Don’t we need data to study urban dynamics?

QuantUrb Seminar

Julien Perret

julien.perret@gmail.com
Laboratoire COGIT
Institut National de l’Information Géographique et Forestière

2015/05/20

IGN
INSTITUT NATIONAL
DE L'INFORMATION
GÉOGRAPHIQUE
ET FORESTIÈRE

Reproducible research

Reproducible research is the idea that scientific claims are published with :

- their data,
- the software code used to analyse them.

“An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.” D. Donoho

“Talk is cheap. Show me the code.” Linus Torvalds

Reproducible research

"Donoho and others have argued that computation presents only a *potential* third branch of the scientific method :"

- Branch 1 (deductive) : mathematics, formal logic,
- Branch 2 (empirical) : statistical analysis of controlled experiments,
- Branch 3 ? (computational) : large scale simulations.

"Computational science as practiced today does not generate reliable knowledge.
Computational science must develop standards for reproducibility before it can be considered a third branch of the scientific method
$\rightarrow$*Data and Code Sharing with publication*" Victoria Stodden

Barriers to Reproducible research

Survey of Machine Learning Community, NIPS Stodden, 2010 and Stodden, 2011

code(%)	reason	data(%)
77	Time to document and clean up	54
52	Dealing with questions from users	34
44	Not receiving attribution	42
40	Possibility of patents	-
34	Legal Barriers (ie. copyright)	41
-	Time to verify release with admin	38
30	Potential loss of future publications	35
30	Competitors may get an advantage	33
20	Web/disk space limitations	29

A Response

The Reproducible Research Standard (RRS) Stodden, 2009
A suite of license recommendations for computational science :

- Release media components (text, figures) under CC BY,
- Release code components under Modified BSD or similar,
- Release data to public domain or attach attribution license.

$\rightarrow$ Remove copyright's barrier to reproducible research and,
$\rightarrow$ Realign the IP framework with longstanding scientific norms.

What about urban dynamics research ?

The study of long-term urban dynamics requires important amount of data to be collected

- very time-consuming data collection
- very few data actually available as open data

But

- the most popular GIS tools are Open Source (PostGIS, QGIS, etc.)
- the most popular geographical database is Open Data (OSM)

So, what are we waiting for ?

What about urban dynamics research ?

OSM is not time-friendly

- arbitrary attributes
- but one point layer, etc.
  - $\rightarrow$ we should hide destroyed objects
  - $\rightarrow$ but then you don't see possible modifications on other timestamped layers
- it's not what it aims at

So, we could

- adapt OSM for geo-historical data (arg)
- propose something else

Context

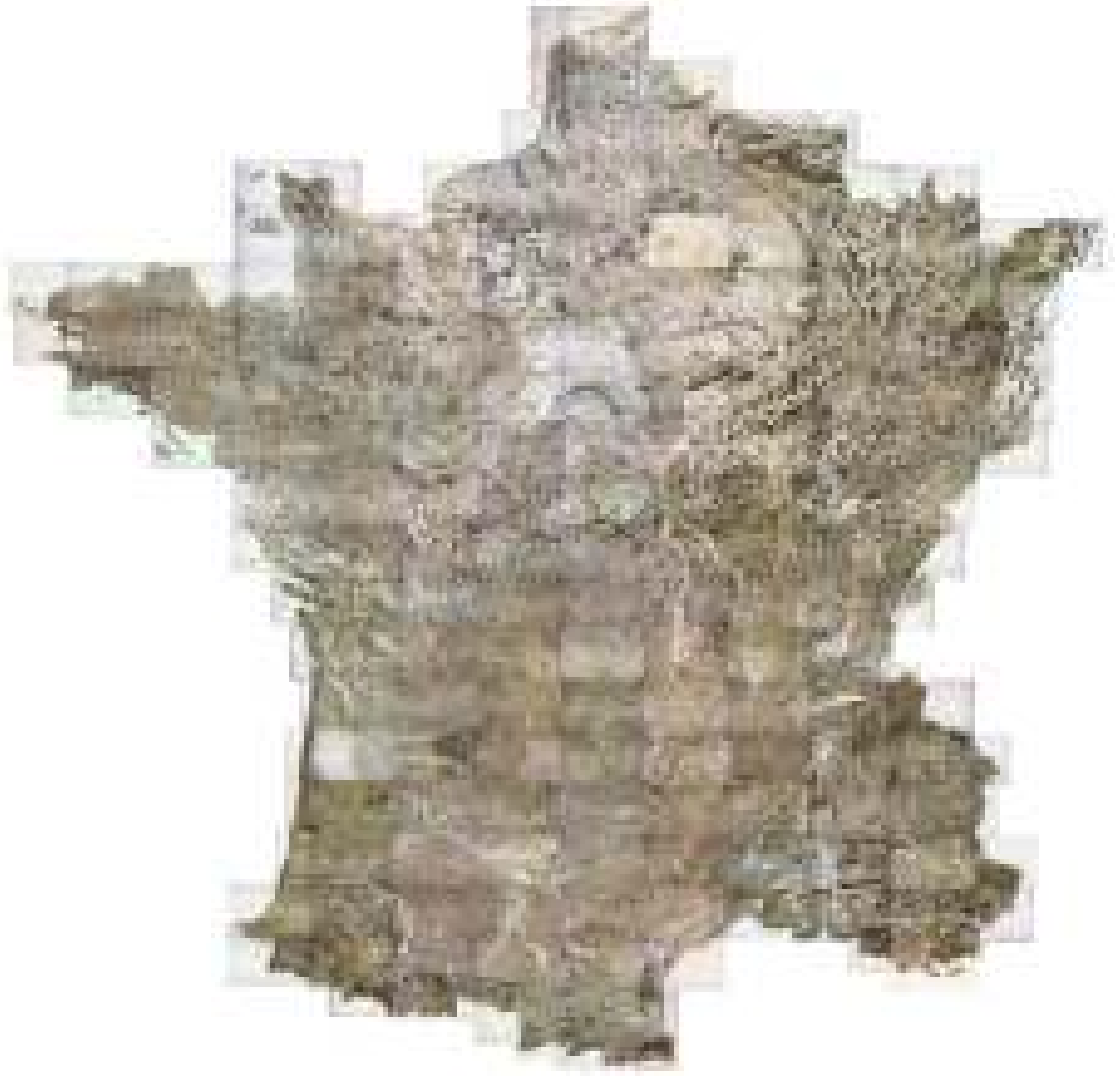
- GeoHistoricalData started in 2013 with CEA, EHESS and IGN
  - by digitizing the Cassini map (roads, cities, etc.) as Open Data
  - other sources have been brought by partners
    - several Paris road layers from SIG-PARIS (EHESS)
    - État-Major maps around Paris from Laurent Costa (ArScAn) and Sandrine Robert (EHESS)
  - new sources started (on Paris mostly)
- about 20 researchers from a dozen institutions
- collaborations with other groups (NYPL, Stanford, PSE, (ENS+Purdue) ?)
- a brand new website : www.geohistoricaldata.org

Goals

Provide tools and repositories to :

- digitize geo-historical sources
- analyse them
- share the data
- share the analysis
- support urban dynamics reproducible research

Collaborative Digitizing (2013/07 - 2014/09)



Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing (2013/07 - 2014/09)

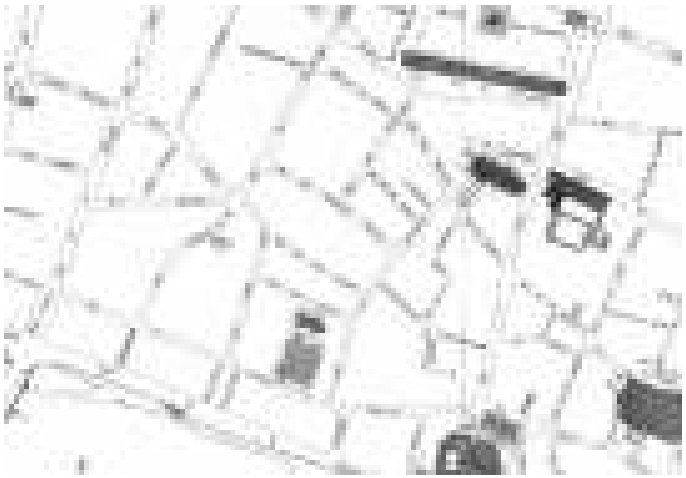
Collaborative Digitizing (2013/07 - 2014/09)

Collaborative Digitizing : contributors

N. Abadie (IGN)
S. Baciocchi (EHESS)
M. Barthelemy (CEA)
C. Bertelli (Charta SRL)
O. Bonin (IFSTTAR)
P. Bordin (Geospective)
B. Costes (IGN)
P. Cristofoli (EHESS)
B. Dumenieu (IGN/EHESS)
J. Gravier (Geographie-Cités)
M. Gribaudi (EHESS)
J.-P. Hubert (IFSTTAR)
P.-A. Le Ny (Le Ny Conseil)
E. Mermet (EHESS)
C. Motte (EHESS)
M. Pardoen (EHESS)
J. Perret (IGN)
A.-M. Raimond (IGN)
S. Robert (EHESS)
M.-C. Vouloir (EHESS)

Other Sources

Atlas National de la Ville de Paris - Verniquet (1785-1791) and Atlas Général de la Ville de Paris - Jacoubet (1825-1836)

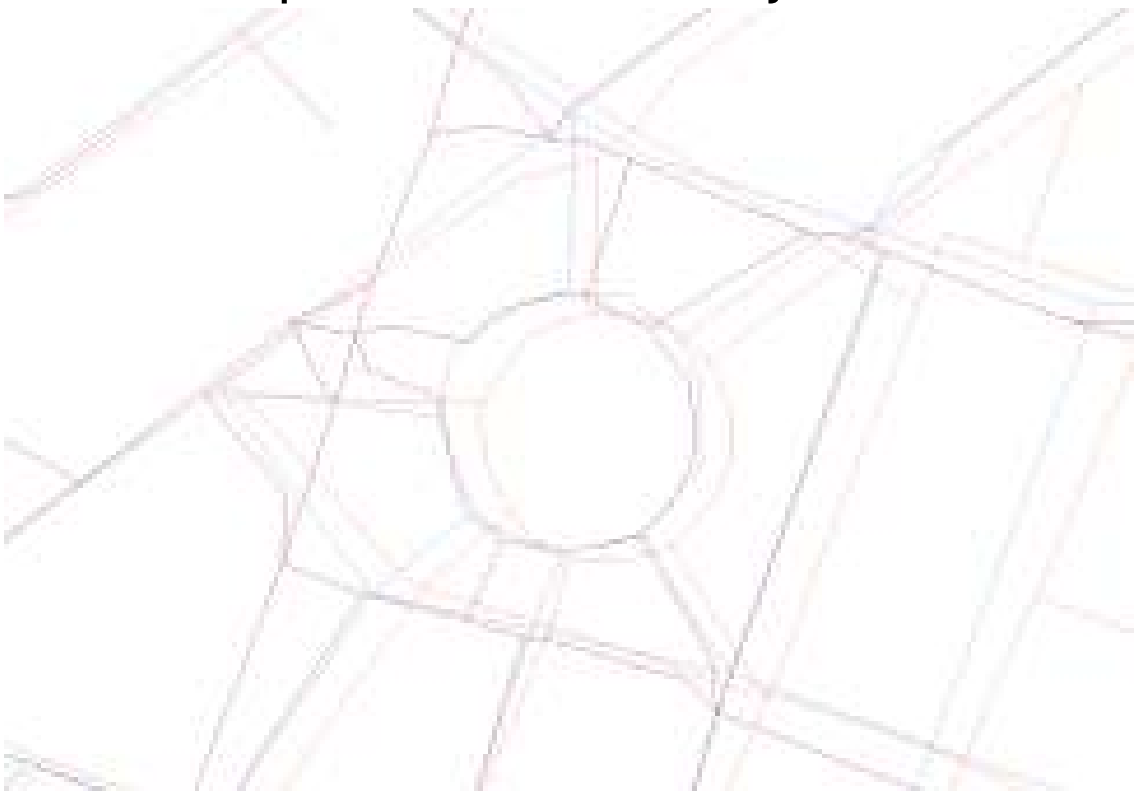


Other Sources

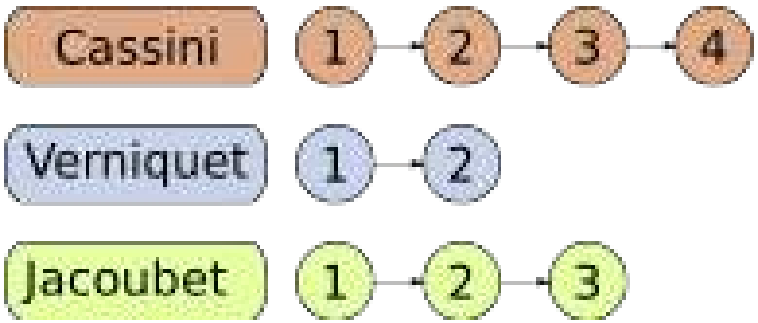
Atlas National de la Ville de Paris - Verniquet (1785-1791) and Atlas Général de la Ville de Paris - Jacoubet (1825-1836)

Data Matching

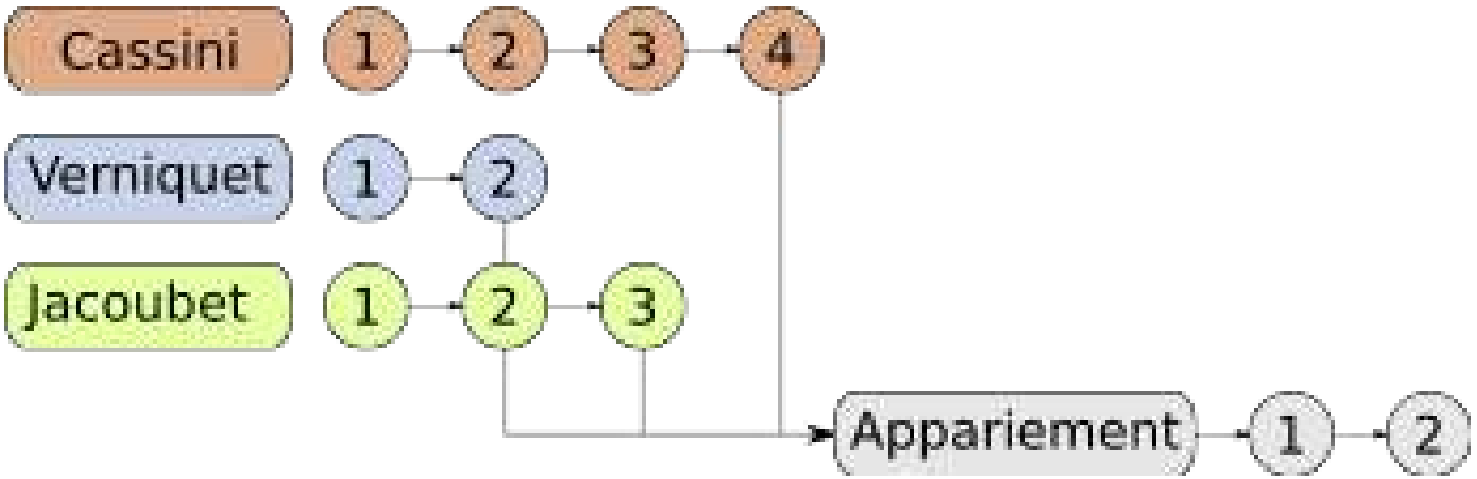
When we have several sources, can we identify the relationships between the objects ?



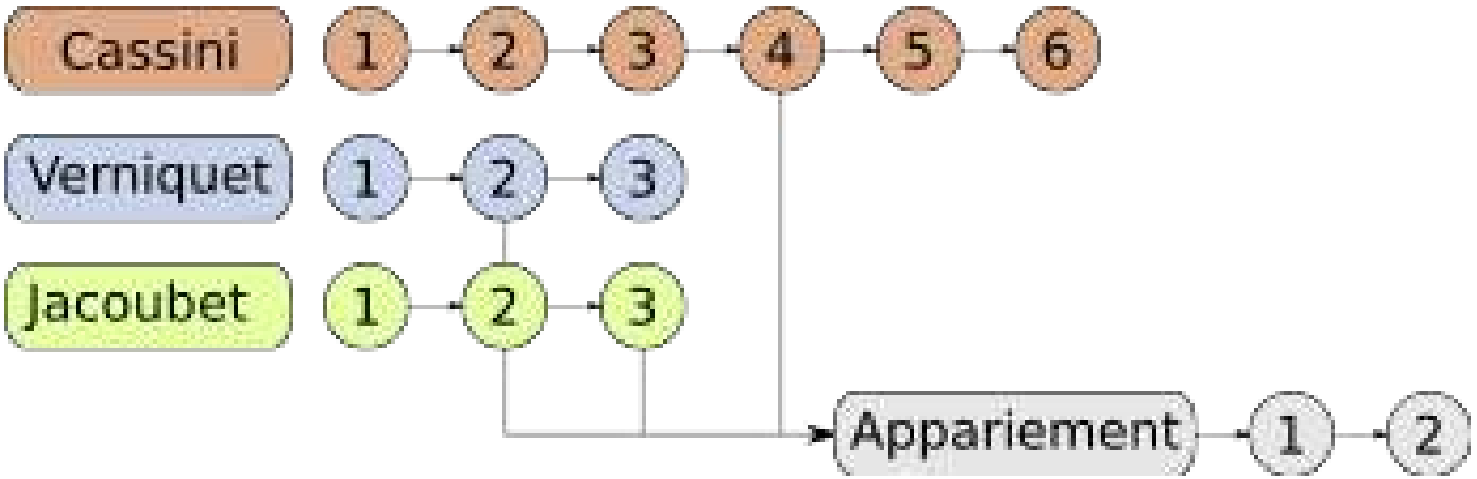
The data lifecycle



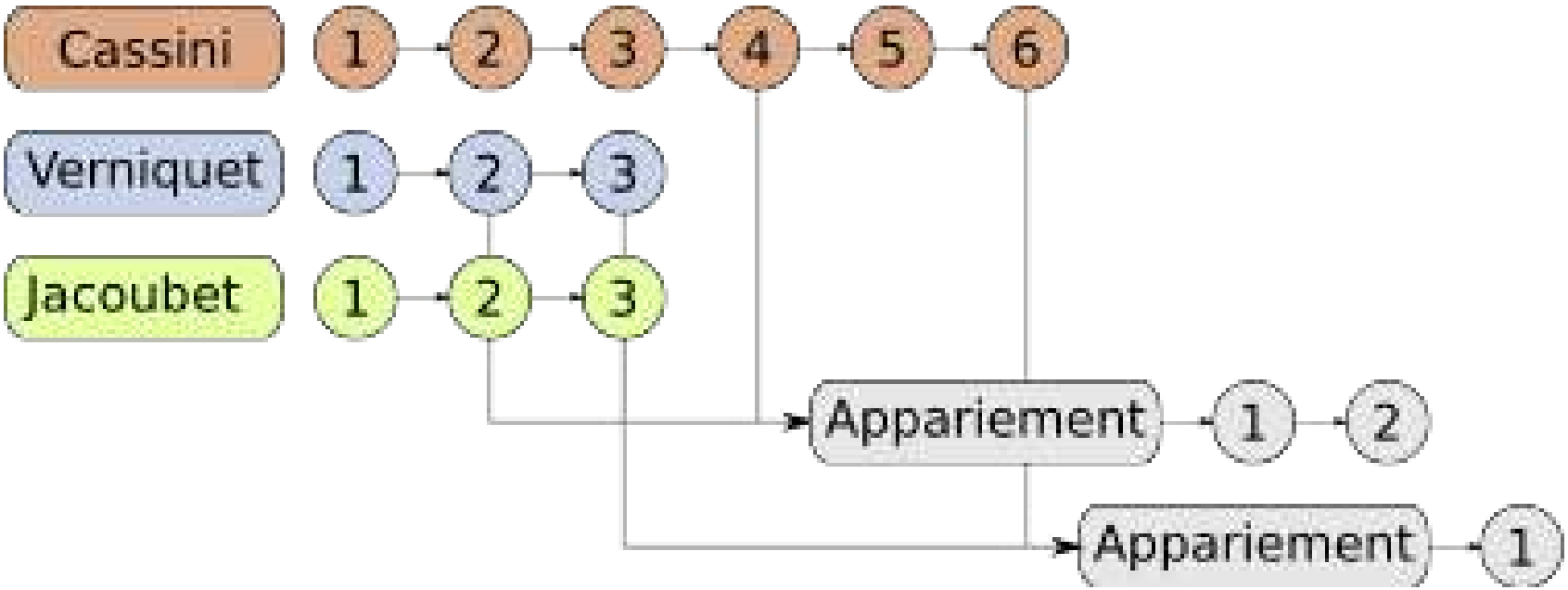
The data lifecycle



The data lifecycle

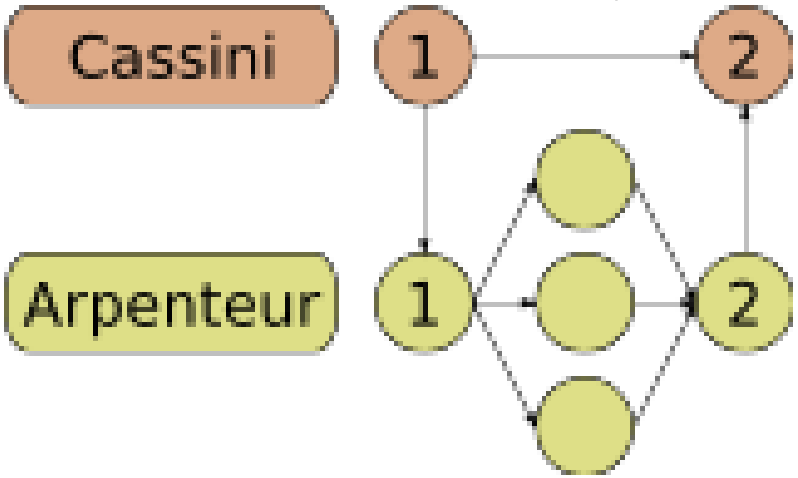


The data lifecycle



L'Arpenteur Topograhe

A fork from the NYPL Building Inspector
Uses exported cities from Cassini and online.
Allows to collaboratively validate and enrich an existing dataset



How to make a network from Cassini

The road network is not connected.
We use PostGIS Topology to create the missing edges.

http ://thedata.harvard.edu/dvn/dv/geohistoricaldata

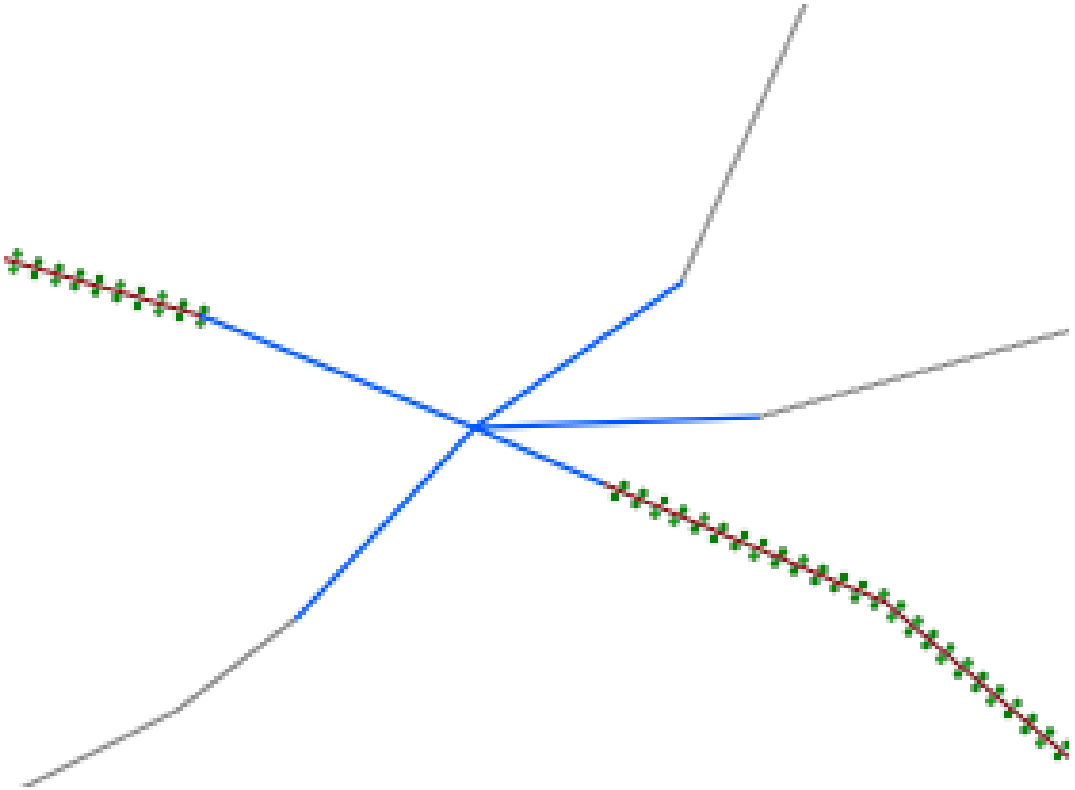
How to make a network from Cassini

The road network is not connected.
We use PostGIS Topology to create the missing edges.

http ://thedata.harvard.edu/dvn/dv/geohistoricaldata

How to make a network from Cassini

The road network is not connected.
We use PostGIS Topology to create the missing edges.



http ://thedata.harvard.edu/dvn/dv/geohistoricaldata

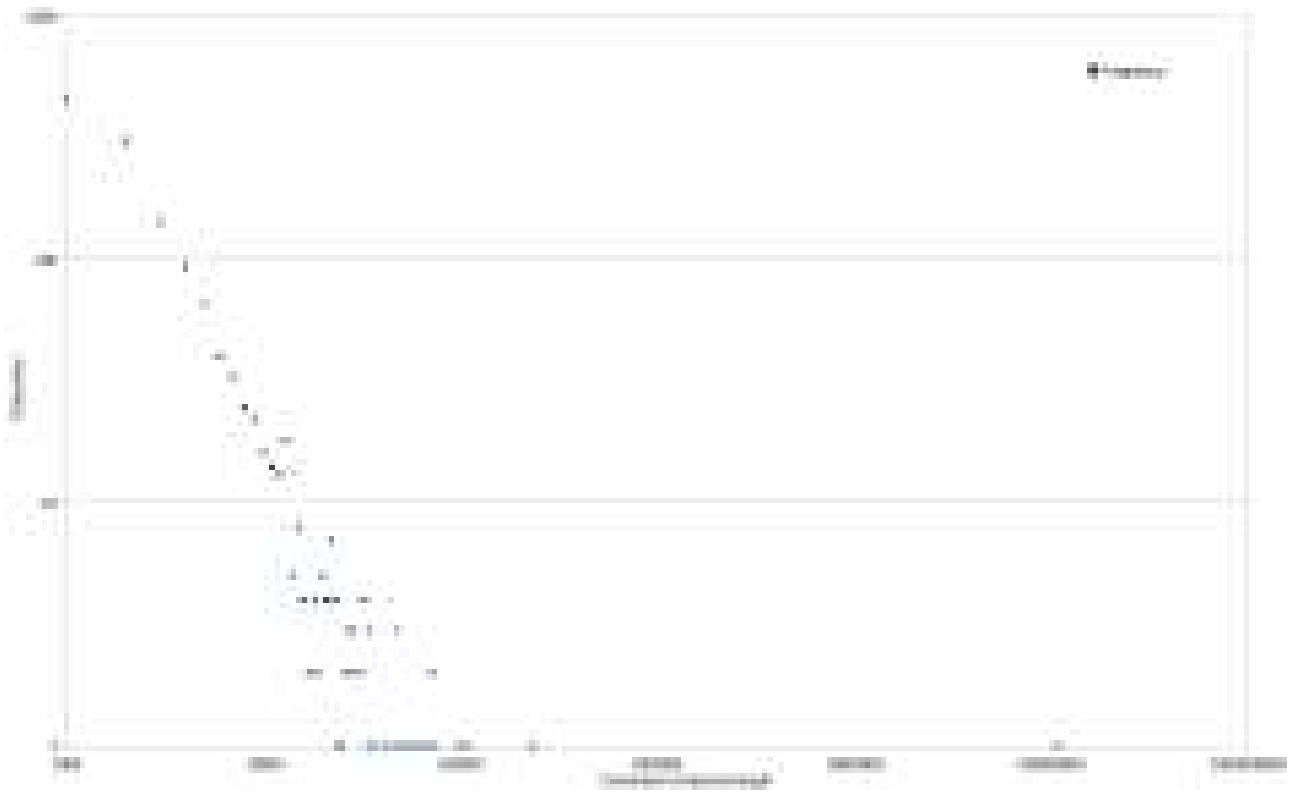
Connected Components

Connected Components

Connected Components

Connected Components

Connected Components



Challenges

- collaborative handling of specifications and their evolutions
- collaborative editing
- collaborative validation
- interaction between data/tools : distributed geo-historical databases
  - QGIS
  - Arpenteur Topographe
  - online editor
- uncertainty, imperfection modeling, etc.
- data matching
- tracking of the data and its analysis

Thank you

for your time, your attention and your ideas...
for your upcoming contributions ?

what tools would you dream of ?

are there good methods to validate a network ?