UNIVERSITÀ DEGLI STUDI FIRENZE

DiSIA
DIPARTIMENTO DI
STATISTICA, INFORMATICA,
APPLICAZIONI "G: PARENTI"

La Qualità dei dati aperti: come organizzarla e come sfruttarla al meglio –

Cristina Martelli
Fulvia Marotta
DISIA
Dipartimento di Statistica., Informatica e Applicazioni

international open data day italia 2015

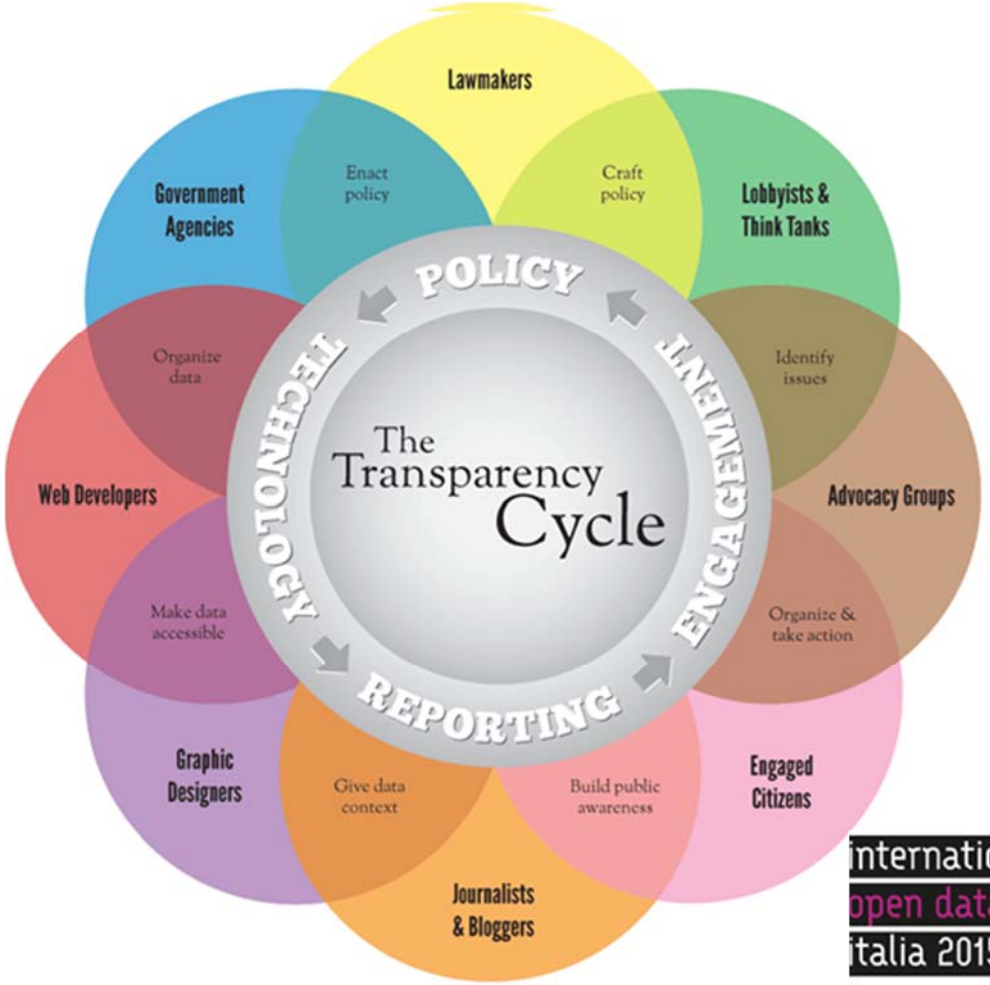
- **Qualità ed open data**

- Ruolo degli open data nella produzione di osservatori e sistemi informativi pubblici
- Gli open data e le qualità statistica delle fonti amministrative
- Il problema del governo delle fonti amministrative
- Quality Report Card for Administrative Data Sources

Informazione e ciclo della trasparenza

- Il ciclo della trasparenza mette in evidenza il ruolo della tecnologia per rendere l'informazione aperta ed accessibile

- I dati aperti si inseriscono a pieno titolo nel ciclo della trasparenza



international open data day italia 2015

Aprire non basta: ruolo della qualità

- Per decidere , comunicare e governare in contesti complessi, si sono organizzati osservatori e sistemi informativi pubblici di istituzioni, governi, agenzie nazionali, soggetti privati

- In tutti questi casi l'assunzione di qualità della informazione diffusa è fondamentale

Open data e qualità delle fonti amministrative

- Le fonti amministrative hanno una lunga tradizione di utilizzo per l'alimentazione degli indicatori dei sistemi informativi pubblici, come gli osservatori

- Si è sviluppato un ricco set di strumenti volti alla misura ed all'assesment della qualità di queste fonti che viene stabilita prima e durante il loro impiego

BLUE ETS

Input quality of administrative data

BLUE-ETS WP4

BLUE ETS

EUROPEAN COMMISSION
European Research Area

SEVENTH FRAMEWORK PROGRAMME

Funded under Socio-economic Sciences & Humanities

- BLUE-ETS Consortium ha coinvolto 16 Partners di 9 paesi europei ed è stato coordinato da ISTAT
-
- the Official Statistics Hub, include 5 EU NSIs: ISTAT-Italian National Institute of Statistics; CBS-Statistics Netherlands; SSB-Statistics Norway; SCB-Statistics Sweden; SORS-Statistics of Slovenia;
-
- the Academic Research Hub, che include 8 Università: UNIBO-Bologna; UL-Ljubljana; UNINA-Naples; UT-Trier; UNIBG-Bergamo; UoS-Southampton; UNIFI-Firenze; UoM-Manchester;
-
- Statistical and Policy Analysis Research Hub, che include 3 Istituti di Ricerca : INFOSTAT-Institute of Informatics and Statistics of Slovakia; CEPS-Centre for European Policy Study; IAB-Institute for Employment Research.
-
- ***"The composition and mix of the consortium implies that the results of the project will be circulated and spread, not only among the consortium and academia but, above all, among key EU actors and policymakers, i.e. the NSIs and EUROSTAT; the European Community; and national governments with, hopefully, feedbacks and benefits for the latter."***

The main goal of BLUE- ETS is to improve the use of administrative sources

• By developing a **standardized** way to determine the **quality** of administrative sources for **statistical purposes**:

• Dimensions of quality
• Indicators for each dimension
• Quality Report Card (QRC)

The concept of administrative data can be broadened by other publi and private organisations owning potential useful data. Technological possibilities can bring accurate, fast, high quality data and a reduction in the respondents' burden.

ETS, EUROPEAN POLICY BRIEF

Data collection in economic and business statistics: New schemes and new sources

This policy brief presents key findings, conclusions and recommendations emerging from the BLUE-ETS project as regards improving or replacing traditional surveys by using data already available in statistical databases, public administration, and alternative or secondary sources. This objective could be reached by textual data analysis and soft computing methods. The research involves analysis of textual data, more selective survey methods and exploitation of recently emerging sources such as social network or mobile device positioning data. It is stressed that these new methods of compiling data while potentially cost-reducing also raise new problems of protecting privacy.

BLUE-ETS is an FP7 research project, G.A. n.244767

international open data day italia 2015

Le dimensione della qualità dei dati amministrativi

1. Technical checks

• Technical usability of the file and data in the file

2. Accuracy

• Extent to which data are correct, reliable and certified

3. Completeness

• Degree to which a data source includes data describing the corresponding set of real-world objects and variables

4. Time-related dimension

• Indicators that are time and/or stability related

5. Integrability

• Extent to which the data source is capable of undergoing integration or of being integrated

accuratezza

- Objects with incorrect Identification numbers (ID's)
- In the Netherlands all people have a Citizen's Service Numbers
- 9-digit number (e.g. 123456782)
- Number has a feasibility check, last digit is a checking digit
- Rule used: sum(9*n1 + 8*n2 + 7*n3 + 6*n4 + 5*n5 + 4*n6 + 3*n7 + 2*n8 – 1*n9)

Remainder of sum/11 should be 0

- In the Social Statistical Database* it was found (in 2000) that:
- 0,3% of all persons in admin. data sources used had an invalid

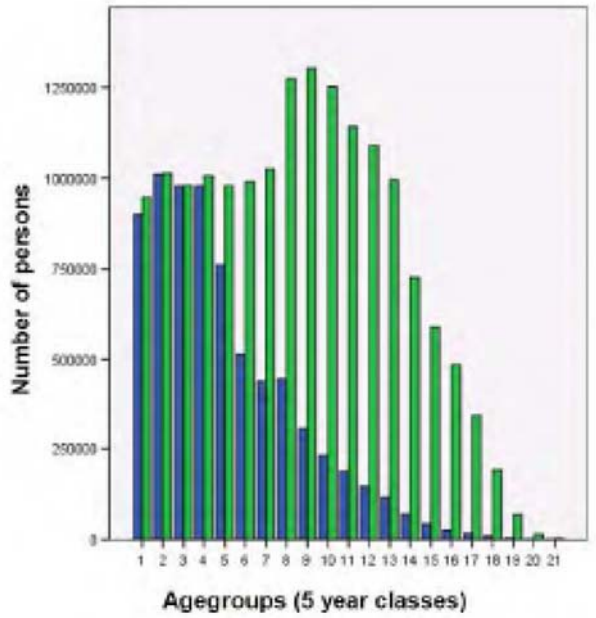
Citizen Service Number

*set of integrated admin. data sources and surveys (then ~100 million admin records)

Arts et al. (2000) *Netherlands Official Statistics* 15, pp. 16-22.

selettività

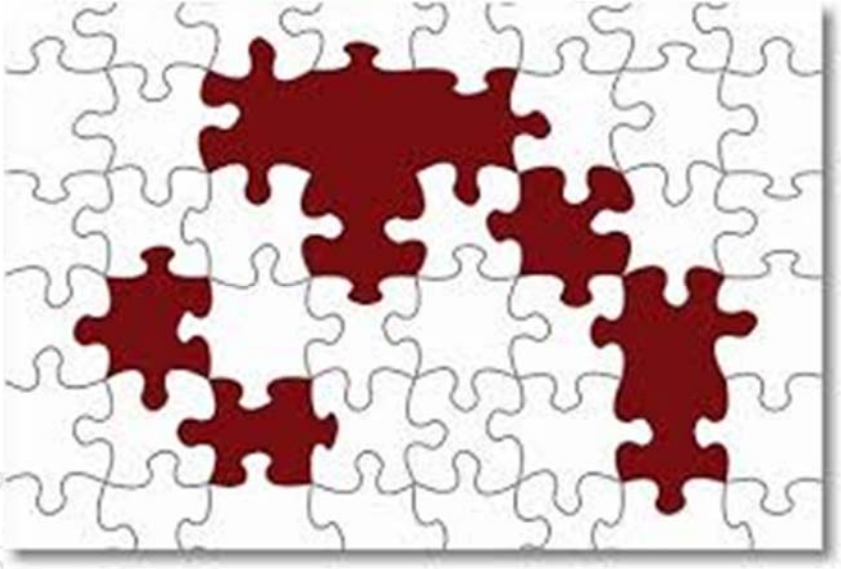
Examples of input data quality indicators:
Completeness: Selectivity



The education register has *age-related* undercoverage of educational attainment (56,3% is missing)

Explanation:
1) Children <15 age have a known level of education
2) Level of education of young adults is usually stored in recently created admin. data sources
3) Information from 'middle-aged' people is obtained from LFS-survey (small compared to admin. data info)
4) Information of 'elderly' people (≥65 year) almost completely missing (not surveyed and hardly registered)

Examples of input data quality indicators: *Completeness: Missing values*



Events recorded some time after they have occurred

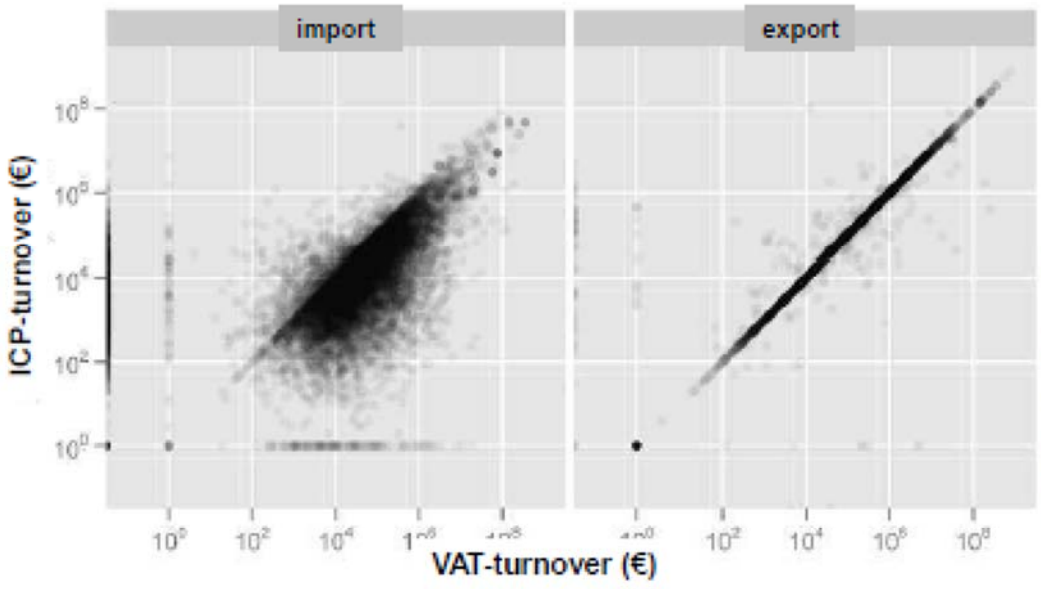
- Events are missing (or erroneously recorded)
- Particularly important for sources used immediately
- Examples:
- Netherlands: Marriages contracted in immigrants' country of origin are sometimes recorded two or three years after the event (Bakker et al. AIOS-paper 2008)
- Norway: Corrections in Persons Register are received over a lengthy period. Even months after the event has taken place (Zhang, presentation in 2011)

~ Netherlands and more: Part of VAT-data is reported later than is needed for monthly estimates (Vlag, ISI-paper 2011)

Examples of input data quality indicators:

Time-related: Delay

Examples of input data quality indicators:
Integrability: Allignment



Differences between two admin. data sources (ICP and VAT) both used for International trade statistics

Export aligns good but import is much more problematic!

Explanation:
-*ICP import* units are difficult to identify and can therefore not always by linked correctly

-ICP export data can be integrated well.

VAT: Value Added Tax data, ICP: Intra-Community Product transactions (EU-countries)

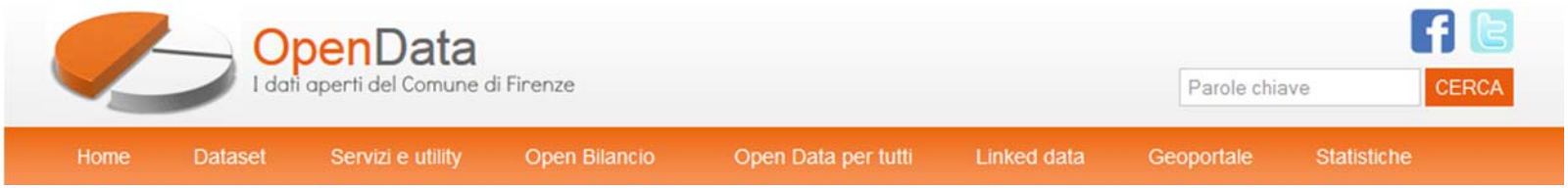
The quality report card

Quality report Card for Administrative Data viene usata per valutare la usabilità statistica delle fonti amministrative

Nella sua versione informatizzata è una lista che obbliga a valutare tutte le fondamentali dimensioni di qualità della fonte: vengono valutati degli indicatori che descrivono la qualità della fonte

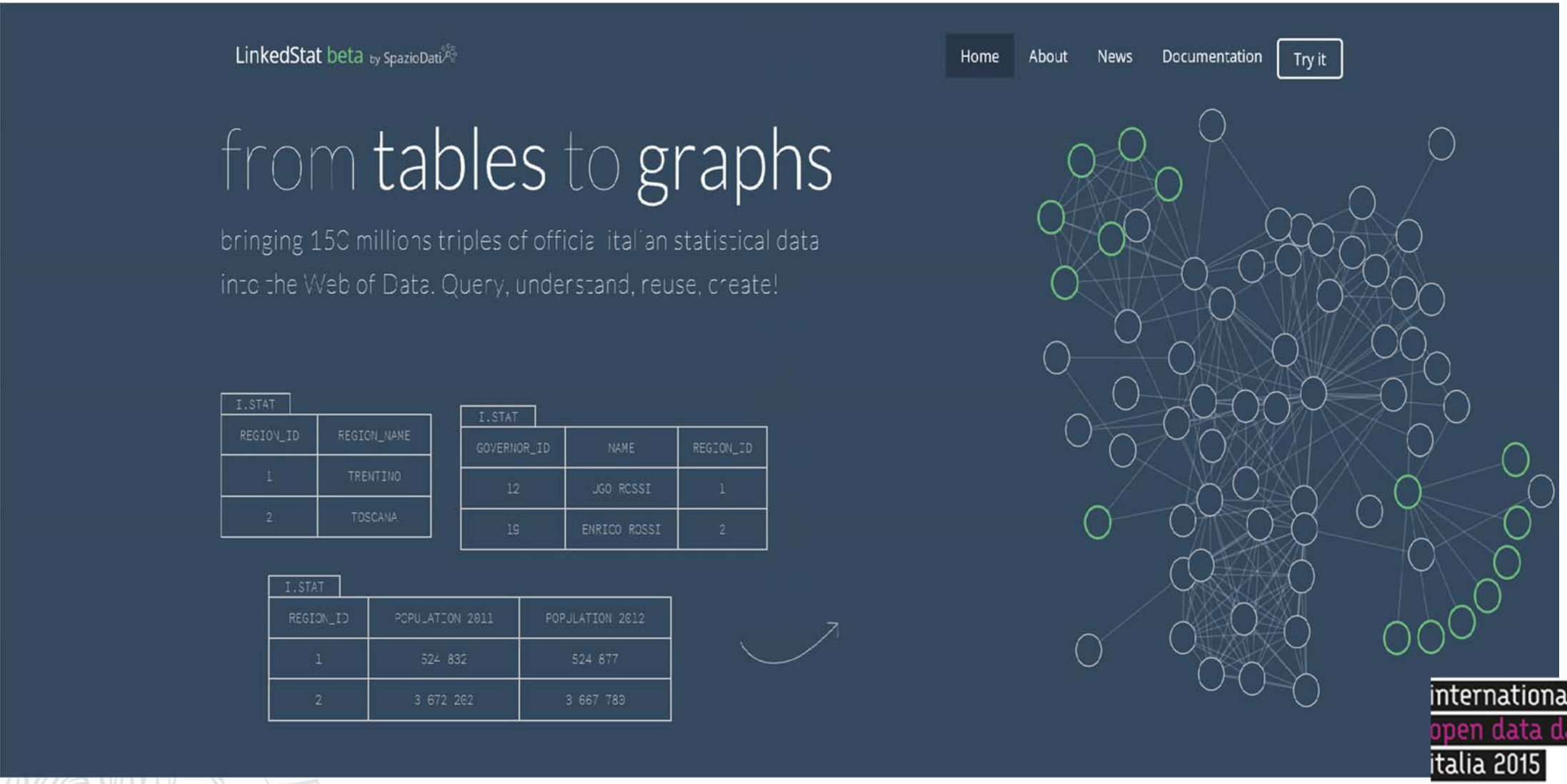
L'outcome del processo di valutazione supporta l'utente nella decisione se utilizzare o meno la fonte nel processo di produzione di un osservatorio o di un Sistema informative statistico

.



	DATI GEOGRAFICI	DATI ALFANUMERICI
RILEVANZA	Esaurienti nella voce *descrizione*.	alcune errori nel titolo e nella voce *descrizione*
ACCESSIBILITÀ	Fascia tre stelle della classifica del w3c.	fascia tre stelle della classifica del w3c
	Analisi dei metadati	
TEMPESTIVITÀ	Voce *data* non presenta errori Voce *gestione dei dati* vengono spesso riportate come risposte indicazioni ambigue. Voce *estensione temporale*	La voce *data di creazione* è sempre presente ma in dei casi porta delle ambiguità.
CHIAREZZA	Non c'è presenza di ambiguità.	Variabili con nomi che si prestano a interpretazioni inesatte. Osservazioni riportate solo per alcuni intervalli di tempo .
VERIFICABILITÀ	L'*ente* di riferimento è sempre il Comune di Firenze, e solo raramente è specificato l'ufficio di riferimento.	La voce *ente* è ben specificata
CONSISTENZA STRUTTURALE	Voci che presentano aspetti discutibili o le cui risposte sono regolarmente insufficienti o mancanti: *Livello di qualità* *Conformità* *Informazioni sulla distribuzione*	Una fonte di fraintendimenti è la difficile interpretazione della definizione delle voci: *data di creazione* *data ultima modifica al dato* *data ultima modifica al metadato*

Un esempio applicativo di uso statistico di open data di qualità statistica



international open data day italia 2015

integrabilità

Once upon a time.
I.Stat statistical data.

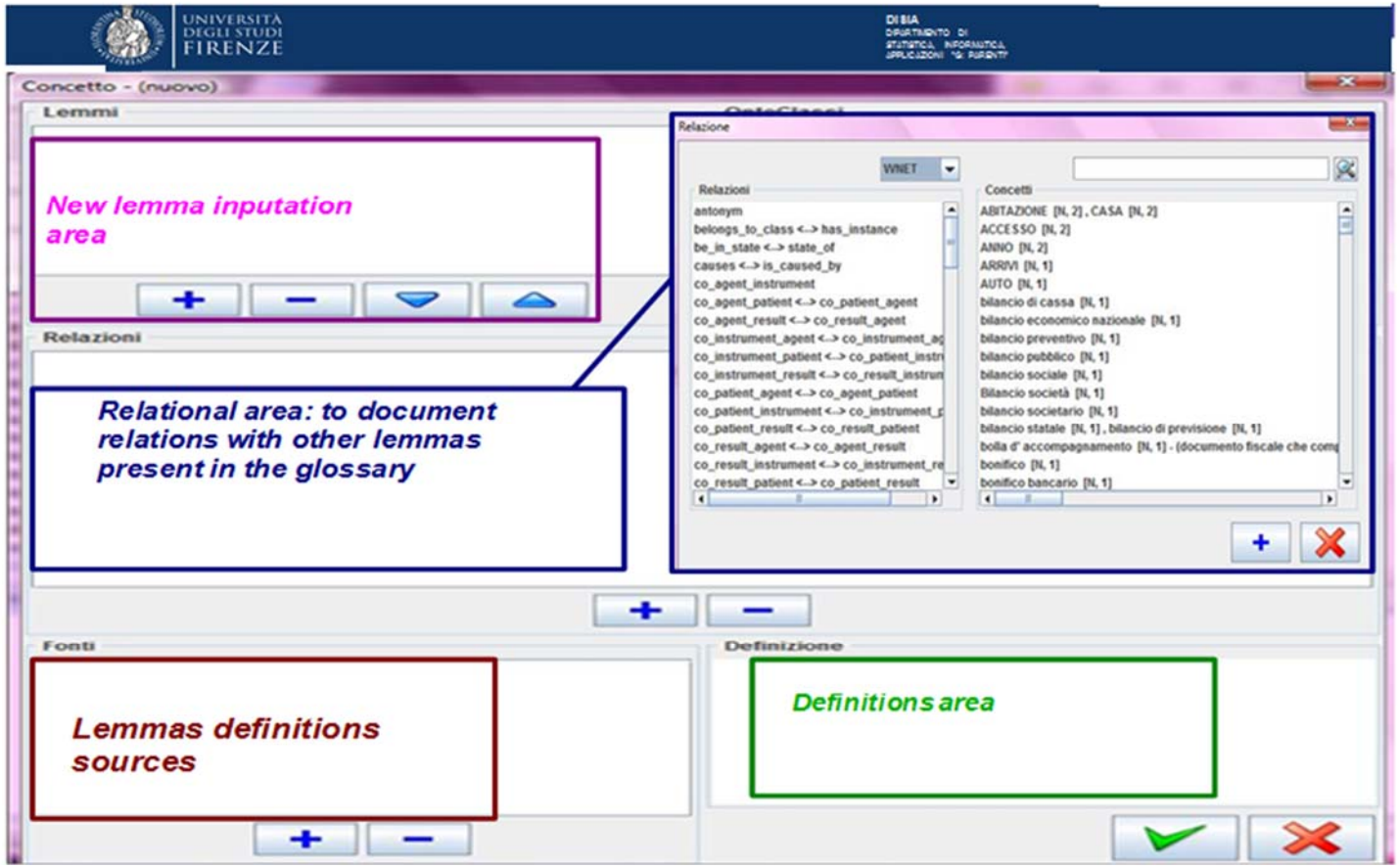
I.Stat is the corporate data warehouse and it is the main channel that Istat uses for data dissemination. Most of Istat's statistics data available on dati.istat.it are disseminated also through SEP, a web service that can be queried to get structured data in SDMX format.

★ ★ ★

international open data day italia 2015

SIS metadata documentation instruments

To properly document a statistical information system it is recomended to use a combination of Glossary, thesaurus and ontological system



concludendo

Metadati aggiuntivi per certificare l'adeguatezza degli open data all'utilizzo nell'ambito di osservatori e sistemi informativi statistici

Sistemi di metadocumentazione semantica che consentano operazioni di linkage con le fonti statistiche ufficiali

Importanza del governo delle fonti amministrative e della committenza pubblica dei gestionali nella prospettiva del riuso statistico e della pubblicazione di open data