(Technical Aspects of)

Harvesting Data from Social Network Sites

aivars glaznieks & egon w. stemle

<{aivars.glaznieks, egon.stemle}@eurac.edu>

Institute for Specialised Communication and Multilingualism

European Academy of Bozen/Bolzano
(EURAC)

EURAC research

February 14th, 2013

Researchers' Night

http://ec.europa.eu/research/researchersnight

"The Researchers' Night is a mega event taking place every year on a single September night in about 300 cities all over Europe."

Why

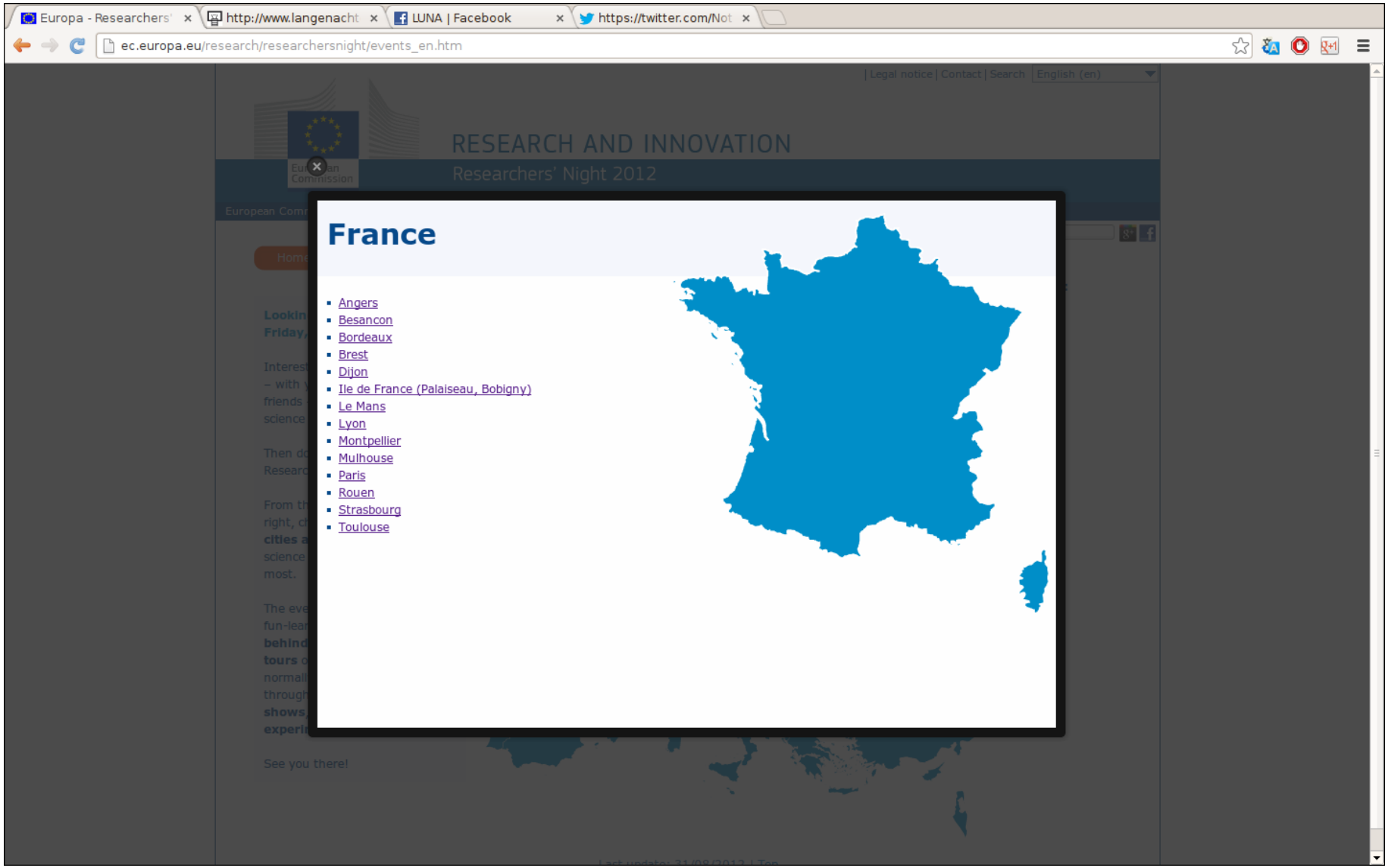
(Among other things) to see what researchers really do and why it matters for our daily life.

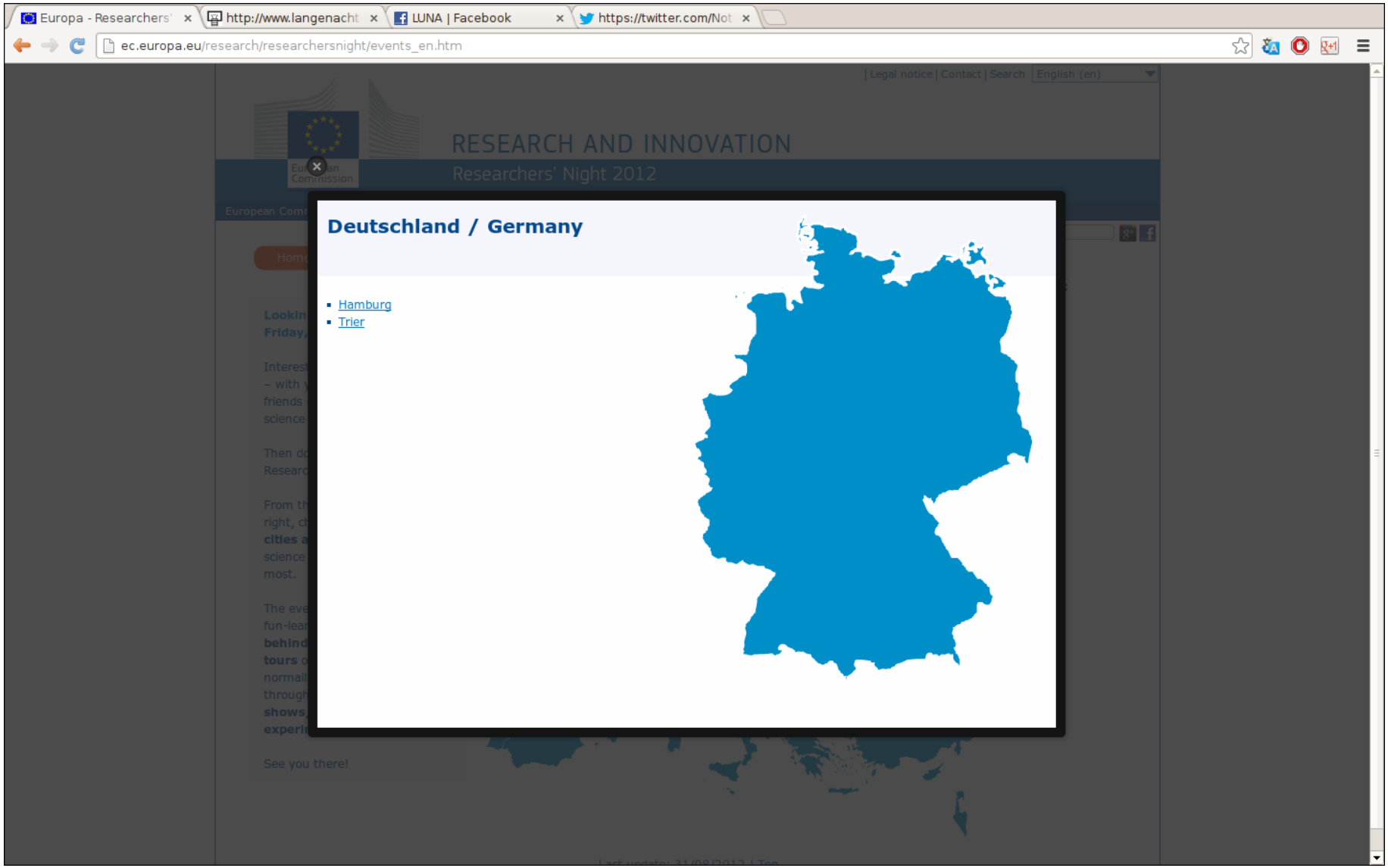
What

Different events offer a wide variety of fun-learning activities, e.g.

- behind-the-scenes guided tours of research labs (that are normally closed to the public),
- interactive science shows, and
- hands-on experiments or workshops.

Europa - Researchers'
http://www.langenacht
LUNA | Facebook
https://twitter.com/Not
ec.europa.eu/research/researchersnight/events_en.htm
| Legal notice | Contact | Search
English (en)
European Commission
RESEARCH AND INNOVATION
Researchers' Night 2012
European Commission > Research & Innovation > Researchers' Night 2012 > Events
Share
Home
Events
Media
About
Looking for things to do on Friday, 28 September 2012?
Interested in some fun learning – with your family, school or friends – through an engaging science show?
Then don't miss this year's Researchers' Night events!
From the events guide on the right, choose among about 300 cities all over Europe to find a science show that appeals to you most.
The events offer a wide variety of fun-learning activities, from behind-the-scenes guided tours of research labs that are normally closed to the public, through interactive science shows, to hands-on experiments or workshops.
See you there!
Find a nearby event in:
BELGIUM
BOSNIA AND HERZEGOVINA
BULGARIA
CYPRUS
CZECH REPUBLIC
ESTONIA
FAROE ISLANDS
FINLAND
FRANCE
FYROM
GERMANY
GREECE
HUNGARY
ICELAND
IRELAND
ISRAEL
ITALY
LATVIA
LITHUANIA

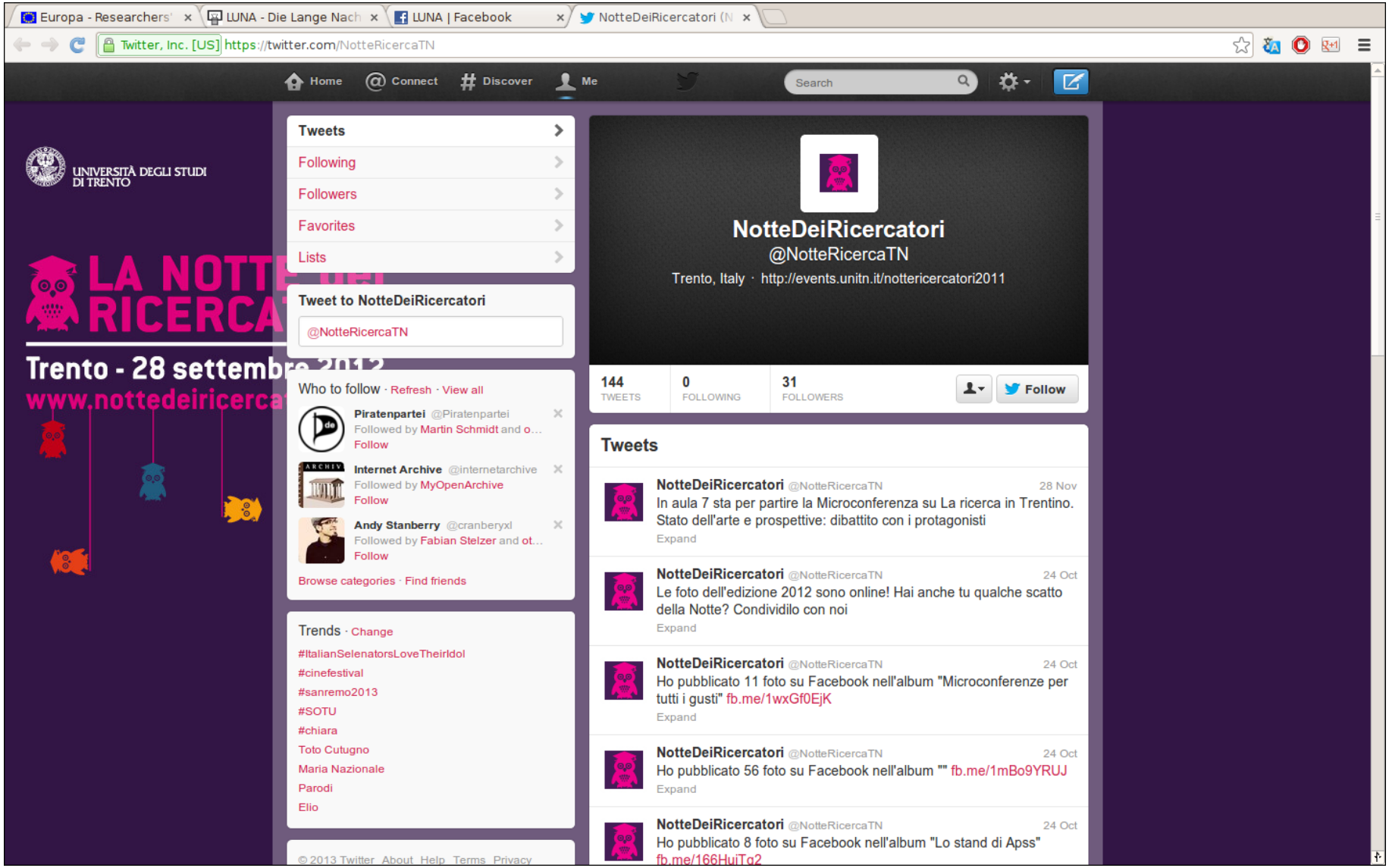

Europa - Researchers'
http://www.langenacht
LUNA | Facebook
https://twitter.com/Not
ec.europa.eu/research/researchersnight/events_en.htm
| Legal notice | Contact | Search
English (en)
RESEARCH AND INNOVATION
Researchers' Night 2012
France
Angers
Besancon
Bordeaux
Brest
Dijon
Ile de France (Palaiseau, Bobigny)
Le Mans
Lyon
Montpellier
Mulhouse
Paris
Rouen
Strasbourg
Toulouse


Europa - Researchers'
http://www.langenacht
LUNA | Facebook
https://twitter.com/Not
ec.europa.eu/research/researchersnight/events_en.htm
| Legal notice | Contact | Search
English (en)
RESEARCH AND INNOVATION
Researchers' Night 2012
Deutschland / Germany
Hamburg
Trier
See you there!

Europa - Researchers'
http://www.langenacht
LUNA | Facebook
https://twitter.com/Not
ec.europa.eu/research/researchersnight/events_en.htm
RESEARCH AND INNOVATION
Researchers' Night 2012
Italia / Italy
Albenga
Alessandria
Benevento
Biella
Bolzano
Brescia
Como
Cremona
Cuneo
Foggia
Genoa
Frascati
Lecco
Lodi
Mantova
Milano
Novara
Padova
Palermo
Pavia
Pisa
Reggio Calabria
Roma (LIGHT' 12)
Roma (RESPEcT)
Sarzana
Torino
Trento
Venezia
Verbania
Vercelli
Verona
Eventi associati
Bologna
Cesena
Firenze
Nuoro
Pisa
Reggio Emilia
Salerno
San Felice sul Panaro
Sienna

Europa - Researchers'
LUNA - Die Lange Nach
LUNA | Facebook
NotteDeiRicercatori (N
www.langenachtderforschung.it
LUNGA NOTTE DELLA RICERCA BOLZANO
LANGE NACHT DER FORSCHUNG BOZEN
28.09.2012
startseite
food & music
wettbewerb
vip lounge
foto gallery 2012
Deutsch Italiano
>> startseite
LUNA - Die Lange Nacht der Forschung 3.0
Was für eine Nacht!!! 13.000 Besucher holten sich den Energiekick bei der LUNA 3.0
Pythagoras war davon überzeugt, die Besucher der Langen Nacht der Forschung scheinen es auch zu sein: drei ist die perfekte Zahl. Die dritte Ausgabe der Langen Nacht der Forschung schlägt alle bisherigen Rekorde und ist das bislang größte Wissenschafts-Event in Südtirol: 13.000 große und kleine Besucher waren am 28. September zwischen 17 Uhr und Mitternacht in Sachen Forschung unterwegs, haben 118 interaktive Stationen an 16 verschiedenen Standorten – von Kardaun bis zum Felsenkeller der Laimburg - besucht.
Findet heraus, wer die tollen Preise gewonnen hat!
Versenden
Drucken
european researchers' night
LUNA ist Teil der Europäischen Nacht der Forscher
innovation festival
INNOVATION FESTIVAL
BOLZANO-BOZEN
wettbewerb - mach mit!
Preise/Premi
1. Mac Book Pro
13" 2.5GHZ i5 4GB/500GB/

http://ec.europa.eu/re
http://www.langenacht
LUNA | Facebook
https://twitter.com/Not
https://www.facebook.com/luna.bolzano
facebook
Email or Phone
Password
Log In
Keep me logged in
Forgot your password?
LUNA
is on Facebook.
To connect with LUNA, sign up for Facebook today.
Sign Up
Log In
28.09.2012
LUNGA NOTTE DELLA RICERCA BOLZANO
LANGE NACHT DER FORSCHUNG BOZEN
LUNA
386 likes · 1 talking about this
Like
Community
Once again, Bolzano's researchers are opening their doors to the public. On the evening of Friday September 28th, you're invited to participate in fascinating scientific experiments and activities thanks to LUNA 2012, the long night of research.
About
Photos
386
Likes
Events
Highlights
Recent Posts by Others on LUNA
See All
Marco Tim
You already have our "I like" I leave our page and invite y...
November 9, 2012 at 5:58pm
Stefano Scodro
sono usciti i risultati del concorso?
October 6, 2012 at 4:26am
Aldo Baglio
Salve a tutti! Complimenti per l'iniziativa e l'organizzazione
Likes
See All
EURAC research
Non-Profit Organization
Like
TIS innovation park South Tyrol
Organization
Like
Innovation Festival Bolzano Bozen
Organization
Like


Europa - Researchers'
LUNA - Die Lange Nach
LUNA | Facebook
NotteDeiRicercatori
Twitter, Inc. [US] https://twitter.com/NotteRicercaTN
Home
Connect
Discover
Me
Search
UNIVERSITÀ DEGLI STUDI DI TRENTO
Trento - 28 settembre
www.nottedeiricerca
Tweets
Following
Followers
Favorites
Lists
Tweet to NotteDeiRicercatori
@NotteRicercaTN
Who to follow · Refresh · View all
Piratenpartei @Piratenpartei
Followed by Martin Schmidt and o...
Follow
Internet Archive @internetarchive
Followed by MyOpenArchive
Follow
Andy Stanberry @cranberyxl
Followed by Fabian Stelzer and ot...
Follow
Browse categories · Find friends
Trends · Change
#ItalianSelenatorsLoveTheirIdol
#cinefestival
#sanremo2013
#SOTU
#chiara
Toto Cutugno
Maria Nazionale
Parodi
Elio
© 2013 Twitter About Help Terms Privacy
NotteDeiRicercatori
@NotteRicercaTN
Trento, Italy · http://events.unitn.it/nottericercatori2011
144 TWEETS
0 FOLLOWING
31 FOLLOWERS
Follow
Tweets
NotteDeiRicercatori @NotteRicercaTN 28 Nov
In aula 7 sta per partire la Microconferenza su La ricerca in Trentino. Stato dell'arte e prospettive: dibattito con i protagonisti
Expand
NotteDeiRicercatori @NotteRicercaTN 24 Oct
Le foto dell'edizione 2012 sono online! Hai anche tu qualche scatto della Notte? Condividilo con noi
Expand
NotteDeiRicercatori @NotteRicercaTN 24 Oct
Ho pubblicato 11 foto su Facebook nell'album "Microconferenze per tutti i gusti" fb.me/1wxGf0EjK
Expand
NotteDeiRicercatori @NotteRicercaTN 24 Oct
Ho pubblicato 56 foto su Facebook nell'album "" fb.me/1mBo9YRUJ
Expand
NotteDeiRicercatori @NotteRicercaTN 24 Oct
Ho pubblicato 8 foto su Facebook nell'album "Lo stand di Apss" fb.me/166HuiTq2

Harvesting Li[vf]e Data I

EURAC research

Goal

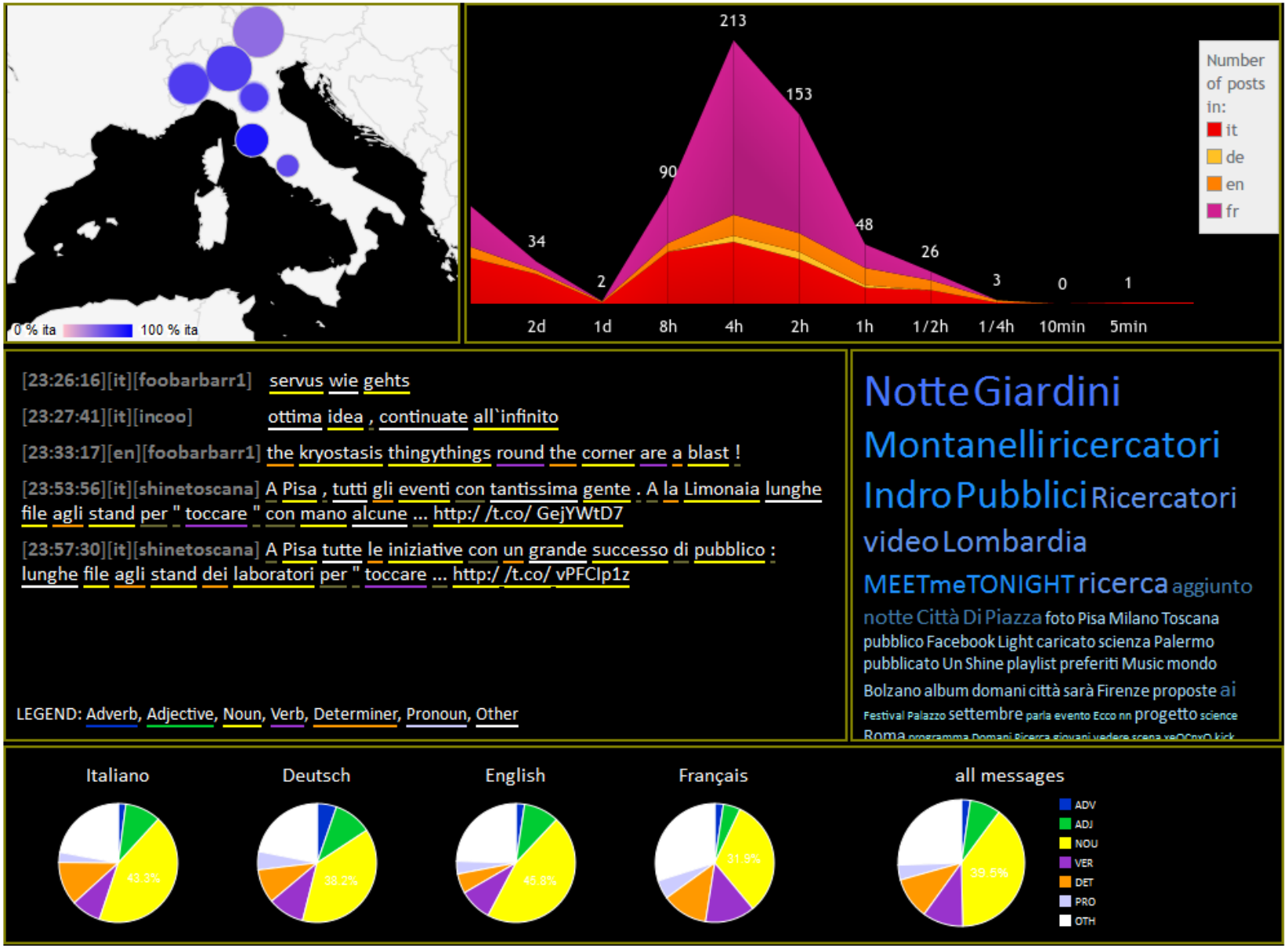
We wanted to show how *HLT* researchers

- process,
- analyse, and
- visualise data.

Means

To this end, we

- collected text snippets (FB Messages, Twitter Tweets) from parties participating in the Researchers' Night 2012,
- processed the data (added Language ID, and POS tags),
- analysed the data (extracted POS distributions, and identified salient terms), and
- used the data for visualisation.



cf. https://bitbucket.org/commul/luna3-www/, https://bitbucket.org/commul/luna3-ws/

Harvesting Li[vf]e Data II

EURAC research

Technical Means

We collected data from

- Facebook (messages) and Twitter (tweets),
- parties participating in the Researchers' Night 2012,
- people posting 'in the vicinity' of the city Bolzano,
- initially, in an asynchronous, and then, a synchronous way.

We used

- the Compact Language Detector embedded in Google's Chromium browser for language identification,
- the IMS TreeTagger for POS tagging, and
- the WaCky corpora (i.e. frequency lists) for detecting salient words.
- Finally, we used readily available (mostly Google Chart) tools for visualisation.

Challenges 1.0

EURAC research

Twitter and Facebook APIs

- The documentation of (and the discussions about) APIs were in dis-synchronisation with the 'current' version of the API.
- We encountered difficulties in 'following too many users' at the same time (+ vicinity restrictions).

EURAC research

World Wide Web 2.0

The term "Web 2.0" was coined in January 1999 by Darcy DiNucci, a consultant on electronic information design (information architecture). In her article, "Fragmented Future", DiNucci writes:

The Web we know now, which loads into a browser window in essentially static screenfuls, is only an embryo of the Web to come. The first glimmerings of Web 2.0 are beginning to appear, and we are just starting to see how that embryo might develop. The Web will be understood not as screenfulls of text and graphics but as a transport mechanism, the ether through which interactivity happens. It will [...] appear on your computer screen, [...] on your TV set [...] your car dashboard [...] your cell phone [...] hand-held game machines [...] maybe even your microwave oven."

EURAC research

It used to be Search Engines I

In 2003 Adam Kilgarriff and Gregory Grefenstette put it like this:

The default means of access to the Web is through a search engine such as Google. Although the Web search engines are dazzlingly efficient pieces of technology and excellent at the task they set for themselves, for the linguist they are frustrating wrt. for example

- maximum number of queries,
- syntactic restrictions on formulating queries,
- obscure(d) selection criteria of results, and
- obscure(d) result figures.

Well, then download the pages (i.e. the former results)

but then, you're in the business of web-page cleaning. . .

EURAC research

It used to be Search Engines II

...and in 2007 Adam Kilgarriff:

Working with commercial search engines makes us develop workarounds. We become experts in the syntax and constraints of Google, Yahoo, Altavista, and so on. We become 'googleologists'. The argument that the commercial search engines provide low-cost access to the Web fades as we realize how much of our time is devoted to working with and against the constraints that the search engines impose.