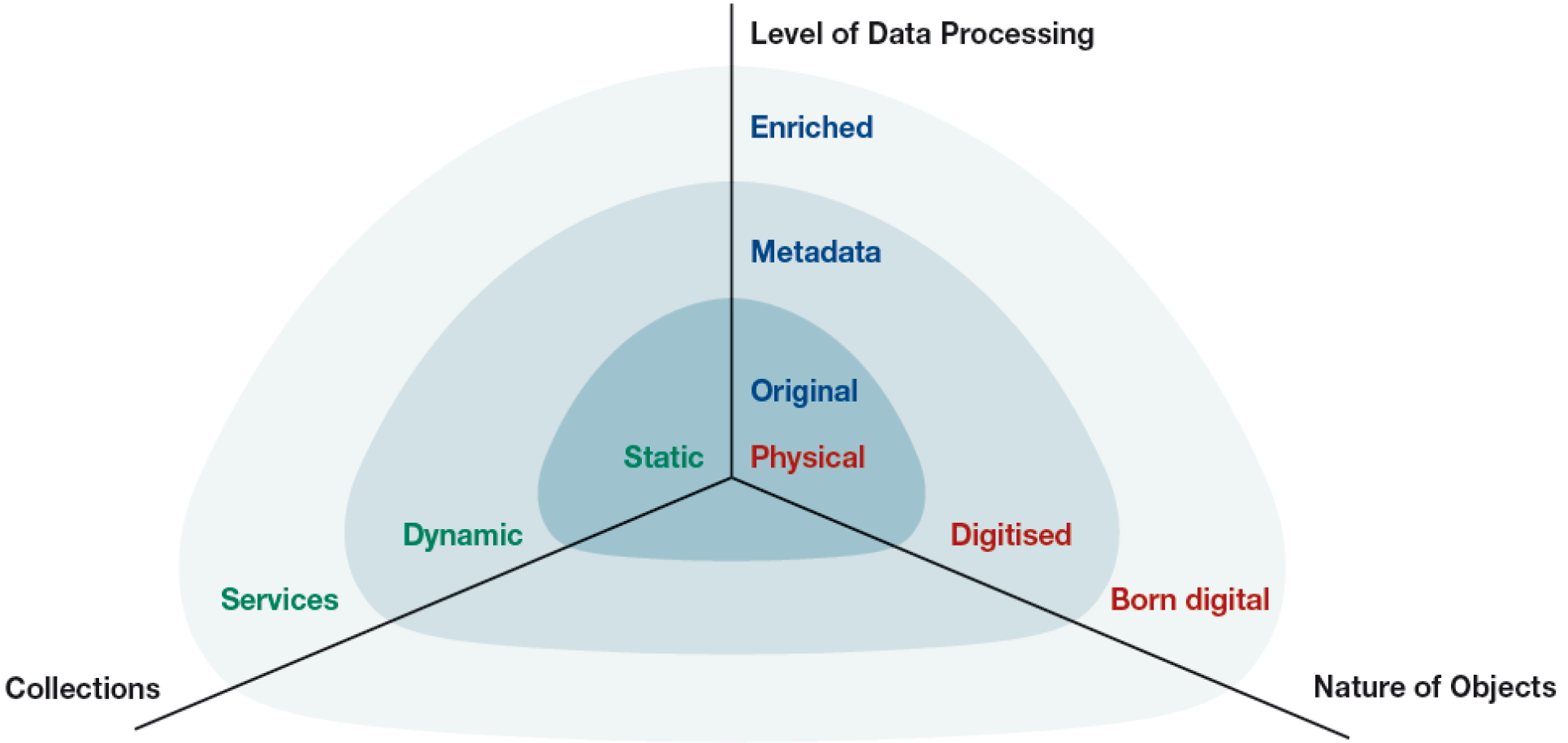
- behind the last meaning of “RI” lays a fundamental epistemological hypothesis: **text is important for HSS**
  - object of research is text itself (and its language/speech)
    - linguistics, all (neo)philologies, phonetics, etc.
  - object of research is mediated through text
    - theory of literature, history, sociology, psychology, cognitive science, information sciences, etc.
- research data for HSS in e-text format: growing rapidly
  - massive digitalisation of traditional texts from paper
  - massive production of new e-text (e.g. Wikipedia, social media...)
- (computational) linguistics (i.e. formal approaches to language)
  - mathematics in HSS (at least for text-based sciences)
- Language Technologies (LT) = its technological application

What is research infrastructure?

- **physical infrastructures**
  - collections of physical objects/installations/vessels/instruments (single-sited or hosted by more than one institution/country)
- **digital data infrastructures**
  - single sited or interconnected data repositories
- **e-infrastructures**
  - networks and/or computing facilities spread over various locations
  - the technical backbone of a given RI: e.g. GRID computing, cluster or cloud computing and the networks that connect them
- **meta-infrastructures**
  - conglomerates of independent RIs, residing in different institutions/ countries with different data formats and data structures (i.e., resulting from different activities)
  - connected using compatible metadata formats or processes, thus enabling access to different data archives

What is research infrastructure?

- Set of concurrent criteria for defining the RI in Humanities



The problems of eHSS

- most of data in digital archives are textually based
  - written language, but also combined with pictures and/or sounds
- many archives known only to local insiders
- archives are mostly not connected
- most of archives have their own standards for storage and access
  - usually only simple retrieval of files (text, audio or video documents) with proprietary indexing and annotation system
  - usually only simple search through files (Ctrl-F or F3)
- HSS researchers are often not aware of the potential benefits of using language technology tools to access e-text
- these tools are hard to use for non-specialist
- usually HSS researchers are not language technologists

The CLARIN Mission

- what?
  - create a Research Infrastructure that makes language resources and technologies (LRT) available to scholars of all disciplines, especially HSS
- how?
  - provide **European federation of digital archives** with language data and tools (text, speech, multimodal, gesture…)
  - target audience: **humanities and social sciences scholars**
  - access: **single uniform sign-in** access to the archives
  - means: **language and speech technology tools** to retrieve, manipulate, enhance, explore and exploit data
  - coverage: **all languages** are equally important
  - location: **all EU and associated countries** included

Why a European RI?

- too much fragmentation
- lack of
  - coordination
  - visibility
  - interoperability
  - sustainability

  achievable by using unifying standards (TEI, IsoCAT,...)
- expertise exists but not in all countries
- language independent tools can be shared
- language dependent tools can often be ported
- most countries not able to bear the cost of infrastructure alone

Why now?

- exponentia
  - e-text en
- maturity of
  - allows fo
  - allows fo
  - allows fo
  - allows fo
- growing int
- ESFRI Roa
  proposals f
  - HSS: CE
- European
  - since 20

Human-Generated Internet Content: Zettabyte of Unstructured BIG Data

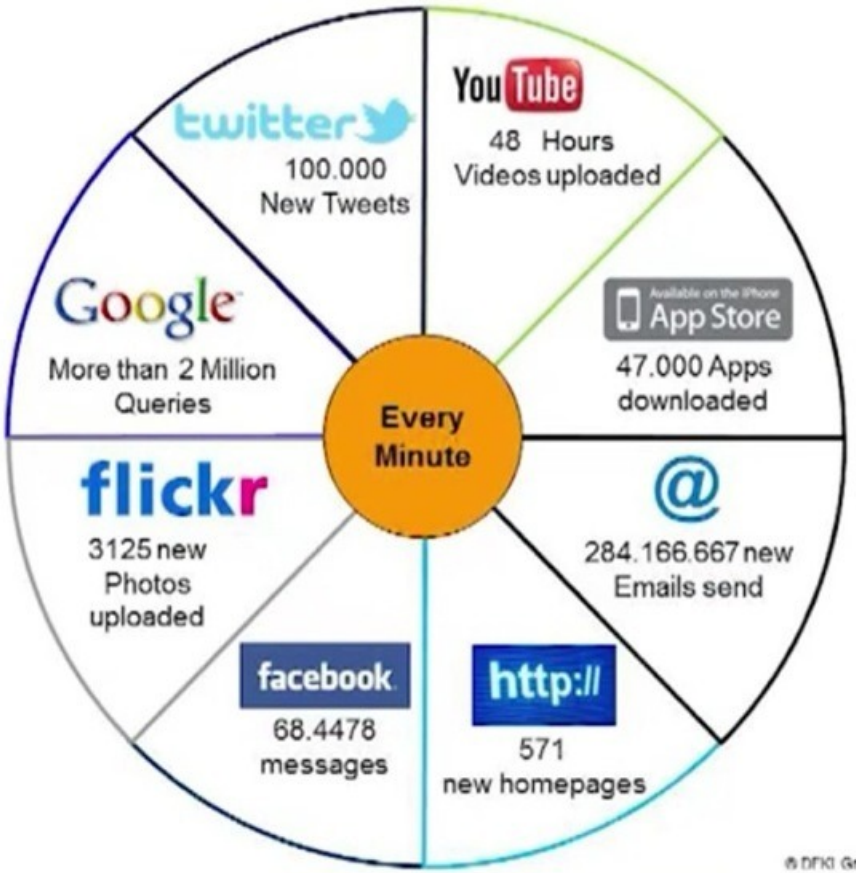
META FORUM 2013

Exponential Growth of Internet Data:

Commercial Spoken Language Access:

Siri (Apple)
Google Now (Google)
S Voice (Samsung)
Cortana (Microsoft)

Still Missing:
Crosslingual Information Retrieval



Wolfgang Wahlster: Language Technology for Big Data Analytics. META-FORUM 2013-09-19, Berlin

What was CLARIN project?

- CLARIN stands for
  - Common Language Resources and Technology Infrastructure (for the Humanities and Social Sciences)
- FP7 RI project
  - started 2008-01-01, ended 2011-06-30
- development of RI in 3 phases
  - preparatory phase
    - 2008-2011: planning, building a prototype
    - budget: 4.1 M€ from EC, ca 5 M€ from participating countries
  - construction phase
    - 2011-2015: building and populating with resources and tools
    - national CLARIN consortia (CLARIN-NL, CLARIN-D, CLARIN-PL...)
  - exploitation phase
    - 2016-: CLARIN infrastructure in full service

What was CLARIN project?

- CLARIN consortium
  - 33 partners from 23 EU and accessing countries
- CLARIN community
  - 151 member from 33 countries
  - interest from Americas, Asia, Australia and Africa
- leading partners
  - Utrecht University (Steven Krauwer, coordinator)
  - Max Planck Institute, Nijmegen (Peter Wittenburg)
  - Oxford University, OTA (Martin Wynne)
  - Hungarian Academy (Tamás Váradi)
  - ...
- CLARIN was based on four previous initiatives
  - LangWeb, EARL, TELRI I & II, DAM-LR

What was CLARIN project?

- CLARIN started as a Preparatory Phase project
  - bottom-up building procedure (incl. national CLARIN teams)
  - large LRT registry (840 LR, 150 LT, 195 languages) produced
  - technical specification defined
  - a series of usage scenarios for HSS as demo prototypes
- procedures
  - using/adapting standards (if necessary also creating new)
  - using metadata (as a “glue” between digital archives)
  - using persistent identifiers (PID), authorisation/autentification
  - offering web-access to LR
  - offering LT as webservices

What CLARIN project wasn't?

- building infrastructure
  - it was only being planned
- creating new resources
  - we were combining and adapting existing resources
- doing new research
  - we were using available results
- focusing on the big languages
  - we found all languages equally important

What is CLARIN ERIC?

- established 2012-04-18
- 9 founding members (countries or international entities)
  - Austria, Bulgaria, Czech Republic, Danmark, Estonia, Germany, Netherlands, Dutch Language Union, Poland
  - some countries signed the memorandum of understanding (Norway, Croatia, Lithuania...)
- each member is supporting a national CLARIN consortium
  - building the infrastructure at national level
  - connecting the infrastructure with CLARIN-ERIC at European level
- http://www.clarin.eu

Summing up

- one day the HSS scholar should be able to use the same unified web interface to infrastructure and ask questions like:
  - "List all uses of enthusiasm in 19th century English novels written by women."
  - "Display all maps between 15th and 18th century showing Danube and Vienna."
  - "Show me all possible translations of saudade in Camões' works and which of them appeared in Lithuanian translations."
  - "Summarize *Le Monde* of June 11th 2009 – in Polish."
  - "Who was mentioned in the same documents with Konrad Adenauer during 1951?"
  - full-blown named entity recognition of web version of daily newspaper for under-resourced language (Croatian) →

Home

Aktualnosti
Vijesti
Crna kronika
Sport
Scena
Kultura
Gospodarstvo
Poslovni svijet
Zanimljivosti
Regije

Free Time
Kompas
Događanja
Kino
TV Vodič
Vrijeme
Lifestyle
Nedjeljni Večernji

NOVO
Vijesti bez slika
Moja karijera
e-Shop

→ Gospodarstvo → Poslovni svijet

25.01.2005 18:45

Hrvatski izvoz još na niskim razinama

90 posto tvrtki uopće ne izvozi!

Autor **Piše Josip Bohutinski**

Hrvatski izvoz napokon je prošle godine počeo rasti brže od uvoza te je, prema podacima za prvih 11 mjeseci 2004. godine, izvoz u kunama rastao 15,7 posto a uvoz 5,7 posto. Iz Hrvatske je izvezeno robe u vrijednosti nešto manjoj od 44 milijardi kuna ili 7,25 milijardi američkih dolara, dok je vrijednost uvoza bila 91,19 milijardi kuna ili više od 15 milijardi dolara.

No podaci o izvozu po glavi stanovnika upozoravaju da je hrvatski izvoz još na niskim razinama u usporedbi s drugim i sličnim zemljama. Prema podacima udruge Hrvatski izvoznici, u 2003. godini vrijednost hrvatskog izvoza po glavi stanovnika bila je samo 1106 dolara.

Koliko je je to mala vrijednost, govori podatak o slovenskom izvozu po glavi stanovnika od čak 4774 dolara. Irska na svakog svoga stanovnika izveze 22.119 dolara roba i usluga. Amerikanci, pak, po glavi stanovnika izvezu robe u vrijednosti 2360

Ostale vijesti
- Novi igrači zajedno protiv T-HT-a?
- Hrvatska ipak nije prezadužena
- Povjerenstva odvaguju veletrgovine
- Državne banke kreću u investicijsko bankarstvo
- Dubrovnik ne bi smio dominirati u promidžbi
- Fokus na Sirelu i Somboled
- Dioničari ne žele novac nego banku

Prijava
Korisničko ime
Lozinka
Prijava
→ Registrirajte se!
→ Zaboravili ste lozinku?

Kolumna

→ Nakon revolucije, slijedi evolucija

→ Devizno tržište
→ Tržište novca
→ Tržište obveznica
→ Tržište kapitala
→ Karijere

Večernji list, 2005-02-17, Gospodarstvo

Hrvatski izvoz još na niskim razinama

90 posto tvrtki uopće ne izvozi!

Autor Piše Josip Bohutinski

Hrvatski izvoz napokon je **prošle godine** počeo rasti brže od uvoza te je, prema podacima za **prvih 11 mjeseci 2004. godine**, izvoz u kunama rastao **15,7 posto** a uvoz **5,7 posto**. Iz **Hrvatske** je izvezeno robe u vrijednosti nešto manjoj od **44 milijardi kuna** ili **7,25 milijardi američkih dolara**, dok je vrijednost uvoza bila **91,19 milijardi kuna** ili više od **15 milijardi dolara**.

No podaci o izvozu po glavi stanovnika upozoravaju da je hrvatski izvoz još na niskim razinama u usporedbi s drugim i sličnim zemljama. Prema podacima udruge Hrvatski izvoznici, u **2003. godini** vrijednost hrvatskog izvoza po glavi stanovnika bila je samo **1106 dolara**.

Koliko je je to mala vrijednost, govori podatak o slovenskom izvozu po glavi stanovnika od čak **4774 dolara**. **Irska** na svakog svoga stanovnika izveze **22.119 dolara** roba i usluga. Amerikanci, pak, po glavi stanovnika izvezu robe u vrijednosti **2360 dolara**.

No vrijednost izvoza velikih zemalja po glavi stanovnika u pravilu je manja od izvoza malih zemalja zbog velikog domaćeg tržišta koje može apsorbirati veliki dio domaće prozivodnje. To potvrđuju i podaci o izvozu po stanovniku i "malih zemalja" poput **Belgije**, **Nizozemske** i **Finske**.

Uz malu vrijednost izvoza po glavi stanovnika, za **Hrvatsku** je nepovoljan i podatak o broju domaćih tvrtki čija godišnja vrijednost izvoza premašuje **milijun kuna**.

Njih je samo **pet posto** od ukupno aktivnih poduzeća. Naime, prema podacima Hrvatskih izvoznika, od 70-ak tisuća aktivnih kompanija u **Hrvatskoj**, svoje proizvode i usluge na strana tržišta izvozi samo njih 6700. Pritom je izvoznika čija vrijednost izvoza premašuje **milijun kuna** samo 3144. Ta grupa izvoznika, prema podacima udruge Hrvatski izvoznici, ostvaruje čak **96 posto** ukupnog hrvatskog izvoza.

Koliko je bitna uloga izvoznika u cjelokupnom hrvatkom gospodarstvu, potvrđuje podatak da 2688 izvoznika izdvaja **83 posto** ukupne dobiti u **Hrvatskoj**, odnosno 16,6 od **19,9 milijardi dolara**.

Upozoravajući na podatke o hrvatskom izvozu po glavi stanovnika, predsjednik Hrvatskih izvoznika **Darinko Bago**, prilikom prošlotjednog potpisivanja Sporazuma o suradnji s **Hrvatskom** bankom za obnovu i razvitak, najavio je sklapanje sličnih sporazuma s drugim udruženjima i institucijama koje mogu pridonijeti afirmaciji hrvatskog izvoza, bez kojeg, naglasio je **Bago**, **Hrvatska** nema budućnosti.

A velike zasluge za prošlogodišnji brži rast hrvatskog izvoza sigurno ima upravo **HBOR** i njegovi programi poticanja izvoza. Preko programa Kreditiranje priprema roba za izvoz i izvoza roba **lani** je odobreno 170 kredita u vrijednosti **1,25 milijardi kuna**, što je čak **448 posto** veći iznos nego **2003. godine** kada su odobrena 52 kredita, ukupno vrijedna nešto više od **279 milijuna kuna**.

I Program osiguranja izvoza zabilježio je **lani** veliki rast. U **2004. godini** osiguran je promet od **580 milijuna kuna**, što je povećanje**180 posto** prema **prethodnoj godini**, a odobreno je 357 zahtjeva, što je povećanje od **306 posto**. **Lani** je **HBOR** osigurao izvoz 67 izvoznika, za razliku od 35 u **2003. godini**. Od početka poslovanja **HBOR** je dosad isplatio 12 odšteta u iznosu **3,2 milijuna kuna**, a od toga je **lani** četvero izvoznika dobilo odštetu od **538.000 kuna**.

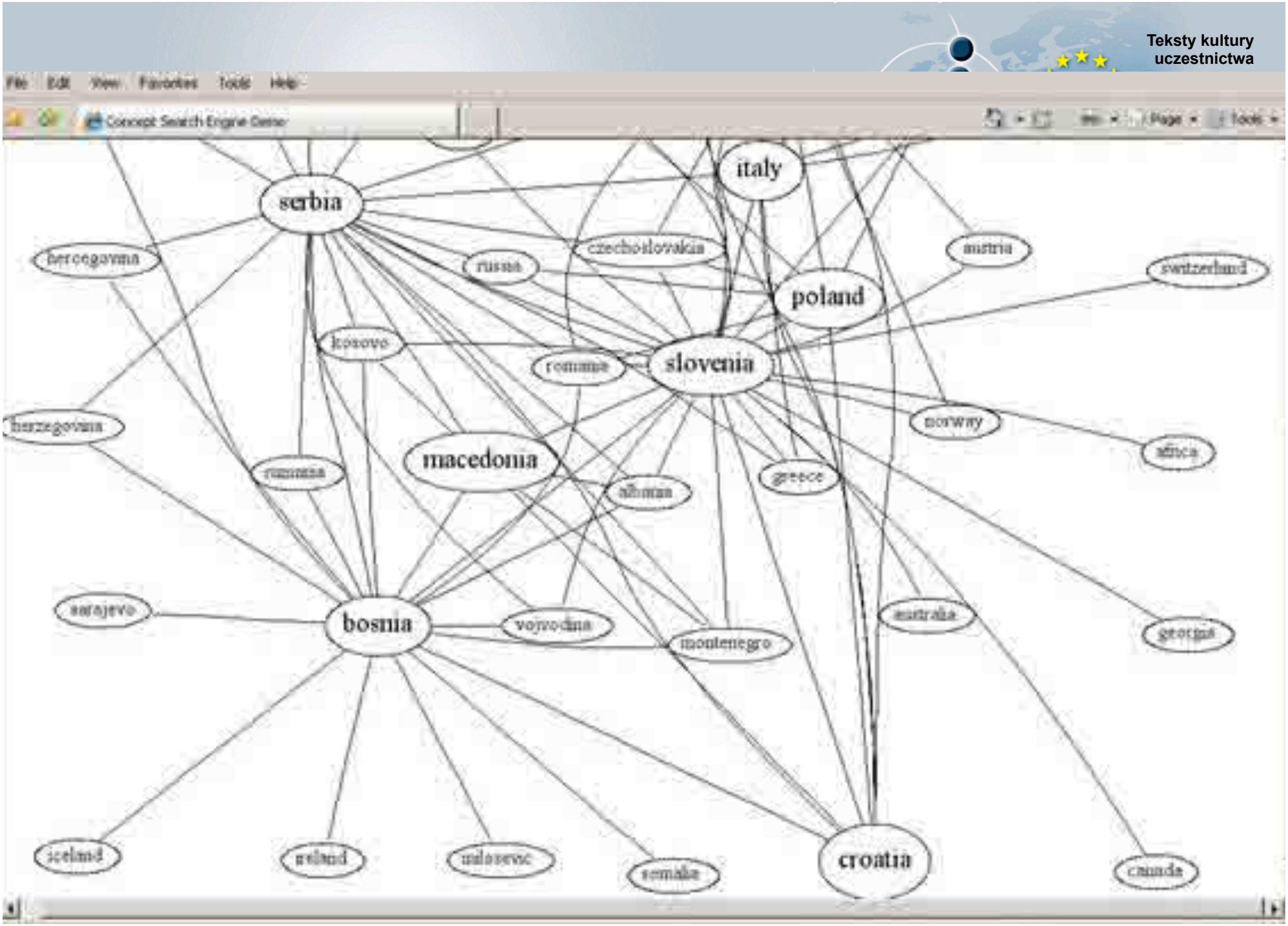
Predsjednik Uprave **HBOR-a** **Anton Kovačev**, potpisujući sporazum s Hrvatskim izvoznicima, rekao je da je **2004.** bila godina izvoza za njegovu banku te da se nada da će ova biti izvozna za cijelu **Hrvatsku**, čemu bi trebao pridonijeti i sporazum o suradnji **HBOR-a** i **HIZ-a**.

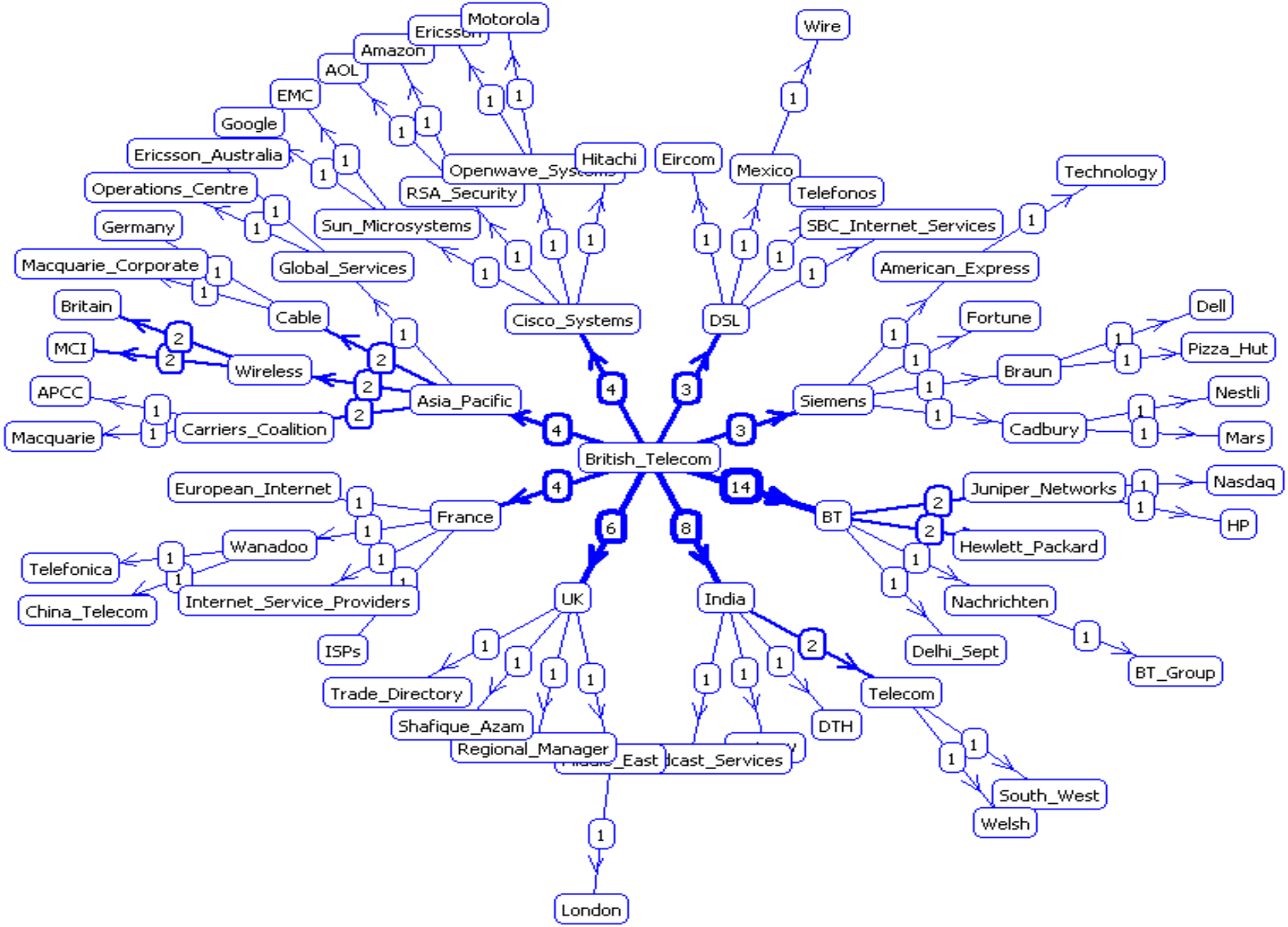
Kovačev je upozorio i da rast hrvatskog izvoza **lani** nije isključivo rezultat brodogradnje.

- Oko **90 posto** kredita koje smo dali za priremu roba za izvoz i izvoz roba odnosi se na prerađivačku industriju, poput prehrambene, metalske, farmaceutske i drvne industrije. A te industrije su ostvarile porast izvoza **6,5 posto**, što je veći rast od prosječnog ukupnog rasta od **15,7 posto** - rekao je **Kovačev**.

.

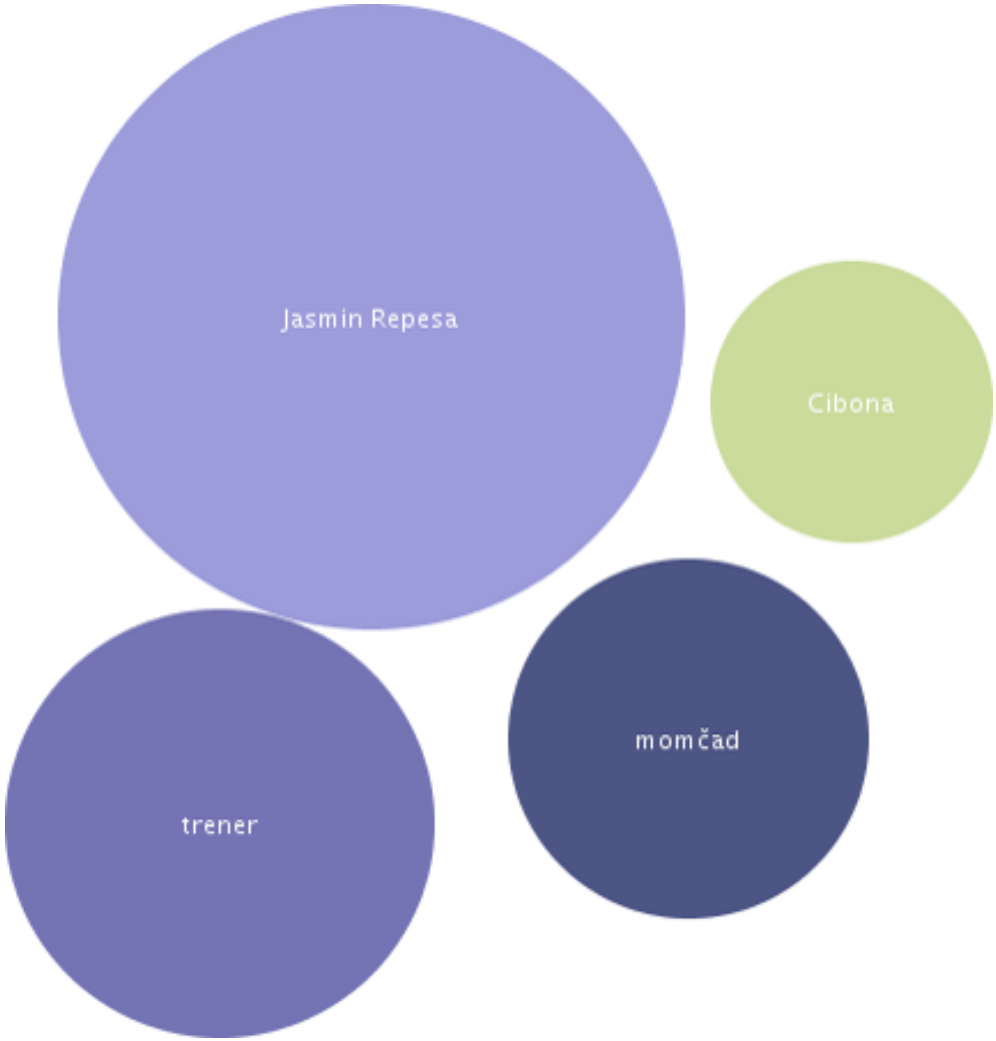
numerical and percentual values
temporal expressions
persons
locations
organizations


File Edit View Favorites Tools Help
Concept Search Engine Demo
serbia
italy
poland
slovenia
macedoma
bosnia
croatia
hercegovina
czechoslovakia
russia
austria
switzerland
kosovo
romania
herzegovina
norway
africa
rumania
albania
greece
sarajevo
vojvodina
montenegro
australia
georgia
iceland
ireland
milosevic
somalia
canada

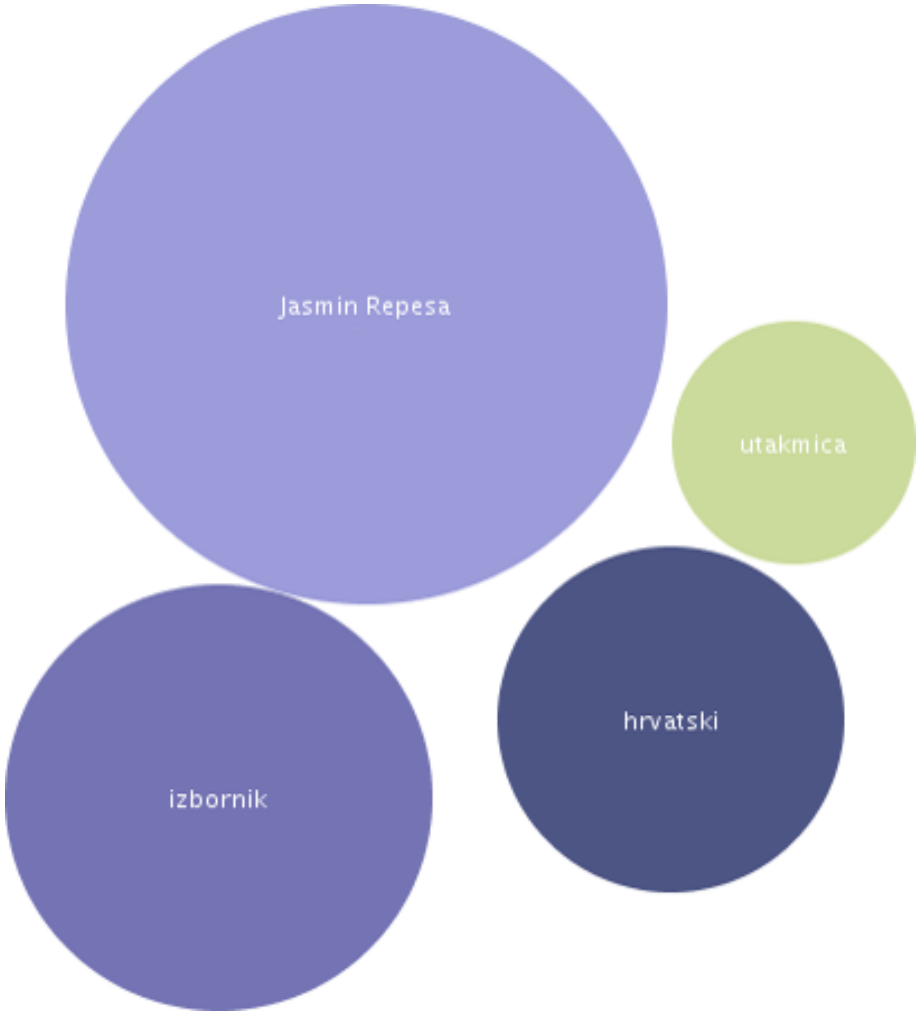

British_Telecom
Cisco_Systems
DSL
Siemens
BT
France
UK
India
Asia_Pacific
Carriers_Coalition
Cable
Wireless
Global_Services
Sun_Microsystems
RSA_Security
Openwave_Systems
Hitachi
Motorola
Ericsson
Amazon
AOL
EMC
Google
Ericsson_Australia
Operations_Centre
Germany
Macquarie_Corporate
Britain
MCI
APCC
Macquarie
Eircom
Mexico
Wire
Telefonos
SBC_Internet_Services
American_Express
Technology
Fortune
Braun
Cadbury
Dell
Pizza_Hut
Nestli
Mars
Juniper_Networks
Nasdaq
HP
Hewlett_Packard
Nachrichten
BT_Group
Delhi_Sept
European_Internet
Wanadoo
Telefonica
China_Telecom
Internet_Service_Providers
ISPs
Trade_Directory
Shafique_Azam
Regional_Manager
London
DTH
Telecom
South_West
Welsh

Tracking NE over time

2001



2009



Weblicht pipelines

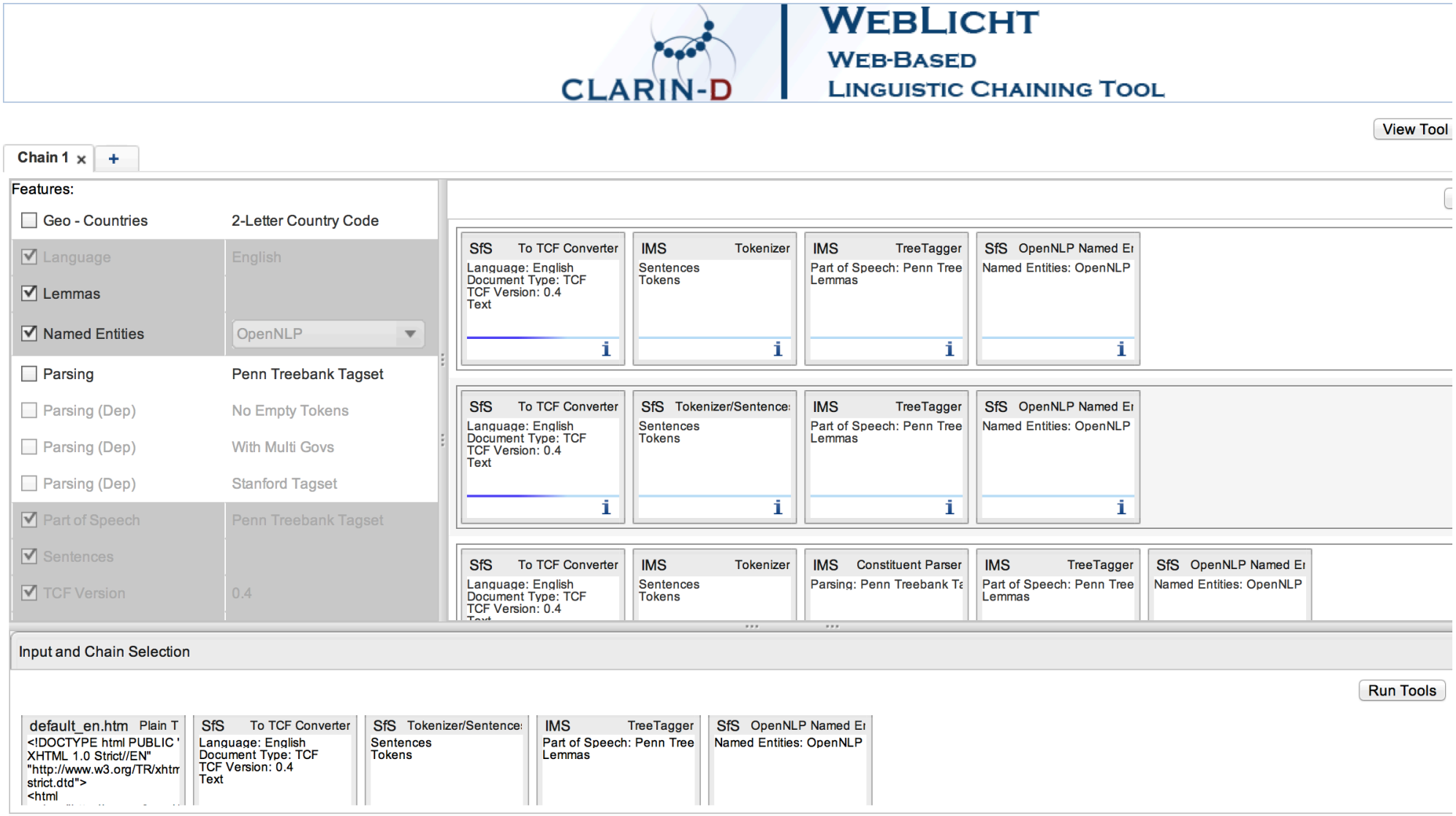
- using LT modules as basic building blocks for pipelines used in procesing texts, e.g.
  - sentence/clause splitter
  - tokenizer
  - lemmatizer
  - POS/MSD-tagger
  - named entities recognition and classification (NERC)
  - parser (subject, verb, object)
- pipelines freely configurable by user

Weblicht pipelines

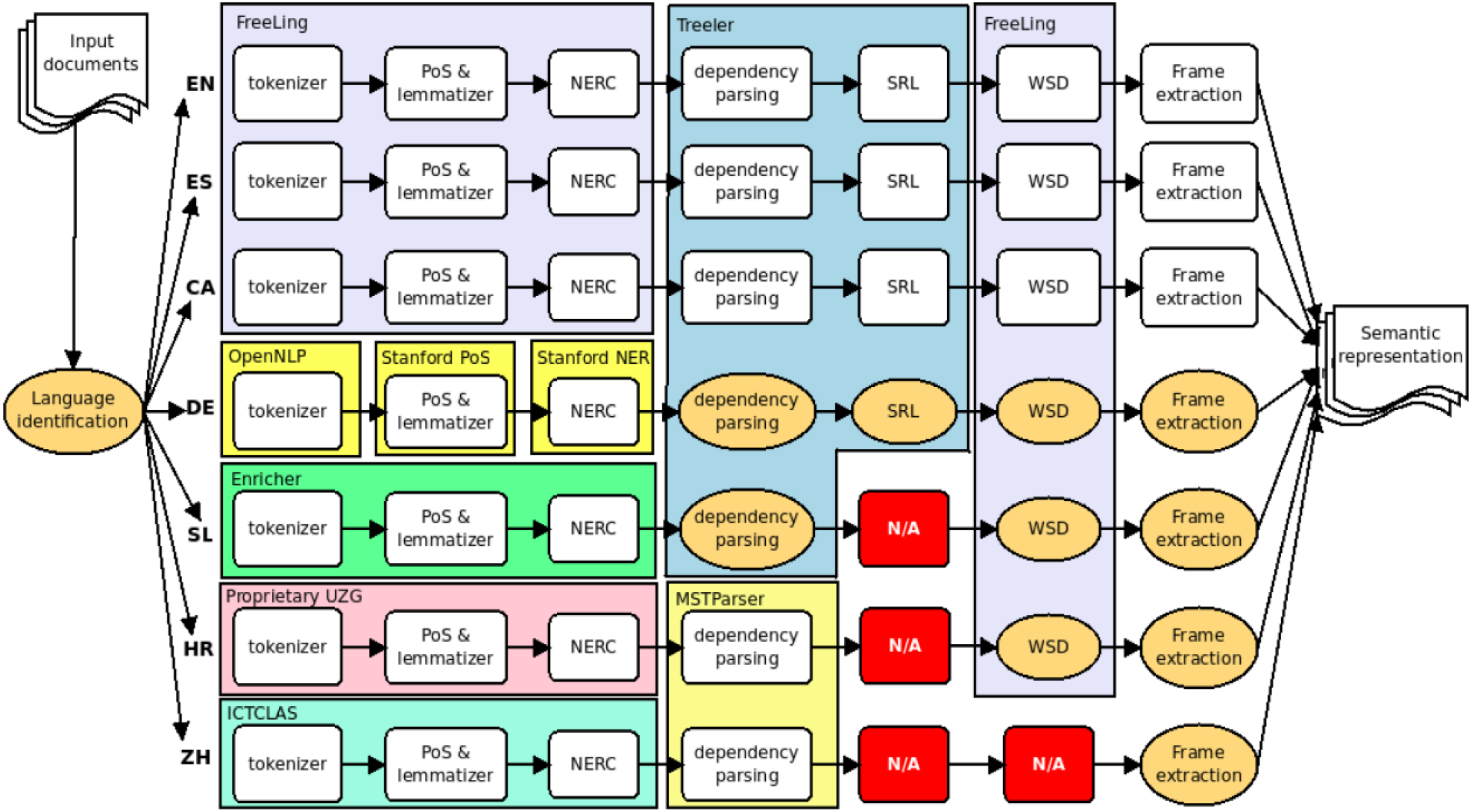
Teksty kultury
uczestnictwa
Warsaw
2013-11-07

CLARIN-D

WEBLICHT
WEB-BASED
LINGUISTIC CHAINING TOOL

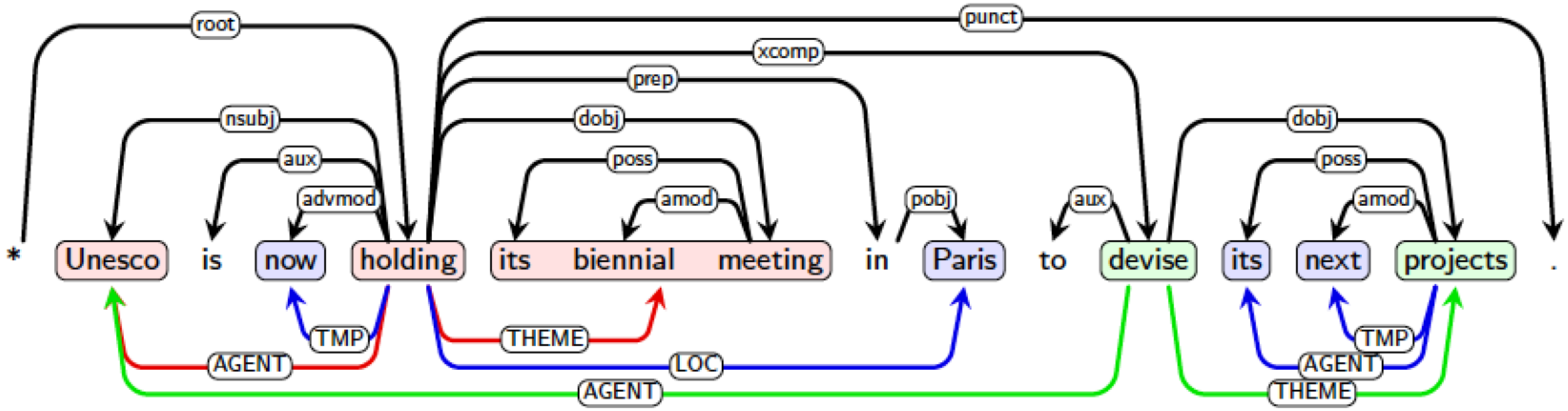


XLike project: multilingual pipelines



XLike project: multilingual pipelines

- starting from
  - “Unesco is now holding its biennial meeting in Paris to devise its next projects.”
- at the end we come to



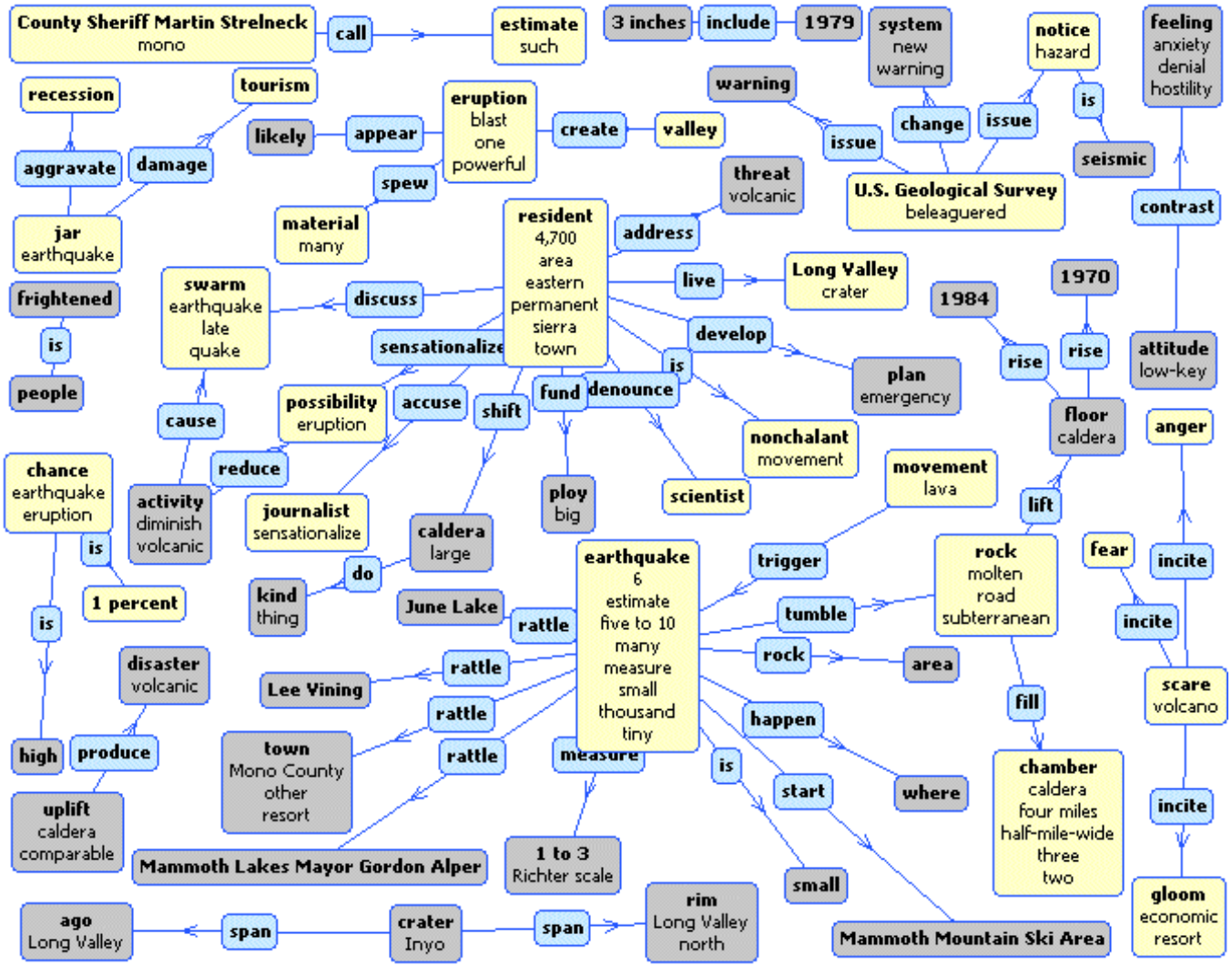
- Two propositions:

```
[EVENT: hold    [AGENT: Unesco] [THEME: meeting]
                [LOCATION: Paris] [TIME: now] ]

[EVENT: devise    [AGENT: Unesco] [THEME: projects] ]
```

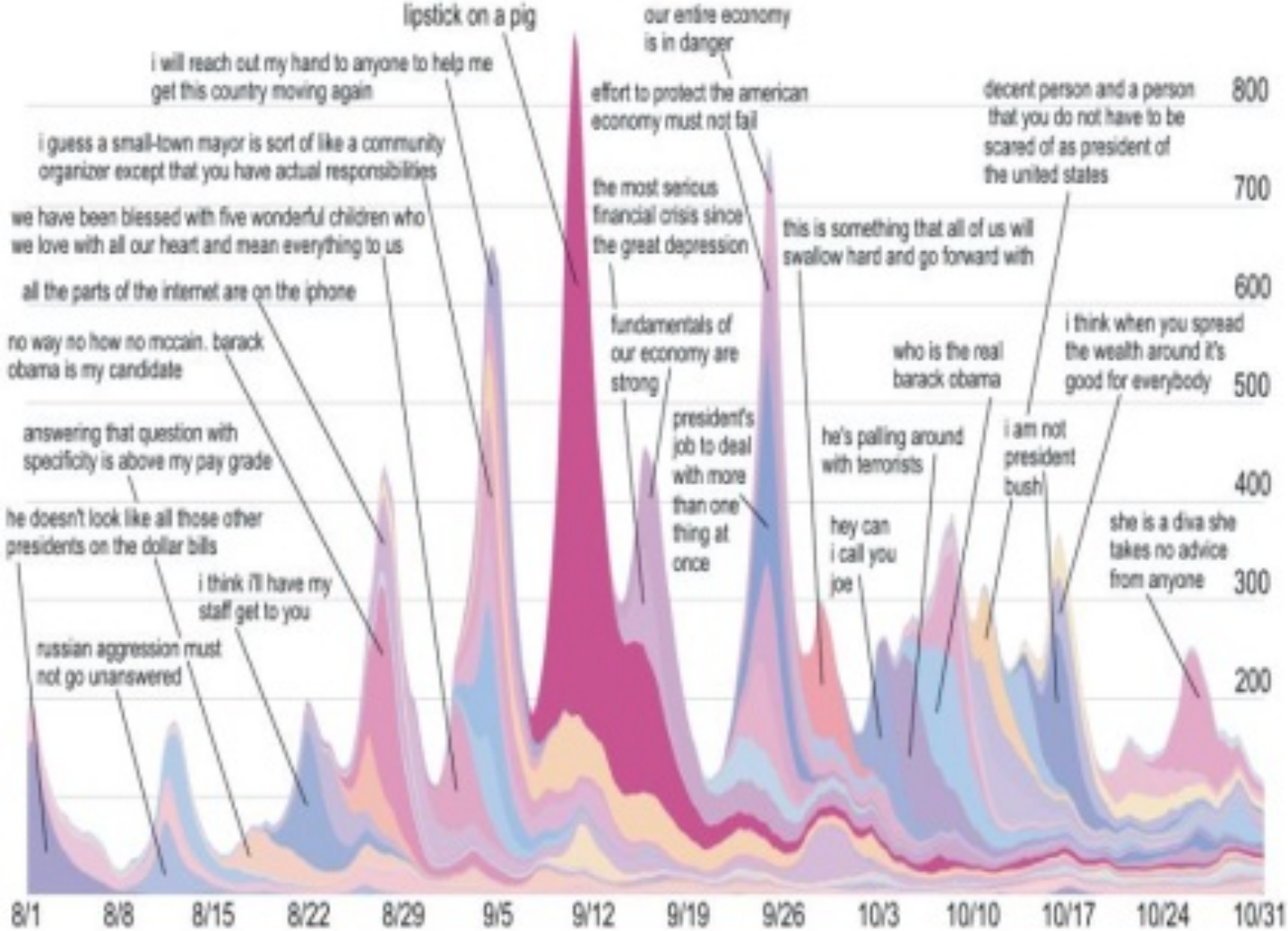
Visualisation of results

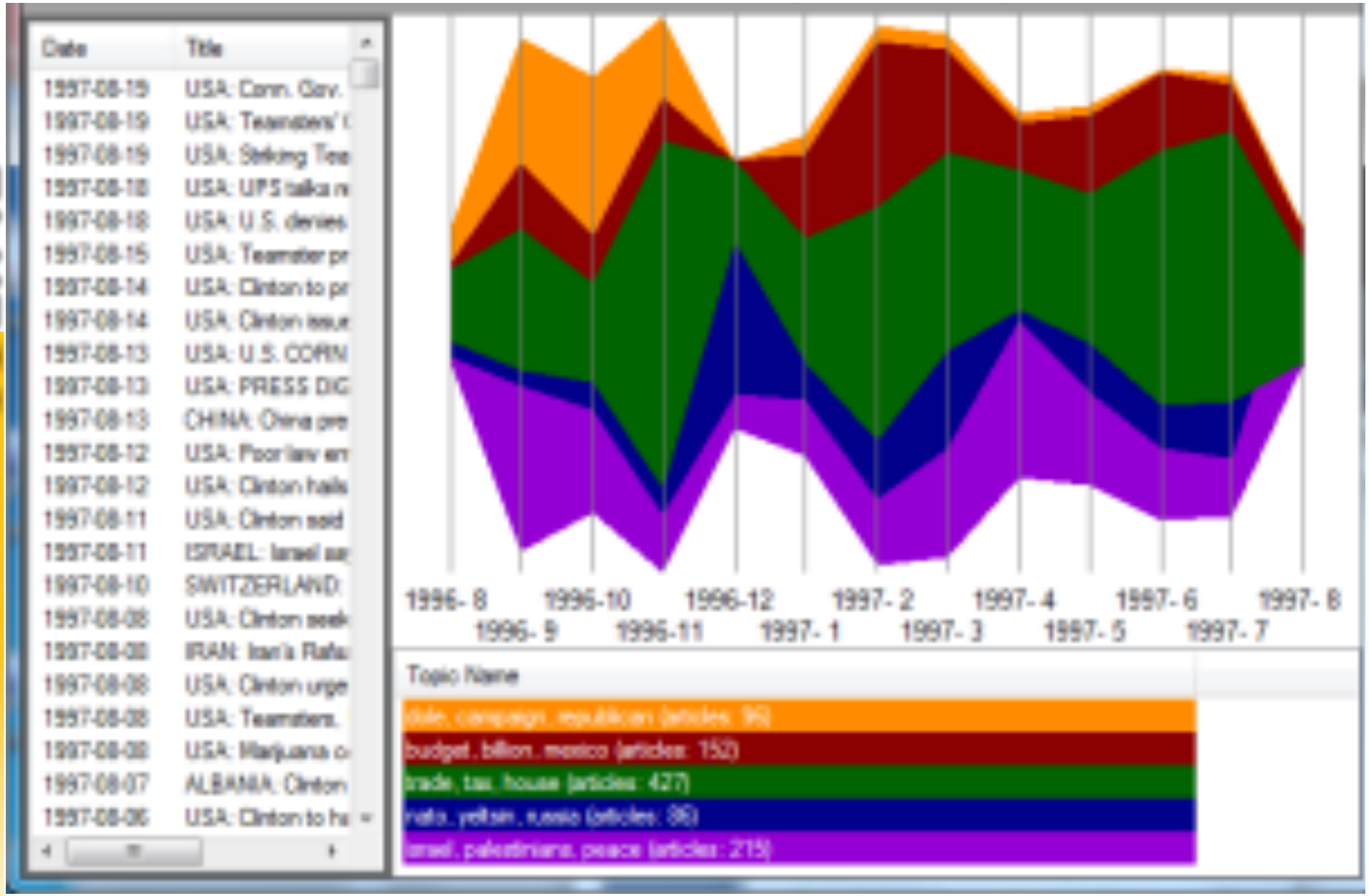
- networks



Visualisation of results

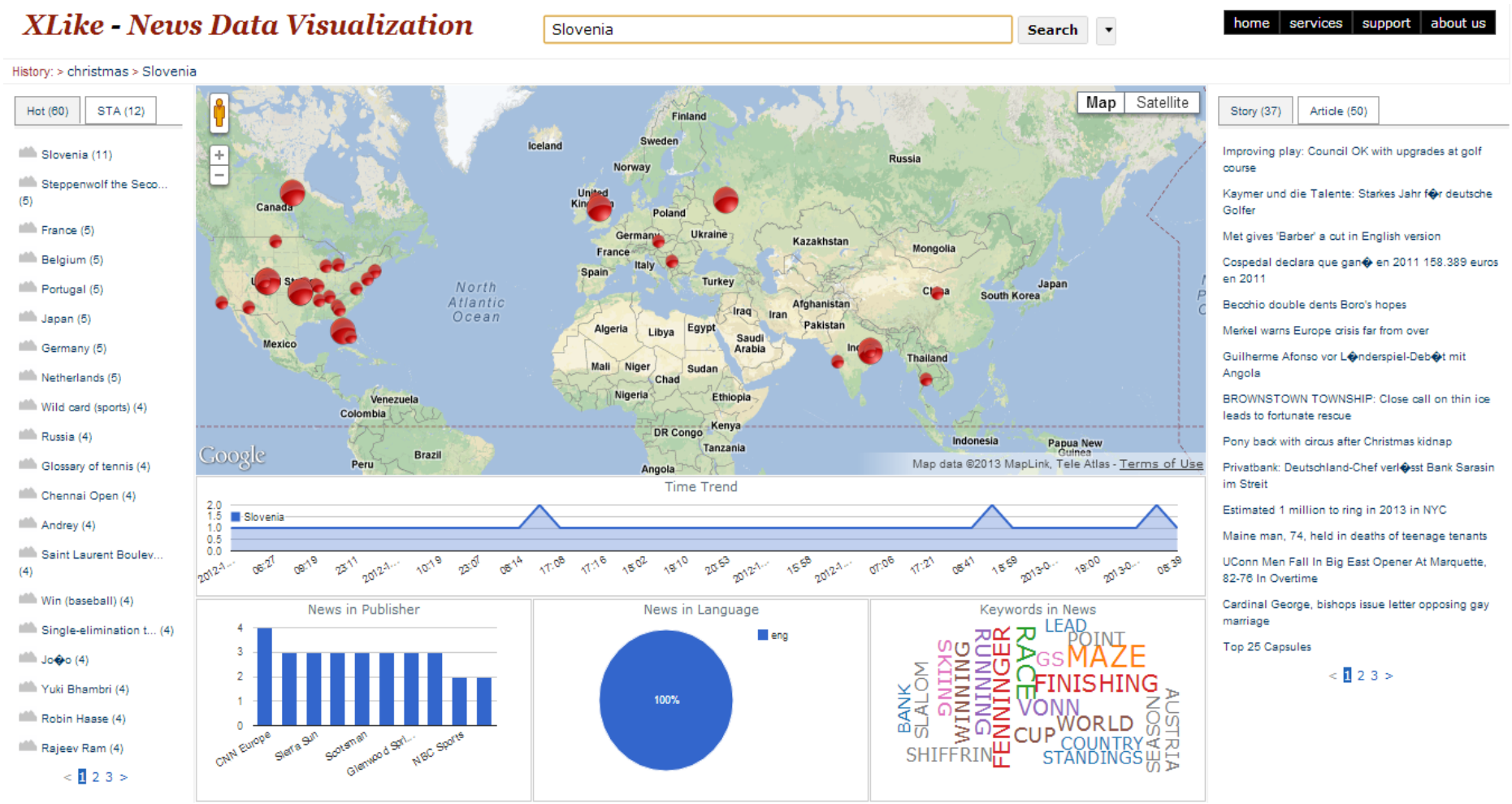
- networks
- graphs





- **timeline mapping of general topics**

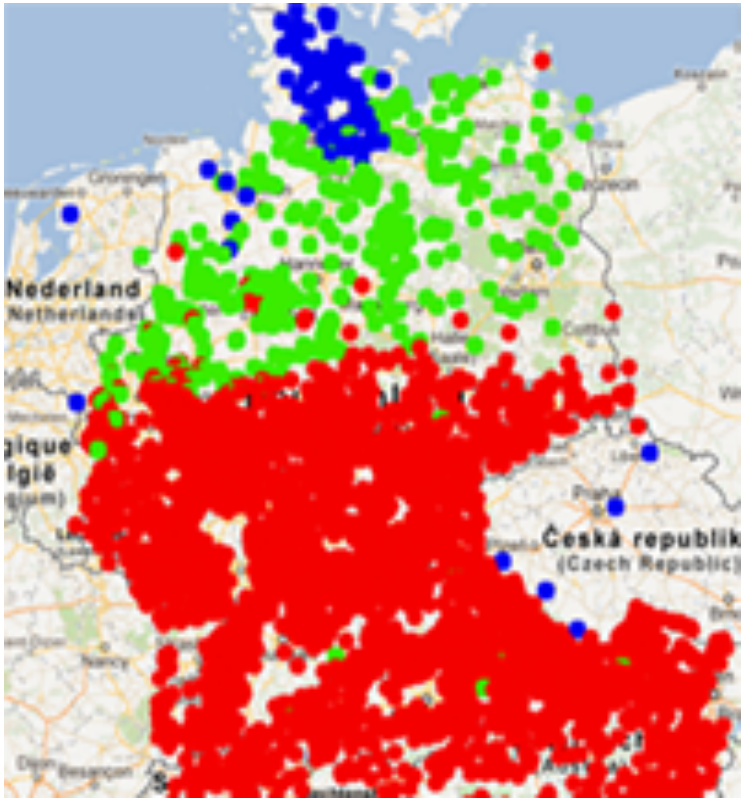
Visualisation of results



- GPS mapping of NEs mentioned (locations, countries,...)

Visualisation of results

- CiNaViz: Visualisation of European City Names
  - location names starting with “Sankt”
  - location names ending in
    - "bach" (red), "beck" (green), "bek" (bek)



CLARIN
Common Language Resources and Technology Infrastructure

Research Infrastructures in the Humanities and the role of Language Technologies

CLARIN: A European Research Infrastructure Project

Marko Tadić

Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb

Croatian Academy of Sciences and Arts

marko.tadic@ffzg.hr

Teksty kultury uczestnictwa
Polish Academy of Sciences, Warsaw, 2013-11-07

CLARA

- Marie Curie ITN
  - training early career researchers for CLARIN-ERIC RI
  - started 2009-12-01, ends 2013-11-30
  - http://clara.b.uib.no